School of Engineering

CAI Centre for Artificial Intelligence

Readings in AI 2022

Publications of the ZHAW Centre for Artificial Intelligence

Volume 2, 2022

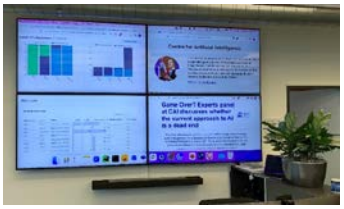
Imprint
Prof. Dr. Thilo Stadelmann
CAI Centre for Artificial Intelligence
ZHAW Zurich University of Applied Sciences
Technikumstrasse 71
8401 Winterthur
office.cai@zhaw.ch

Preface

The ZHAW Centre for Artificial Intelligence (CAI) is a hub for excellence in AI research and application. Our mission is to advance human-centric and trustworthy AI research in Switzerland and beyond, thereby providing students with career opportunities in the AI sector, attracting young talent and addressing the great challenges of our time through innovative use of AI.

We believe in the power of interdisciplinary collaboration and engaging in dialogue with the research community, with our students, and with our partners outside academia. We offer expertise in the areas of *Autonomous Learning Systems* (reinforcement learning, multi-agent systems, and embodied AI), *Computer Vision, Perception and Cognition* (pattern recognition, machine perception, and neuromorphic engineering), *Natural Language Processing* (dialogue systems, text analytics, and spoken language technologies), *Trustworthy AI* (explainable AI, robust deep learning, and AI & society) and *AI Engineering* (MLOps, data-centric AI, and continuous learning). With this, the CAI's vision is to contribute to a society that is worth living in, increasingly supported by AI-driven tools of increased generality that place humans at the centre.

Keen observers will have recognized that I wrote almost the same in last year's "Readings on AI": still true, our mission continues unchanged, but we split our focus topic on XAI into "trustworthy AI" and "AI engineering". Other things grew as well: Two new research groups where created (Frank Schilling's on "Intelligent Vision Systems" and Jasmina Bogojeska's on "Explainable AI"), we re-furnished our office to serve as a nice and practical hub for new (besides novel) work, and we completely re-financed the centre with 3rd-party funds, thereby growing our staff to 24 and counting. Highlights like a colloquium series on the intersections between AI and neuroscience (with plenty attendance from internationally renowned penalists) or DeepMind's CEO Demis Hassabis being interviewed for his Global Swiss AI Award during an AI1 lab in class were only outdone by the massive surge in AI interest due to the release of ChatGPT end of November. We will have to say much more on this next year, but to me, the most important aspect of this rather well-known technology was the "usefulness singularity" it marked: For the first time, a single AI system could be truly used for plenty of purposes by laypeople – a practical (not scientific) revolution here to stay, and we will thoroughly discuss it e.g. in the upcoming new AI2 course.

With this research report, we issue the second annual account of our work. This year, we add noteworthy news items to our scientific publications, and introduce you to our associated faculty members. Organized by research group, you will find a brief overview of the group's development over the year 2022, followed by the full text of the published papers of our staff, in the order of their appearance. I wish you, dear reader, an insightful reading.

Winterthur, Winter 2022/2023

Thilo Stadelmann, Director of Centre for Artificial Intelligence

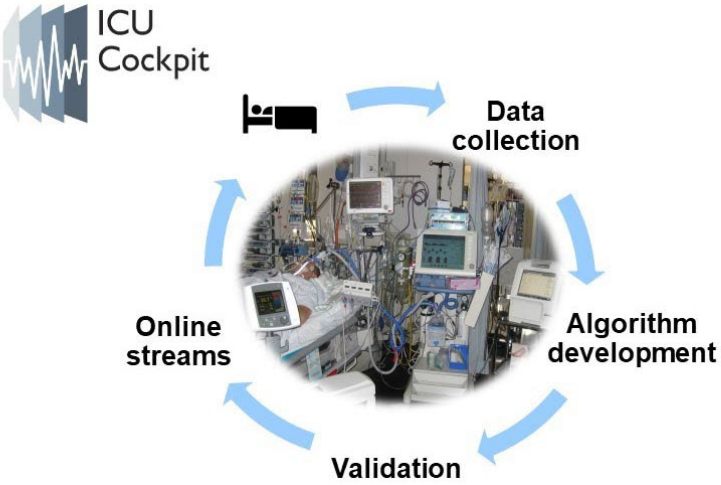
Table of Contents

1 News

Video Analysis for Data-Driven Intensive Care of Patients

07.01.2022 - In a newly-funded DIZH Rapid Action project, CAI and University Hospital Zurich team up to improve automatic data quality monitoring for intensive care interventions.

Monitoring diverse sensor signals of patients in intensive care can be key to detect potentially fatal emergencies. But in order to perform the monitoring automatically, the monitoring system has to know what is currently happening to the patient: if the patient is for example currently being moved by medical staff, this would explain a sudden peak in the heart rate and would thus not be a sign of an emergency.



To create such annotations to the data automatically, the CVPC-Team has teamed up with University Hospital Zurich's Intensive Care Unit (ICU) under the lead of Prof. Emanuela Keller to equip the ICU-Cockpit-Software with video analysis capabilities: based on cameras in the patient room that deliver a constant, privacy-reserving video stream from the patient's bed (i.e., no person can be identified based on the video resolution), the location of patient and medical staff shall be automatically detected and tracked to extract simple movement patterns. Based on these patterns, it shall be classified if and what medical intervention is currently performed on the patient. The research challenge in this project is to realize such a system without access to many labels, i.e., to learn the detection, tracking and classification in mainly un- and self-supervised ways.

Project AUTODIDACT is supposed to start in spring 2022 and deliver results until the year's end. The project is funded by the Digitalization Initiative of the Zurich Universities (DIZH) under the Rapid Action Call in its Innovation Programme. Project lead is USZ's Dr. Gagan Narula.

CAI focuses on autonomous systems as demonstrators for AI in teaching

14.02.2022 - A current bachelor thesis transfers our AI research to four-legged walking robots. These will serve as demonstrators and training objects for teaching.

Deep Learning has revolutionized the way pattern recognition problems like image analysis and natural language processing can be solved today in practice, and is thus permeating industry and society. Yet, the technology behind the success stories remains vague and mysterious to most, to a large degree because it all happens "in the virtual world" (inside a computer). It is less tangible than other technological artifacts involving customer-facing hardware like cars, robots etc.

Claudio Ansaldi now started his bachelor thesis in computer science to build a tangible demonstrator for deep-learning-based computer vision technology like object detection and recognition for the ZHAW Centre for Artificial Intelligence, based on a camera-equipped robotic platform that can move autonomously based on the visual input it receives. This should become the starting point for more use of autonomous systems in AI education at the CAI: It does not just make core AI technology more tangible but offers the possibility to study its side effects when seen in interaction with humans and the physical world, making students aware of the multiple facets of the technology they create. We plan to use several such platforms in future courses in the computer science and data science programmes of the School of Engineering. Additionally, the robotic platforms also offer new opportunities for research into the embodiment of our methods.

Initiative for Large-Scale Development & Testing of Autonomous Systems launched in the Canton of Zurich

14.02.2022 - ZHAW, UZH and ZHdK developed a concept for the safe development and testing of drones and robots together with renowned partners.

The kick-off event organized by the Department of Economic Affairs of the Canton of Zurich last autumn on the topic of autonomous systems in the Zurich Innovation Park lived up to its name: many ideas were pursued after the event and the first results are already visible today. Probably the most concrete result is the successful submission of the project called LINA (A Shared Large-scale Infrastructure for the Development and Safe Testing of Autonomous Systems) unto the structure call of the innovation program of the Digitalization Initiative of the Zurich Higher Education Institutions (DIZH).

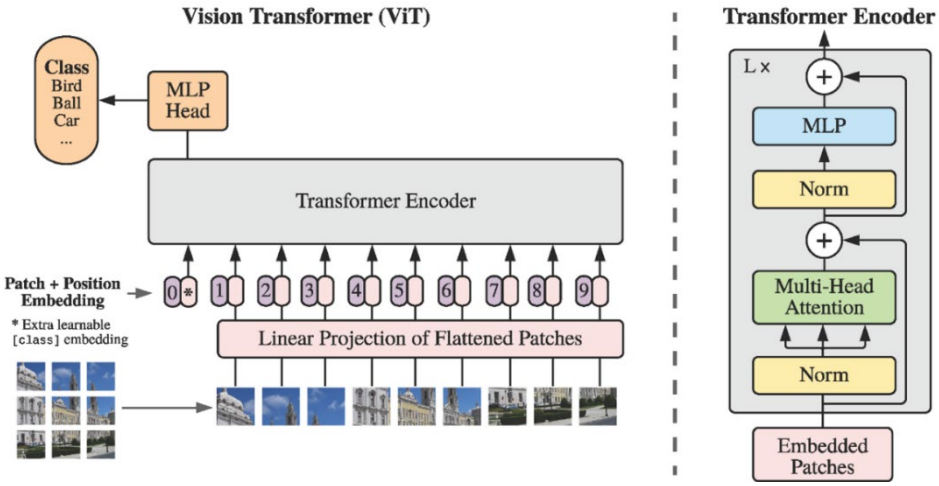
LINA is a comprehensive proposal for the establishment of a test infrastructure for autonomous systems in the Canton of Zurich. The idea to create such a proposal was initiated by the CAI. The elaboration took place in an intensive collaboration between the University of Zurich, the Zurich University of the Arts and the Zurich University of Applied Sciences under the leadership of Michael Guillaume and Peter Lenhart from the ZHAW Centre for Aviation. A total of 21 Letters-of-Intent and 17 Letters-of-Support by renowned companies and organizations from industry, research and administration testify to the usefulness and importance of this proposal.

A decision regarding the launch of LINA is expected in May.

Google Researcher presents Computer Vision Highlights at CAI Colloquium

22.04.2022 - On March 30, 2022, we welcomed a special guest at our CAI colloquium. Dr. Lucas Beyer, senior research engineer at Google Brain, presented the most recent breakthroughs in applying transformer architectures to computer vision.

Deep learning architectures based on so-called Convolutional Neural Networks (CNNs) have proven to be successful in solving computer vision (CV) tasks such as image classification or object detection. On the other hand, recent years have seen breakthroughs in natural language processing (NLP) tasks, e.g. text generation and understanding or automatic translation, by employing so-called Transformers in huge language models such as GPT or BERT.



Such transformer architectures have now also been applied to computer vision, and with great success. This development has been spearheaded by Google research. Therefore, we welcomed Dr. Lucas Beyer, a senior research engineer with Google Brain in Zurich and one of the co-authors of several highly cited research publications in this domain, to our CAI colloquium to give us a deep-dive into the technical details and breakthroughs in the application of transformers in computer vision, e.g. as discussed in the famous ICLR paper “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale” (https://arxiv.org/abs/2010.11929), where the Vision Transformer (ViT) was introduced. He also covered their most recent research, including “How to train your ViT?” (https://arxiv.org/abs/2106.10270), an exploration of its current limits at scale (“Scaling ViT”, https://arxiv.org/abs/2106.04560), and concluded by introducing a recent new alternative to the typical transfer-learning approach, locked image-text tuning (“LiT-tuning”, https://arxiv.org/abs/2111.07991).

The colloquium was thus very relevant to both the NLP and CVPC research groups at CAI, especially in the light of the recent expressed intent of both groups to join forces and work on challenging tasks providing multi-modal input data (combining two or more of text, images, video, audio or tabular data), which could be an ideal application domain of transformer-based architectures. In this context, a recent call for participation has been issued by us via the Databooster platform, calling on potential use case providers to join our collaboration.

Future CAI Colloquia are being planned. The next event will be a panel discussion “Pathways beyond present AI”, introduced by renowned Computer- and Neuroscientist Prof. Dr. Christoph von der Malsburg (Frankfurt Institute of Advanced Sciences, currently guest professor at UZH/ETHZ as well as at ZHAW), and chaired by ZHAW’s Dr. Ricardo Chavarriaga, head of the CLAIRE office Switzerland. The event will take place on Wednesday April 27, 2022.

Panel Discussion "Pathways beyond present AI"

27.04.2022 - Artificial intelligence has made very important and impressive progress in recent years. Yet, in terms of ability to learn from few examples and generalize from them the human (or even animal) brain beats technology by orders of magnitude, as is admitted even by the exponents of that development.

Convinced that the time is ripe for a breakthrough in understanding the nature of intelligence we want to explore this front by organizing, under the title "Pathways beyond Present AI" a short series of interdisciplinary panel discussions that bring together expertise from various fields spanning technology and biology.

The first of these discussions will focus on recent insights into the way the structure of self-organized neural circuits acts as bias tuning the brain to the natural environment.

Format:

1. Introductory presentation by Prof. Dr. Christoph von der Malsburg (FIAS, UZH/ETH, ZHAW)
2. Moderated panel discussion
    - Dr. Ricardo Chavarriaga (Moderator, ZHAW Centre for AI, CLAIRE office CH, AI&Neurotechnologies)
    - Prof. em. Dr. Rodney Douglas (UZH/ETHZ, Neuroinformatics)
    - Prof. Dr. Rudolf Marcel Füchslin (ZHAW, School of Engineering, Applied Complex Systems Science)
    - Prof. Dr. Verena Klamroth-Marganska (ZHAW, School of Health, Occupational Therapy)
    - Prof. Dr. Mike Martin (UZH, Gerontopsychology)
    - Prof. Dr. Thomas Ott (ZHAW, School of Engineering, Computational Life Sciences)
    - Prof. Dr. Thilo Stadelmann (ZHAW Centre for AI, AI and Deep Learning)

You are very welcome to participate and join the discussion, ask questions, and bring in your thoughts and curiosity. This is meant to be an open format in a more informal setting, which should stimulate interdisciplinary discussion and exchange. Of course, even if you just want to listen, you are very welcome as well.

The discussion will be stimulated with a short keynote by Prof. Dr. Christoph von der Malsburg. Christoph von der Malsburg studied physics, with a PhD done at CERN in Geneva. He then reverted to neuroscience and spent 17 years in a Max Planck Institute in Göttingen developing a theory for the wiring of the brain under genetic control and a novel way to interpret neural activity as data structure of the mind, the Dynamic Link Architecture. In 1988 he became professor of computer science at the University of Southern California in Los Angeles, in 1990 in addition professor of systems biophysics at the Institute for Neuroinformatics at Ruhr-University Bochum. Since 2007 he is Senior Fellow at the Frankfurt Institute for Advanced Studies. He has founded two successful companies and has received a number of international awards.

https://www.fias.science/de/fellows/detail/von-der-malsburg-christoph/

https://en.wikipedia.org/wiki/Christoph_von_der_Malsburg

Christoph is currently on a research stay at UZH/ETH, as well as an associate with the ZHAW Centre for AI.

Stanford's Bill Newsome at the CAI

06.05.2022 - As part of the Templeton Lecture Series, Stanford's Prof. Bill Newsome gave a remarkable talk on how free our decisions are and what factors influence our decision making.

Are we free to decide? How do traditional notions of free choice, personal responsibility and religious faith align to each other in terms of neuroscientific understanding of the brain and cognition? As part of the Templeton Lecture Series* and organized by the Christian student campus group (VBG) in cooperation with the Center for Artificial Intelligence (CAI) at the ZHAW, Stanford's Prof. Bill Newsome gave a remarkable talk on the study of decision making and what factors do influence our decision making. From the neurobiological and psychological point of view, he argued that top-down influences, such as goals, beliefs, and aspirations are combined with bottom-up influences, such as individual genetics and conscious experiences to build up the decision-making process.

More specifically, he pointed out that most of our decisions are a combination of biological drives and beliefs together with conscious and subconscious experiences.

Furthermore, Prof. Newsome argued that the knowledge of neurons and synapses is essential for understanding how the brain shapes cognition. But systems-level knowledge, including the level of the unique person, is essential to understand how and why specific behaviours emerge from the interactions of billions of neurons. This multilevel perspective allows us then to think about the relationship about human choice and freedom with the ultimate goal of enhancing human growth and health.

Finally, Prof. Newsome pointed out that having defined moral principles and recognizing situational risk factors, can help us to be the people we most want to be.

* The John Templeton Foundation funds research and catalyses conversations that inspire people with awe and wonder focusing on world affairs and religion. Its vision is to become a global catalyst for discoveries that contribute to human flourishing.

Frank-Peter Schilling appointed Senior Lecturer for Computer Vision and MLOps @ CAI

10.05.2022 - Dr. Frank-Peter Schilling has been appointed as lecturer at the CAI with research focus on computer vision and MLOps, effective June 1, 2022.

Frank-Peter Schilling received his PhD in physics with a focus on elementary particle physics from the University of Heidelberg. He then spent many years in research, including as a senior scientist and project leader at the European research centre CERN in Geneva. There, he was involved in the spectacular discovery of the Higgs particle in 2012. His intensive research activity is reflected in a long list of publications (h-index 150).

He joined ZHAW in 2018, first at the Institute for Applied Informatics InIT and since its foundation in 2021 at the Centre for Artificial Intelligence CAI. His research focus lies on artificial intelligence and machine learning, especially in the area of computer vision, including applications in medical imaging. In addition, with his group he will represent the topic of MLOps (machine learning operations), as well as coordinate all continuing education activities of the centre and represent AI in teaching. University-wide, Frank-Peter Schilling is involved in the ZHAW Digital initiative, in the ZHAW Datalab, as well as in the promotion of doctoral students.

Interdisciplinary panel discussion at the CAI: What is intelligence and how to advance present AI

24.05.2022 - Artificial intelligence (AI) has made impressive progress in recent years. Yet, its ability to learn from few examples and generalize pales compared to human intelligence. An interdisciplinary expert panel at CAI discussed what is intelligence and how can we improve AI from various perspectives including technology, neuroscience, and biology.

Artificial intelligence (AI) has made very important and impressive progress in recent years. Yet, in terms of ability to learn from few examples and generalize from them the human (or even animal) brain outperforms AI by orders of magnitude. A crucial way to reduce this gap is to better understand the actual nature of intelligence. That is the goal of the interdisciplinary panels "Pathways beyond Present AI", organized by the ZHAW Center for Artificial Intelligence (CAI).

The first one of these panels took place within the frame of the CAI colloquia on April 27th. It focused on recent insights into the way the structure of self-organized neural circuits acts as bias tuning the brain to the natural environment.

The event started by a brief on this topic by Prof. Dr. Christoph von der Malsburg (Senior Fellow at the Frankfurt Institute for Advanced Studies, and guest professor at UZH/ETH and ZHAW). His hypothesis, developed together with Prof Dr. Thilo Stadelmann (ZHAW CAI) and Prof Dr. Benjamin Grewe (UZH/ETHZ) in this paper, states that the structures of brain and natural environment are closely related. In particular, the structural regularity of the brain takes the form of net fragments (self-organized network patterns) and that these serve as the powerful inductive bias that enables the brain to learn quickly, generalize from few examples and bridge the gap between abstractly defined general goals and concrete situations. Preliminary studies on computer vision applications provide supporting evidence to this hypothesis.

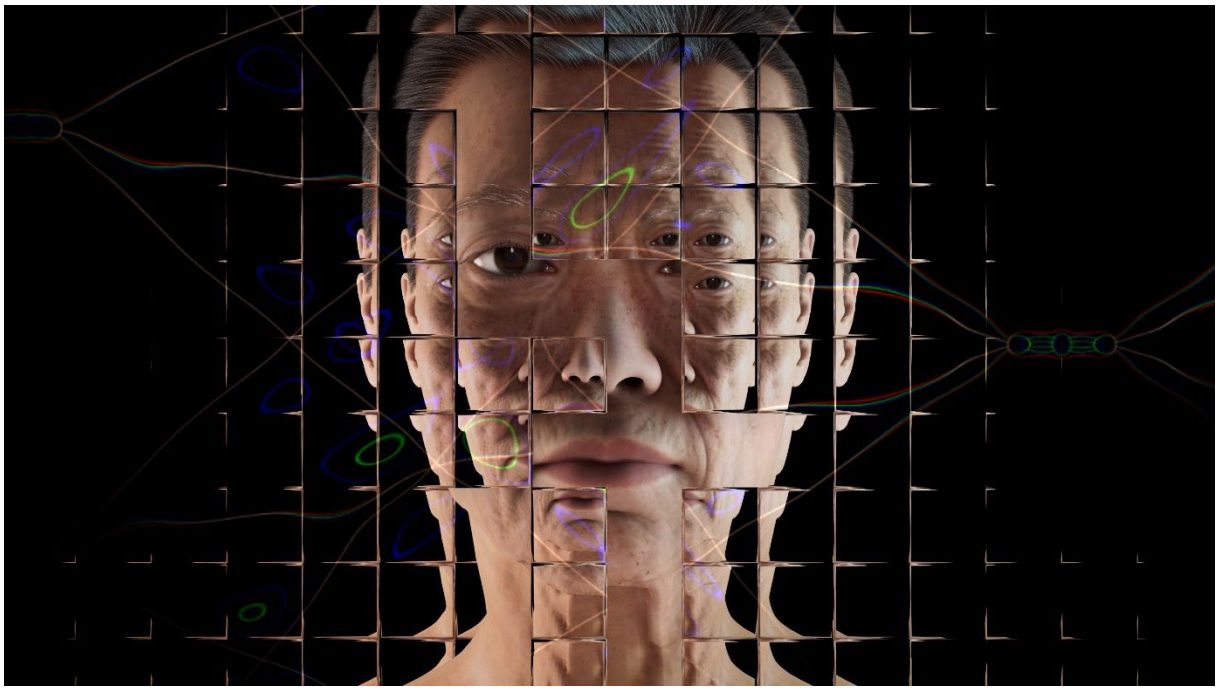
The presentation was followed by a panel where Profs. von Der Malsburg, Stadelmann, and Grewe were joined by experts from various fields spanning technology and biology. The present panellists were Dr. Matthew Cook (UZH/ETHZ, Neuroinformatics, Cortical Computation), Prof. Dr. Rudolf Marcel Füchslin (ZHAW, School of Engineering, Applied Complex Systems Science), Prof. Dr. Mike Martin (UZH, Gerontopsychology), Prof. Dr. Thomas Ott (ZHAW, ZHAW School of Life Sciences and Facility Management, Computational Life Sciences). Prof. Dr. Verena Klamroth-Marganska (ZHAW, School of Health, Occupational Therapy), was also invited as panellists but could not attend due to a last minute contingency. The panel was moderated by ZHAW's Dr. Ricardo Chavarriaga, head of the CLAIRE office Switzerland.

The panel covered multiple topics, starting from the fundamental question of what is intelligence. Besides the ability of generalization mentioned above, panellists mentioned the capacity defining (self-generated) goals as one of the main signs of intelligent. This is complemented by common-sense - achieving semantic meaning of things-and intentionality (that goes beyond simple activities and behaviours) which were all considered as particular characteristics of intelligence that AI has not yet fully achieved.

Another important feature that was mentioned is the ability to establish connections between concrete scenes and these abstract goals, which enable e.g. intentionality. Additionally, the possibility of taking a very complex idea and transform it into a concrete stimulus (e.g. a picture) is an important element for intelligence. This process can be interpreted as inverse problem to processes characterised in the opening lecture. last but not least, it was mentioned that the fact that intelligence develops within a cultural environment should not be neglected.

While addressing which directions AI development should go, the discussion ensued addressing the need for having system that learns while embedded and interacting in an environment, as well as the need of asking the question of what would we like to use the AI for; e.g., Some applications may not require human-like intelligence to be efficiently solved, while others we may prefer to be solved by humans instead of machines. Interestingly, it was pointed out that increase in computing power may be hindering our possibility of creating artificial intelligence as we lose capacity to learn from small datasets.

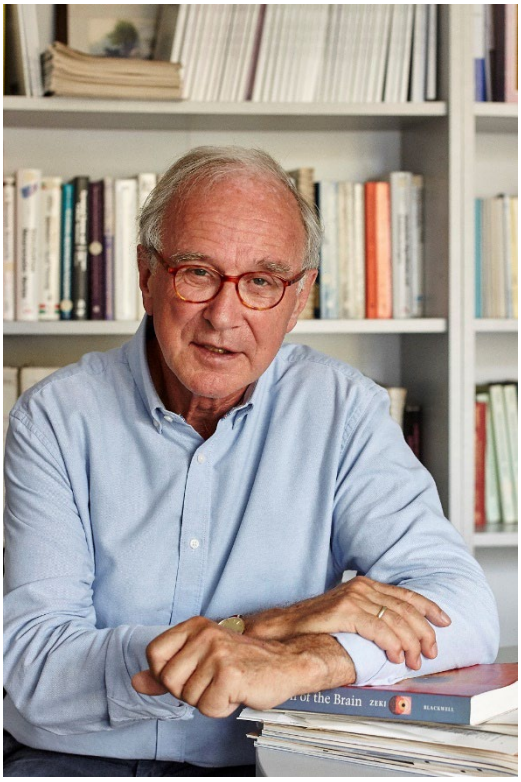
Interaction between panellists and audience delve into different aspects of learning and the role of collective intelligence in the advancement of this field. As well, as the possibility of identifying universal learning rules that, combined with specialised brain-like structures and architectures, can lead us closer to understanding intelligence and more powerful AI. We will continue exploring these topics in future panels in this series and in the CAI Colloquium.



New Theory Explains Intelligence as Emergent Phenomenon

25.05.2022 - Does intelligence amount essentially to the learning and imitation of examples, as the current paradigm of machine learning suggests, or is it dominated by structure that is given a priori? In a new manuscript, a team of ZHAW and ETH/UZH presents convincing arguments for the latter.

To this day the human and animal brain with its incredible ability to pursue vital goals in complex natural environments defies understanding.



In the as of yet unpublished article "A Theory of Natural Intelligence" (Lead author Christoph von der Malsburg; see https://arxiv.org/pdf/2205.00002.pdf) a team of the ZHAW's CAI Group Computer Vision, Perception and Cognition, ETH/UZH's INI and the Frankfurt Institute for Advanced Studies proposes a totally novel explanation for this phenomenon. According to it the structure of the brain arises by emergence and is characterized by strong regularity that mirrors, in a natural way and ahead of all learning, the basic structure of the environment. Implicit in this structure is the capacity to enact genetically encoded abstract behavioural schemata in concrete situations.

Besides putting Epistemology on a new grounding, the paper lays the basis for the artificial emulation of animal and human intelligence, and with that for a technology of autonomous vehicles and robots. First applications of the new theory in the context of machine vision are currently attempted in a master thesis project.

Increased Storage for CAI GPU Infrastructure

31.05.2022 - The CAI is continuously improving its compute infrastructure. After investing in additional computing power in recent years, now the storage system is being renewed. This will allow running machine learning experiments considerably easier and faster.

The Centre for Artificial Intelligence (CAI) at the ZHAW specializes in the areas of Autonomous Learning Systems, Computer Vision, Perception and Cognition, Trustworthy AI, AI Engineering, and Natural Language Processing. All these areas utilize deep learning, a methodology that optimizes complex machine learning systems with enormous compute power and many examples in the form of large data sets. For these computationally intensive optimization problems, graphics processing units (GPUs) are typically used because they can process many calculations in parallel. In addition to GPUs, CPUs are needed to load the data from the file system, as well as memory (RAM) to temporarily store the large amounts of data on the system.

For this purpose, the CAI, together with the InIT, maintains a state-of-the-art infrastructure that is constantly being renewed and optimized. Currently, staff researchers have access to several systems with a total of 124 GPUs, 2648 CPU cores and 21.5 TB of RAM. After a massive increase in compute power in recent years, the storage system has now been renewed. This improves the management of deep learning training data sets that typically consist of millions of tiny files such as text snippets, images or short video sequences. To this end, a new storage system from DALCO that offers 216 TB HDD and 15 TB SSD storage replaces the existing system. The new storage server is connected to the compute cluster with two redundant 100 GB network connections and thus allows lightning-fast data access.

With this new storage server, the CAI is equipped to train deep learning models even more efficiently and thus accelerates research and industrial applications.

CAI launches “Live Impact Series” of events for personal growth of students

03.06.2022 - With a first afternoon on entrepreneurship, the CAI started a series of events on personal growth for its students and staff. It took off on May 4 at CAI premises with two invited speakers: Maurice Gonzenbach, successful founder of Caplena, and Matthias Rosenthal from the School of Engineering’s entrepreneurial initiative.

The ZHAW Centre for Artificial Intelligence (CAI) has started a series of events to further the personal growth of its students (and, by extension: its staff). Its purpose is to help everyone realize their potential and have respective impact in life. To this end, the “Life Impact Series” will offer workshops and tutorials to improve non-technical skills and develop one’s career, e.g., by focusing on topics such as self-management, leadership, personality etc. The first event now focused on the topic of creating socio-economic impact as an entrepreneur: what it’s like to found a company, what it takes to be an entrepreneur, and where to find support.

The first presentation was given by Maurice Gonzenbach, co-founder and machine learning engineer at Caplena, a Swiss company for AI-powered user feedback analysis. The company was founded almost 5 years ago and is now successfully established in the market, with large and important customers worldwide. Maurice gave valuable insights into his personal lessons learned and what to focus on when joining the endeavour of an own company.

The second speaker was Matthias Rosenthal, co-founder and former CTO of sonic emotion ag and now one of the leaders behind the entrepreneurship initiative at the ZHAW School of Engineering. He explained the roadmap of the initiative and how it is embedded into the existing landscape of funding and supportive instruments for young entrepreneurs.

The Life Impact Series will continue with a roughly quarterly schedule. Specific events might be open also to external participants. The next event will take place on July 05 (later afternoon / evening) and shed light on the topic of personality types and their impact on one’s work and life.

CAI successfully co-organised the SwissText 2022 Conference in Lugano

22.06.2022 - The 7th Swiss Text Analytics Conference (SwissText) took place from June 8 to 10. More than one hundred participants from industry and academia exchanged on new and exciting developments in Natural Language Processing (NLP).

SwissText 2022 was organised jointly by SUPSI, SwissNLP and the ZHAW Centre for Artificial Intelligence (CAI). The annual conference is a forum for researchers and practitioners in NLP to meet and discuss. After two online editions, it was finally held again as a physical conference at the SUPSI East Campus in Lugano.

The pre-conference day included interactive workshops on keyword extraction from scientific documents, Swiss German (Speech-to-Text and lexical normalisation), and NLP for Insurance, plus a co-located event on a programme for upskilling linguists for technical professions.

During the two main conference days, there were a total of 18 presentations organised into thematic tracks such as “Speech-to-Text and Swiss German”, “Legal Applications”, “Generation and Parsing” and “Business Applications”. The newly created Junior Track, designed to provide a platform to young researchers, featured six insightful presentations.

Two interactive events made sure that participants were able to further showcase their work and provided valuable networking opportunities:
first, the exhibition on Thursday afternoon featured 21 research posters, 4 system demonstrations, as well as 11 booths by the conference sponsors and affiliated academic institutions.

Second, in the “Battle of NLP Ideas” on Friday, participants discussed in small groups to come up with ideas for new NLP projects. In subsequent rounds, groups were merged in a pyramid fashion and selected the most promising suggestions before presenting them in a plenary meeting where the audience could award votes. The three winning ideas were on synthetic data generation, identifying bot-generated content and data anonymization for NLP. All participants were then able to sign up for those ideas where they are interested in a follow-up meeting.

The rich and varied programme was complemented by three keynotes by renowned experts:

Raul Rodriguez-Esteban**,** Senior Principal Scientist at Roche, discussed Quantitative Social Media Listening, a relatively new trend in health care where social media are used to identify not yet documented symptoms of a disease or populations with unmet medical needs, or find the best location for a clinical trial.

Google's Enrique Alfonseca introduced their ongoing work on integrating structured knowledge into large language models. This is important in the context of enhancing the reasoning capabilities of these models and creating more factually correct responses.

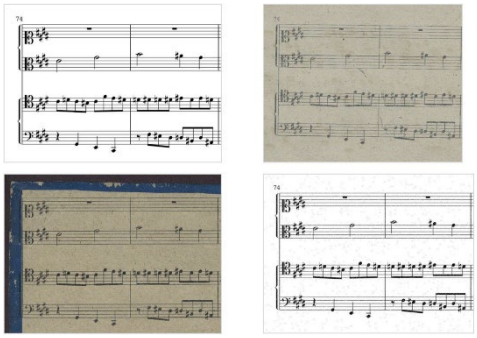
Marco Passarotti, professor at the Catholic University of Milan, talked about the benefits of Linked Data Interoperability when creating language resources, which he illustrated with his ongoing ERC project «LiLa: Linking Latin».

Overall, it was a very successful conference and the beautiful Ticino weather was a big bonus! The slides and recordings of the talks will be made available on the conference website in the coming weeks. SwissText 2023 will continue the journey across Switzerland's language regions: it will be held at Haute École Arc Ingénierie (HE-ARC) in Neuchâtel.

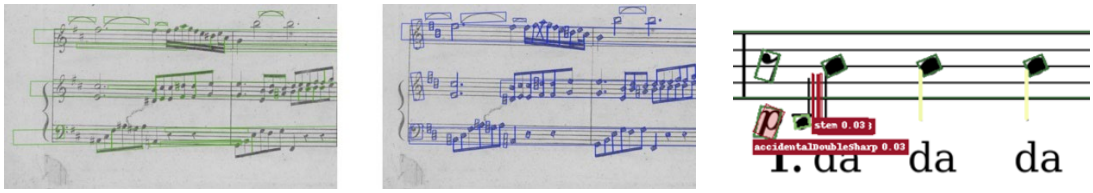
Novel methods for Domain Adaptation and Confidence-Rated Predictions enable Digitalization of Real-World Sheet Music

24.06.2022 - CAI researchers make smart-phone-based digitalization of real-world musical scores possible by equipping the world's most advanced optical music recognition system with a novel domain adaptation mechanism based on deep neural nets and confidence rated output for a sleek human user interface.

ScorePad AG developed a music digitalization pipeline with a variety of applications, ranging from digitalization services for large-scale preservers of score collections like libraries to feeding a user-facing app with fresh and individual content that can be used by music students and professionals alike to practice and perform music alone and in groups. Specifically, it sets free from handling printed scores by displaying and manipulating the scores in computer-readable format (MusicXML) on a tablet or computer.



This computer-readability of digitized music, as opposed to just displaying scanned images, enables novel and highly demanded features for the use cases described above, like ensemble coordination or automatic page turning for orchestra musicians, or enabling music analytics for scholarly users of digital sheet music collections in libraries. It builds on real digitization of the scores through the world's most advanced optical music recognition (OMR) system, whose foundation has been laid during the predecessor CTI project "DeepScore" that outperformed the state of the art in musical symbol recognition by a large margin.



Goal of the RealScore project has been to enable the music digitalization pipeline by extending the use of the predecessor technology, which has been confined to high-quality (synthetic) musical scores as input, to real-world scans of sheets that may have lingered in the musician's gig bag for an extended period of time and have seen many rehearsals. Dealing with such artifacts like yellowed pages, stains and tears requires breakthroughs in applied R&D for symbol recognition (to better detect less frequent musical symbols, the technology needs to be extended to detect dynamically-shaped symbols like slurs at arbitrary rotated angles), domain adaptation (from perfectly produced score PDFs to messy scans or photos) and confidence rating (to mark a potentially non-perfect recognition result with specific colours to indicate where the system is likely to be right and wrong with its detections, according to neural network outputs). These ambitious goals could be achieved by a team of researchers around technical project lead Lukas Tuggener within Prof. Thilo Stadelmann's Computer Vision, Perception and Cognition Group.

The results of project RealScore are two-fold: (i) Transitioning the pipeline to a S^2A-Net-based system with rotated detection capabilities and designing an array of domain adaptation techniques based on (i.a) advanced input data augmentation (“ScoreAug”, see Fig. 2) that combine artificial data degradation with real world wear and tear, (i.b) specific neural network training regimes and (i.c) an adversarial domain adaptation algorithm (see Fig. 1), together improving the music symbol recognition (MOR) on real-world noisy data by more than 50%. (ii) Confidence-rated output (see Fig. 3) has been achieved by adapting Snapshot Ensembles successfully to the S^2A-Net architecture for the first time in an efficient manner, improving the average precision of the MOR task by 4.6 pp and speeding up subsequent manual post-processing of results by a factor of 3 through a through a user tailored and optimized digitalization toolchain.

The training data has been released as an open research data resource. The final models are in productive use as ScorePad AG, Erlenbach, Switzerland.

CAI appoints four new associated faculty members

11.07.2022 - Early July, four well-known ZHAW researchers from several departments and with diverse scientific backgrounds related to AI have been appointed as new associated faculty members at the CAI.

The newly appointed CAI associates will be integrated into the life of the AI centre, engage in common research projects with CAI members and internal or external partners, contribute to teaching and student supervision, and help establishing new or intensifying existing research areas in AI. The associates have an excellent research track record and demonstrated their qualification through a thorough selection process.

The following four ZHAW researchers have been appointed for a three-year period:

Dr. Elena Gavagnin (Senior Lecturer, ZHAW School of Management and Law, Institute of Business Information Technology IWI): Elena brings a strong interest in astrophysical applications of deep learning, where a common ZHAW project embedded in the Swiss consortium within the Square Kilometre Array Observatory has just started. Her research interests extend into other applications of computer vision as well as AI for good and AI for society.

Prof. Dr. Rudolf Marcel Füchslin (Professor for Complex Systems Science, ZHAW School of Engineering, Institute of Applied Mathematics and Physics IAMP): Ruedi has a long-standing research track record in AI, e.g., with evolutionary and morphological methods. On top of existing joint research projects in medical imaging, he is interested in physics-informed machine learning, ethics and philosophy of AI, and dynamical processes.

Dr. Manuel Doemer (Senior Lecturer, ZHAW School of Engineering; Programme Director, BSc. Data Science; Head, ZHAW Datalab): Manuel is interested in computer vision as applied to earth observation (satellite) data, multimodal AI, as well as MLOps. His background as a senior data scientist in industry makes him also proficient in natural language processing.

Prof. Dr. Christoph Heitz (Professor, ZHAW School of Engineering, Institute of Data Analysis and Process Design IDP): Christoph has a background in operations research and predictive maintenance. He is president of the largest Swiss innovation network, the data innovation alliance, where he also leads the Data Ethics group. His research interest is in algorithmic fairness, a topic that he also drives forward in joint teaching activities with the CAI.

New projects aim at enabling the practical adoption of trustworthy AI

14.07.2022 - Three new projects at the CAI address key challenges for practical adoption of reliable AI, comprising quality control, development, testing, and certification.

Three newly awarded projects reinforce the mission of the ZHAW Center for Artificial Intelligence (CAI) to enable the practical adoption of Artificial Intelligence (AI). These projects address key aspects of safe and reliable AI deployment comprising quality control, development and testing and certification of AI systems.

In the first project, termed **DISTRAL** (“Industrial Process Monitoring for Injection Molding with Distributed Transfer Learning”) and funded by Innosuisse, ZHAW CAI partners with Kistler Group and ZHAW Institute of Embedded Systems (InES) to develop a distributed machine learning system to sort out defect plastic parts during production. This project will develop a solution based on distributed machine learning that saves costs, improves usability, and improves production quality. The project entails specific research in transfer learning and federated learning: Using a novel data-centric development process for deep neural networks, it will achieve a semantic transfer of process knowledge that goes far beyond the current state of the art. The resulting model will be able to run on edge devices as well as in the cloud.

The second project, **LINA** ("Shared Large-scale Infrastructure for the Development and Safe Testing of Autonomous Systems"), will build the largest European infrastructure for research, development, and safe testing of autonomous systems such as drones or service robots. This infrastructure, to be established in the Kanton of Zurich, will comprise a large-scale indoor flight-testing arena, an outdoor physical cage, as well as an outdoor digital cage. Together, these infrastructures will cater to the needs of different stakeholders in the autonomous systems space, providing facilities for research, development and testing from technology readiness levels (TRL) 1, observing basic principles, to TRL 9, prove of actual system in operational environment. This project, funded by the DIZH Innovation program, will be developed in collaboration with the University of Zurich, the ZHAW Center for Aviation (ZAV), and the Zurich University of the Arts (ZHdK) with the support of more than 35 practice partners.

Finally, upcoming regulations will require certain types of AI systems to be certified. However, certification bodies currently lack means that allow them to evaluate all aspects of an AI system, including dimensions such as autonomy and control, transparency, reliability, and safety. In the Innosuisse-funded **certAInty** project, the CAI addresses this gap by developing a comprehensive framework for evaluation of AI systems comprising the processes for its development and operation (e.g., document management, change management process), as well as the requirements, technical criteria, measures, and actions directly related to the product. As major innovation, our framework will include technical methods, developed in this project, for verifying relevant properties of the system (such as data management, model validation, verification and explainability). Our partners in this project are the ZHAW Institute of Applied Mathematics and Physics (IAMP) and CertX, the first Swiss certification body for functional safety and cyber security of industrial systems. Outcomes of this project will strengthen the recently launched AI certification program (CertAI) launched by CertX, Fraunhofer IAS and MunichRE.

Through these endeavours, CAI addresses key challenges in the development and deployment of reliable, trustworthy AI-powered systems. It leverages its expertise in applied science to advance the state-of-the-art in AI by developing novel methods and infrastructure for improving the reliability of industrial processes and for the certification and validation of AI systems. Altogether, these projects will benefit CAI partners and society by enabling and accelerating the deployment of more efficient, trustworthy products and systems.

Panel Discussion "Pathways beyond present AI" at CAI on July 20

20.07.2022 - On April 27, the first interdisciplinary panel discussion with Prof. Christoph von der Malsburg in the mini series "Pathways beyond Present AI", organized within the frame of the CAI colloquium, took place. We would like to deepen some of the ideas developed in this first discussion in a follow-up event and 2nd panel discussion:

Date & Time: Wednesday, 20.7., 11:00-12:30

Place: ZHAW Winterthur, Technikumstrasse 71, Room TS O1.19 and Zoom

Title: Artificial Intelligence: Game Over?

Abstract:

Machine Learning with its avalanche of sweeping success stories (Go Zero, BERT, GPT-3, DALL-E, LaMDA, Gato) is broadcasting the impression that the problem of intelligence is solved (see, e.g. here or here) and all that is left is to scale systems. As a result, academia with all its accumulated wisdom and current research results seems to be pathetically left behind, its role reduced to preparing students for earning humongous salaries at Google and their ilk. Is that the end of the story? Is intelligence just an issue of learning from massive amounts of human-generated samples? Even some of the ML gurus send signals that essential ingredients seem to be missing (e.g., here). What could those missing ingredients be? This question that will be addressed in our panel discussion.

Panelists:

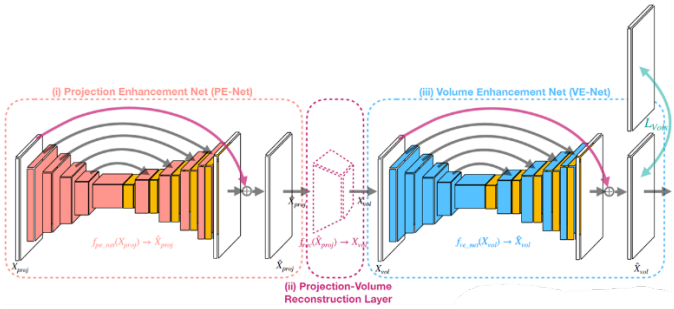
- Prof. Dr. Christoph von der Malsburg (FIAS, UZH/ETH, ZHAW; Neuroinformatics)
- Prof. Dr. Benjamin Grewe (UZH/ETZH, Neuroinformatics)
- Dr. Yulia Sandamirskaya (Neuromorphic Computing, Intel Labs, Munich, Germany)
- Prof. em. Dr. Rodney Douglas (UZH/ETHZ, Neuroinformatics)
- Prof. Dr. Thilo Stadelmann (ZHAW Centre for AI, AI and Deep Learning)
- Moderator: Dr. Ricardo Chavarriaga (Moderator, ZHAW Centre for AI, CLAIRE office CH, AI&Neurotechnologies)

See here a news article on the first event in this series: https://www.zhaw.ch/en/about-us/news/news-releases/news-detail/event-news/interdisziplinaere-diskussionsrunde-am-cai-kolloquium-eroertert-was-intelligenz-ist-und-wie-man-die-ki-der-gegenwart-voranbringen-kann/

See also: https://www.zhaw.ch/en/engineering/institutes-centres/cai/colloquium/

Improving the quality of CT images with AI and Deep Learning

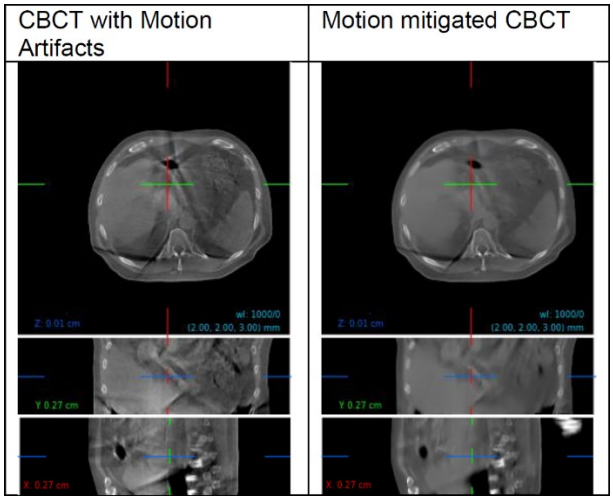
25.08.2022 - In the recently concluded Innosuisse project DIR3CT, researchers from ZHAW's Centre for AI (CAI) could improve the quality of CT images with the help of AI and Deep Learning. The results of the project, which was carried out in collaboration with Varian Medical Systems, as well as with ZHAW's Institute for Applied Mathematics and Physics (IAMP), could improve radiation therapy applied to patients with cancer.



Cone beam computed tomography (CBCT) is widely used in clinical radiation therapy to quickly acquire volumetric (3D) images of the patient's anatomy during treatment, e.g. for the correct positioning of the patient. On-board CBCT devices suffer however from reduced image quality compared with diagnostic CT scans, as well as from artefacts induced by patient motion (e.g., due to breathing, heartbeat, muscle relaxation or digestion).

Varian Medical Systems (now a Siemens Healthineers company), a world market leader in radiation therapy, teamed up with researchers from two ZHAW institutes, the Centre for AI (CAI) and the Institute for Applied Mathematics and Physics (IAMP), in order to reduce the motion artefacts and thus improve the CBCT image quality with Artificial Intelligence (AI) and Deep Learning.

In the project DIR3CT (funded by Innosuisse), a motion mitigation solution was developed, in which deep Neural Networks are trained to correct the artifacts. The developed "dual-domain" approach can significantly improve CBCT image quality, while operating on both the 2D X-ray projections, as well as the reconstructed 3D CBCT image. This allows end-to-end training of the model **(see left figure)**, embedded within the CBCT image reconstruction. An example of the original and motion mitigated CBCT is shown in the **right figure**.



The quantitative evaluation, as well as clinical experts, confirmed the superiority of the motion mitigated CBCT images over the original images. A clinical study is currently underway.

In addition, time resolved 4D-CBCT images were investigated, which require physically plausible concepts to model the anatomical motion. A motion model could be learned, which can use external conditioning data (e.g., a breathing signal) to predict the changes (so-called displacement vector fields) between different motion states in a single forward pass.

In summary, the project could demonstrate that motion mitigated CBCT images are valuable in the clinical workflow, also for advanced applications such as adaptive radiation therapy. The results are being finalized for submission to a peer-reviewed scientific journal and were presented as a poster at the recent AAPM (American Association for Physicists in Medicine) scientific conference in Washington DC.

The research collaboration between Varian and ZHAW continues in the recently started follow-up project AC3T, now also including additional project partners from South Korea.

Game Over? Experts panel at CAI discusses whether the current approach to AI is a dead end

31.08.2022 - Experts discussed whether we are at the verge of achieving artificial general intelligence in the second panel of the series "Pathways beyond present AI". They provided Insights on the differences between natural and machine intelligence, and research lines that can help us understand the former and improve the latter.

"The game is over" claimed triumphally a recent tweet announcing the release of "Gato", a generalist AI approach aimed at performing hundreds of different tasks. According to the announcement, artificial general intelligence (AGI) was now at reach and the key to achieve it laid on continuing to scale up current systems: i.e., using larger datasets and networks. Similar claims are now more and more common, based on sweeping success stories in the last years (Go Zero, BERT, GPT-3, DALL-E, LaMDA, Stable Diffusion). They purport the vision that access to a massive amounts of data and computing power is the direct way to achieve truly intelligent systems, with little role for further theoretical or formal research on the topic.

Not everybody shares this vision. Voices from academia and industry argue that further foundational research on artificial intelligence is needed (see here and here). The second instalment of the ZHAW Center for Artificial Intelligence series of interdisciplinary panels on "Pathways beyond Present AI" focused on this topic. Panellists discussed whether the game was actually over and if not, what could be the missing elements and the roadmap towards AGI.

The panel took place on July 20th 2022, gathering the following panellists: Prof. Dr. Christoph von der Malsburg (Senior Fellow at the Frankfurt Institute for Advanced Studies, and guest professor at UZH/ETH and ZHAW), Prof. em. Dr. Rodney Douglas(UZH/ETHZ, Neuroinformatics), Prof Dr. Thilo Stadelmann (ZHAW CAI), Prof. Dr. Rico Sennrich (Natural Language Processing, UZH), Dr. Yulia Sandamirskaya (Neuromorphic Computing, Intel Labs, Munich, Germany), and Prof Dr. Benjamin Grewe (UZH/ETHZ). The panel was moderated by ZHAW's Dr. Ricardo Chavarriaga, head of the CLAIRE office Switzerland.

Prof von der Malsburg pointed to the differences between machine and natural intelligence, noting that the learning environment of AI is composed by streams of human-provided samples while natural intelligence relies on self-driven exploration. Additionally, natural intelligence serves the pursuit of innate goals. This last point was supported by Prof. Grewe. He argued that a characteristic of intelligence is the ability of self-generating goals, a vision that was discussed in the first panel of this series.

Panellists also addressed ways towards AGI. Prof Stadelmann argued that achieving AGI may not intrinsically need to be related to understanding natural intelligence. As pointed out by Prof. Sennrich, we may not need self-awareness in systems devoted to solving some specific tasks. Nonetheless, panellists agreed that cross-fertilization of these fields can significantly benefit each other. Dr. Sandamirskaya reminded the importance of research and stated that the field may benefit from supporting not only large-scale "flagship" projects but numerous smaller projects allowing to diversify the objects of study and try multiple approaches. She proposed as an alternative to generalist models, the idea of starting from simple models, like insect brains, and evolve to be able to solve more complex tasks.

Regarding the differences between machine and natural intelligence, Prof Douglas challenged the common conception that the initial states of AI systems is "tabula rasa", clarifying that AI systems already encode some knowledge by the choice of architectures, and computational machinery. Interestingly, the role of the network architecture and learning biases is an important element of the "Theory of Natural Intelligence" advanced by Profs von der Malsburg, Stadelmann and Grewe. However, how does this theory maps into modern deep learning architectures is still an open question.

Besides the architectures, panellists also mentioned the importance of the learning rules. They stated that AI's reliance on backpropagation enforces consistency between the network output and the teacher's expected values. In contrast, bio-inspired self-organizing networks can potentially promote consistency in all levels, across connections and sensory inputs. Consistency over time and modalities was deemed of particular importance, highlighting the need for AI systems to interact with open environments, instead of treating their inputs as independent samples.

The depth and richness of the panel makes it clear that the game is far from being over and there exist yet multiple gaps that need further research to achieve AGI (or any considerable advance in AI systems towards “common sense”). The consensus was that only scaling up current systems is not likely to be the solution. The discussion led to several research directions that could provide insights into how to improve current artificially intelligent systems and get a better grasp on the principles that rule natural intelligence.

More specific aspects of these directions will continue to be discussed in future panels. Please be attentive to the CAI colloquium webpage and our news page for further announcements on the panel series: "Pathways beyond Present AI".

Rieter and the Johann Jacob Rieter Foundation Sponsor Professorship for Artificial Intelligence at the ZHAW

06.10.2022 - The Rieter Group is constantly expanding its technology leadership. Together with the Johann Jacob Rieter Foundation, the company is therefore supporting a new Endowed Professorship for Industrial Artificial Intelligence (AI) at the ZHAW School of Engineering. The Professorship is dedicated to teaching and research in the field of industrial applications of Artificial Intelligence and will be announced later this year.

The new Endowed Professorship will be established at the Center for Artificial Intelligence (CAI) of the ZHAW in Winterthur. It will focus, in particular, on the application of machine learning methods and knowledge-based systems in connection with processes in production and service. “The use of artificial intelligence in industry is becoming increasingly important, especially with regard to the potential of data for evaluation and control of complex processes. The support of the Johann Jacob Rieter Foundation and the Rieter Group will allow us to further expand AI research in the field of industrial applications,” explains Prof. Dr. Dirk Wilhelm, Director of the ZHAW School of Engineering.

For Rieter, the commitment is related to the implementation of its technology leadership strategy. “The use of Artificial Intelligence will make a significant contribution to automation and process optimization, and thereby advance sustainability in the textile industry. This makes it an important element of the leading technology that Rieter offers,” emphasizes Rieter CEO Dr. Norbert Klapper.

The contribution of the Johann Jacob Rieter Foundation to sponsoring the Professorship is in line with the Winterthur Cluster Initiative. The increasing digitalization of production processes opens up new perspectives for Winterthur as a business location. “The Smart Machines cluster is growing in importance,” says Thomas Anwander, member of the Foundation Board, and adds: “The Endowed Professorship for Industrial AI at the ZHAW aims to promote Winterthur as a technology location by pooling locally available strengths in mechanical engineering and Industry 4.0.”

Building expertise in the field of Industrial AI

The Endowed Professorship will serve to build expertise in the field of Industrial AI and will oversee a group that will focus on teaching and research pertaining to trustworthy machine learning. This involves, for example, the deployment of artificial intelligence with the aim of optimizing production processes in relation to the use of raw materials and energy, and making expert knowledge more readily available.

In addition to research, for the purpose of knowledge transfer, the new professorship will also be active in teaching, in the bachelor's degree programs in Computer Science and in Data Science, in the Master of Science in Engineering, and in continuing education.

The annual commitment of CHF 300 000 over a period of six years will be financed equally by the Rieter Group and the Johann Jacob Rieter Foundation.

Frank-Peter Schilling appointed as an adjunct professor at Victoria University of Wellington

21.11.2022 - As part of the collaboration between Victoria University of Wellington and the ZHAW School of Engineering, Frank-Peter Schilling from the Centre for Artificial Intelligence (CAI) has been appointed as an adjunct professor. In this role, the ZHAW lecturer will, among other things, work to establish a joint PhD programme between the two universities.

The agreement signed in February this year between the ZHAW School of Engineering and Te Herenga Waka – Victoria University of Wellington provides for collaboration in the fields of data science and artificial intelligence (AI). In addition to a joint PhD programme, cooperation on research projects and the mutual exchange of students, the partnership is aiming to establish bridge professorships. The two universities are now taking a major step towards achieving this objective by appointing ZHAW lecturer Dr Frank-Peter Schilling as an adjunct professor. The Head of the Intelligent Vision Systems Group at the Centre for Artificial Intelligence (CAI) conducts research in the area of artificial intelligence and deep learning and is also Academic Coordinator of the PhD Programme in Data Science, which exists as a collaboration between the ZHAW and the University of Zurich.

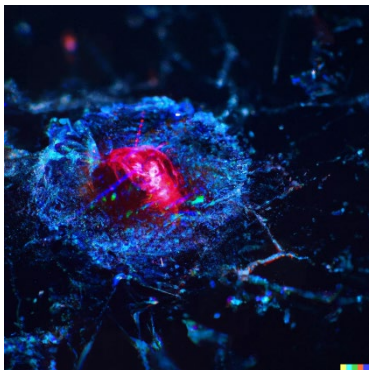
Frank-Peter Schilling was enthusiastic upon receiving news of his appointment as an adjunct professor: “I am looking forward to the academic collaboration with Victoria University of Wellington and am motivated to advance the joint exchange in the areas of research and teaching.” As an adjunct professor, Frank-Peter Schilling will in future be involved in joint research projects and work to establish a PhD programme between the New Zealand university and the ZHAW School of Engineering.

New European research project for automated detection of disinformation campaigns in social media

21.11.2022 - A new ChistERA Project tackles the problem of detecting organized intentional misinformation campaigns in social media. This project is a collaboration with international research partners from Spain, Estonia, and France.

The NLP group at the ZHAW Center for Artificial Intelligence (CAI) has acquired a new ChistERA project called **HAMiSoN**. The goal of the project is to tackle the socially crucial problem of detecting organized intentional misinformation campaigns.

Misinformation campaigns are not only limited to single instances of fake news. They are to be seen in a more holistic way: one must take into account the agents that introduce the misinformation, the supporting media that propagates it, and the social network dynamics, which lead to the adoption of this information. Furthermore, misinformation is not only a single instance of stating wrong facts, they are more about spreading narratives with a specific intention or goal. Finally, this problem is not limited to text; with current technologies it has become a multimodal problem including videos, images, and audio alongside text.



In this project, we address these issues by taking all these aspects into account, that is, by taking a holistic approach to the problem detecting misinformation campaigns. The project will span over a period of three years, and includes partners form the National University of Distance Education in Spain, the University of Tartu in Estonia, and Synapse Developpement in France.

We hope that thought this endeavour, the CAI can contribute to a more trustworthy and just consumption of information.

Thilo Stadelmann receives Fellowship and Impact Award

28.11.2022 - Prof. Stadelmann, head of CAI, receives a DIZH Fellowship 2022 for research excellence and the ZHAW digital shaper award 2022 in the category “impact”

DIZH Fellowships are awarded in a competitive selection process to ambitious ZHAW researchers. The university-wide funding scheme for research excellence sends its fellows to the joint research cluster of the Digitalization Initiative of the Canton of Zurich, DIZH, located at the University of Zurich. DIZH Fellowships last for up to 2 years and can fund a fellow plus a PhD student.

Prof. Stadelmann’s research endeavour for the fellowship builds on top of the CAI panel discussions on pathways beyond present AI and the recently released "Theory of Natural Intelligence" that proposes a possible key to the emergence of intelligence in biological learners. Goal of the fellowship is to develop a technical implementation of the concept of self-organizing net fragments within contemporary deep artificial neural nets. Together with a PhD student, the team will start ca. mid-2023, supported by colleagues from neuroinformatics and theology in the context of techno-ethics and society.

The Impact Award of ZHAW digital recognizes outstanding contributions through projects and initiatives that are impacting the digital transformation at the ZHAW and/or at the societal level. The price comes with an

With this award, Prof. Stadelmann’s impact at the three levels of teaching, research (exemplary, in the DeepScore/RealScore projects) and building the respective environment for such endeavours to thrive (e.g., co-creation of the ZHAW Datalab, data innovation alliance, and the ZHAW Centre for AI) have been recognized.

The award comes with a cash price of CHF 1’000 that will be used to further enhance the dialog and discussion with fellow scientists. In Absence of Thilo Stadelmann, who was on interdisciplinary speaking assignment, Prof. Frank-Peter Schilling accepted the prize and gave a short address on the kind of impact we a re striving for at the CAI.

ZHAW and Mindfire seal partnership for AI

19.12.2022 - The newly signed cooperation involves the ZHAW Center for AI and other ZHAW units as well as Lab42, a globally connected AI Lab operated by Mindfire, which spans joint activities in research, education, start-ups, events, and dialogue with the general public.

Building on joint initiatives since 2019, the Zurich University of Applied Sciences (represented by ZHAW digital and the ZHAW Centre for AI) and the Mindfire Foundation recently signed an agreement to collaborate more closely in the context of Mindfire's Lab42. Cornerstones of this initiative are scientific advice and support for Lab42, joint research and innovation projects, involvement of ZHAW students in Lab42 activities, creation of joint education and training activities, promotion of start-ups and innovation in the field of AI, and dialogue with the general public and specialist audiences on AI in the context of joint events.

The festive setting for the ceremonial sealing of the partnership was provided by the Swiss AI Gala Dinner, which took place on December 15 at the historic Zunfthaus zur Zimmerleuten in Zurich. As part of the program, renowned scientist Jürgen Schmidhuber and DeepMind CEO Demis Hassabis addressed a select audience of AI lab heads, scientists, entrepreneurs, investors and journalists, and the question of what is missing in current AI systems was discussed passionately.

The Swiss AI Gala Dinner 2022 offered the perfect setting for the ceremonial sealing of the cooperation.

Director of CAI Prof. Dr. Thilo Stadelmann and Mindfire president Pascal Kaufmann shake hands after signing the memorandum of understanding.

2 Student Affairs

2 Student Affairs

The ZHAW Centre for Artificial Intelligence offers method-oriented courses in both the BSc and MSc programmes of the ZHAW School of Engineering. In addition to their academic coursework, our students are exposed to applied research in our core topics and get opportunities to also develop their non-technical capabilities and skills to advance their careers.

At the Bachelor level we teach the fundamentals of machine learning, data mining, artificial intelligence, and various specialisations in the focus areas of the CAI. Our goal is to ensure that the latest advances in research and development are incorporated in our teaching. As one of the contributing activities, we offer you the opportunity to carry out one of the limited Research Project Theses in the final year of your studies.

The CAI also offers Master's profiles in Computer Science and Data Science. During their studies, our students are directly incorporated into one of our research groups and personally mentored by a senior researcher in that group. This gives them direct exposure to the latest AI methods at the highest scientific level, and their application to solve challenges within real-world projects in collaboration with practice partners, e.g., industrial, or clinical.

In 2022, 10 Master students were hosted at the CAI while 14 students made their BSc thesis with us. Besides their coursework, they developed practical projects on fields including language modelling, chatbots, machine translation, speaker verification, medical imaging, brain-machine interfacing, computer vision for autonomous cars, and others.

The CAI offers tools for developing technical and non-technical skills.

Besides their academic credentials, Master students at the CAI will also develop other key competences for advancing their career. Through our ***Life Impact*** event series, we offer regular seminars and workshops aimed at inspiring and equipping young researchers with non-technical knowledge and skills that allow them to develop their full potential. Throughout this series we have organized activities on, e.g., entrepreneurship, personality types and leadership skills.

Want to know more?

Student projects at the CAI

BSc Project thesis at the CAI

Why doing a MSc at the CAI?

CAI Students in 2022:

MSc students: Raphael Emberger, Nicola Good, Livia Lüscher, Janick Michot, Sydney Nguyen, Pascal Sager, Samuel Stucki, Manuel Weiss. *Not pictured*: Sebastian Salzmann, and Juan Ribera.
BSc Students: *Not Pictured*: Nathalie Achtnich, Aurora Alitjaha, Nico Ambrosini, Lukas Bamert, Lukas Boner, Benjamin Berli, Manuel Berweger, Michael Häseler, Gian Hellinger, Maurice Hostettler, Besmir Kadrii, Kevin Kläger, Mario Küng, Tenzin Samdrup Langdun, Urban Lutz, Alexandre Manai, Kai Mannhart, Lars Mosimann, Lucian Nicca, Martin Oswald, Benjamin Stern, Marvin Tseng, Patrik Randjelovic, Mika Ruch, Ralph Scheu, Samuel Stucki, Florian Witschi, and Simon Zaugg.

Completed Student Theses 2022

Student work at the CAI typically has a method-oriented focus and contributes its own research in the context of an exciting application. We are very proud of the achievements of our students, who regularly outdo themselves, achieve excellent results, and often contribute to the latest research through their publications, as also shown through various prices. In the following, we briefly present works from 2022 in descending order by date and thesis level.

When	Type	Title	Group
Fall 2022	Master Thesis	Entropy-Aware Active Vision Through Voxelized Octree Exploration of 3D Scenes	CVPC
Fall 2022	Second Master Project	Supporting DARPin binder selection through deep learning	CVPC
Fall 2022	First Master Project	Evaluation of Data Augmentation Strategies for Motor Imagery BCI Classification Tasks Based on EEG Data / top grade, Student: Manuel Weiss	CVPC
Fall 2022	First Master Project	The Practical Impact of Data-Centrism on the Example of Autonomous Driving	CVPC
Fall 2022	First Master Project	Reproducing a large-scale Speaker Verification System	CVPC
Fall 2022	First Master Project	Base System for a Language Learning Chatbot / top grade, Student: Janick Michot	NLP
Fall 2022	Bachelor Project Thesis	Dialekte raten – Erweiterung einer Web-Applikation / top grade, Students: Kai Mannhart, Nathalie Achtnich	NLP
Fall 2022	Bachelor Project Thesis	Digitalization of Chess Score Cards	NLP
Fall 2022	Bachelor Project Thesis	Corpus Evaluation for Automatic Speech Recognition 2.0 (“CEASR 2.0”)	NLP
Fall 2022	Bachelor Project Thesis	Machine Learning-Based Analysis of Data from the ZHAW Movement Analysis Laboratory for Fatigue Detection during Sports Exercises	CVPC
Fall 2022	Bachelor Project Thesis	Speak your mind! Brain Computer Interfaces for Communication	CVPC
Fall 2022	Bachelor Project Thesis	Building a Vision-Based AI Demonstrator with Unitree A1 Quadruped Robot / top grade, Students: Tenzin Samdrup Langdun, Martin Oswald	CVPC
Spring 2022	Master Thesis	Leveraging Neuroscience for Deep Learning Based Object Recognition	CVPC
Spring 2022	First Master Project	Automatic extraction of anthropometric features and body composition parameters from computer tomography images enables improved BMI prediction at scale	CVPC
Spring 2022	Bachelor Project Thesis	Hitting the Jackpot: Optimizing Neural Networks with Composite Pruning Strategies / top grade, Students: Urban Lutz, Alexandre Manai	CVPC
Spring 2022	Bachelor Project Thesis	Improved Speech Translation for Swiss German Using a Hybrid DynamicWindow Approach	NLP

When	Type	Title	Group
Spring 2022	Bachelor Project Thesis	Was hast du gesagt? Gespräche automatisch vereinfachen	NLP
Spring 2022	Bachelor Project Thesis	Auswirkungen von Speech Enhancement auf die automatische Spracherkennung	NLP
Spring 2022	Bachelor Project Thesis	Digitalization of Chess Scorecards	NLP
Spring 2022	Bachelor Project Thesis	Ensemble Methods for Speech Recognition	NLP
Spring 2022	Bachelor Project Thesis	Summarize This! Automated Generation of Meeting Highlights	NLP
Spring 2022	Bachelor Project Thesis	Exploring Wav2Vec2 Pre-Training on Swiss German Dialects Using Speech Translation and Classification	NLP

3 Research Output of the Computer Vision, Perception and Cognition Group

3 Research Output of the Computer Vision, Perception and Cognition Group

The CVPC group, led by Prof. Stadelmann, conducts pattern recognition research on a wide variety of tasks relating to image, audio, and generally signal data. It focuses on deep neural network and reinforcement learning methodology, inspired by biological learning.

Each studied task has its own learning target (e.g., detection, classification, clustering, segmentation, novelty detection, or control), corresponding use case (e.g., predictive maintenance, speaker recognition for multimedia indexing, document analysis, industrial quality control, automated game play or building control), and research challenge (e.g., domain adaptation, practical robustness, or limited data), which in turn sheds light on different aspects of the learning process. This experience is used to create increasingly general AI systems built on neural architectures.

In 2022, the group underwent some natural change with research associate Dr. Javier Montoya being appointed senior lecturer at HSLU, senior researcher Dr. Frank Shilling becoming senior lecturer at ZHAW CAI and adjunct professor at the University of Wellington, and intern Adhiraj Ghosh leaving to pursue further studies at Tübingen University, while doctoral student Mohammadreza Amirian handed in his thesis at Ulm University. With most research projects ending in early summer (e.g., *RealScore* on optical music recognition and *DIR3CT* on CT motion artifact reduction) and many new ones starting from summer on (e.g., the *Mobile Inclusion Lab* on brain-computer-interfaces, *LINA* on the safe development and testing of autonomous systems, *DISTRAL* on transfer learning in industrial anomaly detection, and *Master3D* on engineering sketch understanding), new researchers joined the team, among them Paul Luley and Benjamin Meyer as research assistants, Peng Yan as doctoral student (co-supervised at the UZH/ETH Institute of Neuroinformatics), and two master students.

The group's publications in 2022 reflect the diverse use cases that are being worked on, including some highly interesting side projects: Stadelmann et al.'s work on *data centrism as the foundation of data science* and, together with co-authors von der Malsburg and Grewe, on a "*Theory of Natural Intelligence*", and Dr. Chavarriaga's continued work on *data governance* and *cognitive workload*. Pascal Sager (and colleagues) was able to turn his second master project into a journal publication on *vertebrae detection with semi-supervised domain adaptation*, and co-supervised PhD student Felix Schmitt-Koopmann published his first paper in IEEE Access, together with open research data to build *document analysis systems to help the visually impaired with assistive technology*. Finally, final year PhD student Lukas Tuggener published his brilliant analysis of *ImageNet as a basis to derive CNN architectures* in Frontiers of Computer Science.

We thank our project partners, students, and funding bodies for their generous support and effort, without which these results (and the results forthcoming in future years) would not have been possible!

The CVPC 2022 team

Thilo Stadelmann, Ricardo Chavarriaga, Lukas Tuggener, Mohammadreza Amirian, Peng Yan, Pascal Sager, Raphael Emberger, Paul-Philipp Luley, and Benjamin Meyer (along with three Master students, two external PhD students and three associated faculty members)

Workshops of the Eighth International Brain-Computer Interface Meeting: BCIs: The Next Frontier

Authors (after Dr. Huggins as lead author and Dr. Krusienski as the **wo**rkshop coordinator, authors are listed in reverse alphabetical order.):

corresponding author: Jane E. Huggins

Department of Physical Medicine and Rehabilitation,
Department of Biomedical Engineering,
Neuroscience Graduate Program
University of Michigan, Ann Arbor, Michigan, United States
325 East Eisenhower, Room 3017;
Ann Arbor, Michigan 48108-5744,
734-936-7177,
janeh@umich.edu
ORCID: 0000-0001-8709-4350

Dean Krusienski

Department of Biomedical Engineering
Virginia Commonwealth University
Richmond, VA 23219
djkrusienski@vcu.edu
ORCID: 0000-0002-4668-5784

Mariska J Vansteensel

UMC Utrecht Brain Center
Dept of Neurosurgery
University Medical Center Utrecht
The Netherlands
m.j.vansteensel@umcutrecht.nl
ORCID: 0000-0002-9252-5116

Davide Valeriani

Neurable, Inc
davide.valeriani@gmail.com
ORCID: 0000-0001-9866-0063

Antonia Thelen

eemagine Medical Imaging Solutions GmbH, Berlin, Germany
athelen@ant-neuro.com

ORCID: 0000-0002-3989-5388

Sergey Stavisky

University of California, Davis
sstavisky@ucdavis.edu
ORCID: 0000-0002-5238-0573

James J.S. Norton

National Center for Adaptive Neurotechnologies
US Department of Veterans Affairs
113 Holland Ave
Albany, NY 12208
jamesjsnorton@gmail.com

Anton Nijholt

Faculty EEMCS
University of Twente
Enschede
The Netherlands
a.nijholt@utwente.nl
ORCID: 0000-0002-5669-9290

Gernot Müller-Putz

Institute of Neural Engineering, GrazBCI Lab, Graz University of Technology
Stremayrgasse 16/4
8010 Graz, Austria
gernot.mueller@tugraz.at
ORCID: 0000-0002-0087-3720

Nataliya Kosmyna

Massachusetts Institute of Technology (MIT)
Media Lab
E14-548
Cambridge, MA 02139
Unites States
nkosmyna@mit.edu
ORCID: 0000-0003-1272-0470

Louis Korczowski

Siopi.ai
Grenoble, France
louis.korczowski@gmail.com
ORCID: 0000-0002-7982-5739

Christoph Kapeller

g.tec medical engineering GmbH
Phone +43 7251 22240 20
kapeller@gtec.at
ORCID: 0000-0002-7330-895X

Christian Herff

School of Mental Health and Neuroscience
Maastricht University
Maastricht, The Netherlands
c.herff@maastrichtuniversity.nl
ORCID: 0000-0002-5610-2618

Sebastian Halder

University of Essex
s.halder@essex.ac.uk
ORCID: 0000-0003-1017-3696

Christoph Guger

g.tec medical engineering GmbH/Guger Technologies OG,
Austria, Sierningstrasse 14,
4521 Schiedlberg, Austria,
+43725122240-0,
guger@gtec.at
ORCID: 0000-0001-6468-8500

Moritz Grosse-Wentrup

Research Group Neuroinformatics,
Faculty of Computer Science;
Vienna Cognitive Science Hub;
Data Science @ Uni Vienna University of Vienna
moritz.grosse-wentrup@univie.ac.at
ORCID: 0000-0001-9787-2291

Robert Gaunt

Rehab Neural Engineering Labs,
Department of Physical Medicine and Rehabilitation,
Center for the Neural Basis of Cognition,
University of Pittsburgh, Pittsburgh, PA, USA
3520 5th Ave, Suite 300
Pittsburgh, PA 15213
412-383-1426
ORCID: 0000-0001-6202-5818
rag53@pitt.edu

Aliceson Nicole Dusang

Department of Electrical and Computer Engineering, School of Engineering
Brown University
Carney Institute for Brain Science
Brown University
Providence, RI
Department of Veterans Affairs Medical Center, Center for Neurorestoration and Neurotechnology, Rehabilitation R&D Service,
Providence, RI
Center for Neurotechnology and Neurorecovery, Neurology, Massachusetts General Hospital
Boston, MA
aliceson_dusang@brown.edu
ORCID: 0000-0002-1480-1533

Pierre Clisson

Timeflux Research Group
pierre@clisson.net

Ricardo Chavarriaga

IEEE Standards Association Industry Connections group on neurotechnologies for brain-machine interface
Center for Artificial Intelligence,
School of Engineering,
ZHAW-Zurich University of Applied Sciences, Switzerland, Switzerland
ORCiD: 0000-0002-8879-2860
r_chavarriaga@ieee.org

Charles W. Anderson

Department of Computer Science
Molecular, Cellular and Integrative Neurosience Program
Colorado State University
Fort Collins, CO 80523
chuck.anderson@colostate.edu

Brendan Allison

Dept. of Cognitive Science,
Mail Code 0515,
University of California at San Diego,
La Jolla, United States,
619-534-9754
https://orcid.org/0000-0001-7729-982X.
ballison@ucsd.edu

Tetiana Aksenova

University Grenoble Alpes, CEA, LETI, Clinatec, Grenoble 38000, France
tetiana.aksenova@cea.fr
ORCID: 0000-0003-4007-2343

Erik Aarnoutse

UMC Utrecht Brain Center
Department of Neurology & Neurosurgery
University Medical Center Utrecht
Heidelberglaan 100
3584 CX Utrecht, The Netherlands
e.j.aarnoutse@umcutrecht.nl
ORCID: 0000-0001-7648-250X

Workshops of the Eighth International Brain-Computer Interface Meeting: BCIs: The Next Frontier

Abstract

The Eighth International Brain-Computer Interface (BCI) Meeting was held June 7-9th, 2021 in a virtual format. The conference continued the BCI Meeting series' interactive nature with 21 workshops covering the bread of topics in BCI (also called brain-machine interface) research. Some workshops provided detailed examinations of methods, hardware, or processes. Others focused on BCI applications or user groups. Several workshops continued consensus building efforts designed to create BCI standards and improve comparisons between studies and the potential for meta-analysis and large multi-site clinical trials. Ethical and translational considerations were the primary topic for some workshops or an important secondary consideration. The range of BCI applications continues to expand, with more workshops focusing on approaches that can extend beyond the needs of those with physical impairments. This paper summarizes each workshop, provides background information and references for further study, summarizes discussions, and describes the resulting conclusion, challenges, or initiatives.

Keywords: brain-computer interface; brain-machine interface, neuroprosthetics; conference;

Introduction

The field of brain-computer interface (BCI) research has many names, most historically originating from related research domains with converging objectives. The terms BCI and brain-machine interface (BMI) are quite common and the term neuroprosthetic also applies. In general, a BCI is a device that interprets information directly from the brain to provide a means of interacting with technology. Brain activity can be measured using either implanted electrodes or external sensors. The technology can be operated through a variety of methods, including a direct connection between the brain and the effector (e.g., to operate a prosthetic), or a secondary interface such as a keyboard display (e.g., for communication). Recent work has also used electrical stimulation of the brain itself to "close the loop" and provide sensory feedback about the state of the technology. The defining feature of a BCI is that the brain activity itself is interpreted, the

information to control a device is not derived from activity propagated through peripheral nerves. Many BCIs were initially developed for use by people with physical impairments, but the current broad range of applications also targets other neurological and cognitive impairments, abled-bodied users, and even opportunities for human enhancement. The 8th International Brain-Computer Interface Meeting provided a venue for exploration of the breadth of BCI topics and this paper is designed to provide a window into the workshops that occurred at that Meeting.

The BCI Meeting Series

The 8th International Brain-Computer Interface Meeting was originally scheduled to be held in 2020. However, due to travel restrictions and health concerns during the global pandemic, the 2020 in-person meeting was postponed to June 7-9th, 2021 and ultimately converted to a virtual meeting format. The goal of the BCI Meeting Series (1999 [1] 2002 [2], 2005 [3], 2010 [4], 2013 [5, 6], 2016 [7-9], and 2018 [10, 11]) is to create a single venue for people representing all the diverse backgrounds, disciplines, expertise, and application areas necessary for successful and practical BCI research and development.

The Eighth International Brain-Computer Interface (BCI) Meeting was hosted in the Pheedloop platform (Toronto, Ontario, Canada), which managed individual sessions using the Zoom platform (San Jose, California, USA). Poster sessions and social events were held on the GatherTown platform (gather.town). This Meeting was attended by 395 delegates from 35 countries, a significant growth from the 50 delegates in 1999 [1], although not quite as many as the previous in-person meeting in 2018. Respondents to the 2021 BCI Meeting evaluation survey identified themselves as 40% students, 13% postdocs, 25% faculty members, and 22% other. The BCI Meeting Series is intentionally designed to promote interaction between different groups and different career stages and has advanced the careers of numerous BCI researchers. Many activities are designed to provide educational content and networking opportunities for students and early-career investigators. The 2021 BCI Meeting had a theme of "BCIs: The Next Frontier." The workshops of the BCI Meeting Series provide examples of how BCIs are advancing the frontiers of science and details on both how close we are to realizing new applications and the challenges that remain to be overcome. The workshop summaries presented here serve as an overview of the current status of BCI research and development and present a roadmap to the next steps needed to advance that frontier.

Organization of Workshop Summaries

Workshops for the BCI Meetings are proposed by members of the BCI community, then evaluated and curated by the Program Committee. For the virtual BCI Meeting of 2021, the workshops were assigned to four different schedule slots with three to four workshops running concurrently. In addition, six of the workshops volunteered to run as part of a five-month preliminary series of “BCI Thursdays.” These workshops were the same length and format as the workshops that

occurred during the Meeting, but did not overlap with other BCI Society events and had a separate registration structure. However, they retained the strong emphasis on attendee participation that is central to workshops of the BCI Meeting series. The BCI Thursday series also included free events designed to provide technical background for students on cutting-edge topics in BCI research.

The workshop summaries presented here are divided into three themes and ordered to provide a progression of topics. They can be read sequentially as an overview of the field or separately to provide detail on a topic of interest. However, acronyms are only defined on their first use. For each summary, we report the primary organizer, who is also a co-author of this paper, and list all additional presenters. Each summary is designed to introduce the workshop topic, the latest developments or central ideas presented in the workshop, and the topics of discussion and eventual conclusions. Of course, nothing will substitute for the actual experience of being part of an interactive workshop, even a workshop in a virtual platform. However, the summaries are intended to at least provide an overview and pointers to the information that workshop attendance would have provided. Further, the summaries provide the key points, conclusions, or consensus opinions that resulted from the workshop discussions and may include opportunities to participate in ongoing discussions or collaborations.

Each workshop focused on a specific topic area, yet these topics overlap and complement each other, so that the summaries sometimes create a mosaic examining related ideas from different angles and at other times build on each other. For example, the workshops "*Toward an international consensus on user characterization and BCI outcomes in settings of daily living"* and *"On the need of good practices and standards for Benchmarking Brain-Machine Interfaces"* examine different aspects of standards. Similarly, BCI use for children and people with congenital disabilities are examined in the pair of workshops "*The design of effective BCIs for children"* and "*Non-invasive BCIs for people with cerebral palsy."*

Three general themes provide the structure for this article, although many alternative organizations could be proposed. The themes are independent of the time slot in which the workshop occurred. The first theme is Tools and Methods and contains workshops providing detailed examination of a particular hardware, software, or analysis method. The second theme is BCIs for Specific Populations or Applications and is less concerned with hardware and software than with the outcome produced or the common considerations for working with a specific group. The final theme is Expanding BCI Usability and Availability. The workshops in this theme focus on big picture topics such as standards, translational issues, and ethics as well as the expansion of BCIs into the broad consumer market through applications such as entertainment and human enhancement.

The trajectory of these three themes, and the workshop summaries presented here, creates a progression from foundational topics to translational efforts for standardized clinical applications and BCIs for the population at large. Together these workshops show the diversity of BCI applications and intended users and the complexity of the issues that must be solved to make BCIs into useful tools for the many intended user groups.

Tools and Methods

Focal Bi-Directional Brain Computer Interfacing with Concentric Electrode Technology

Organizer: Charles Anderson (Colorado State University)
Additional Presenters: Walter Besio (University of Rhode Island and CREMedical), Barry Oken (Oregon Health & Science University), Myles McLaughlin (KU Leuven)

This workshop focused on EEG BCI experiments and stimulation studies using tripolar concentric-ring electrodes (TCREs) and the advantages of this technology over conventional disc electrodes. Compared to conventional disc electrodes, TCREs have significantly better spatial resolution and signal-to-noise ratio [12-14]. TCREs increase signal bandwidth for high-frequency signals useful for localizing epileptic brain regions and possibly imagined movements [15, 16]. Imagined movement BCI improved significantly with TCREs [17, 18]. TCREs' increased spatial resolution and signal-to-noise ratio may enable discrimination between finger movements, currently only possible with implanted electrodes. Experiments involving real and imagined finger movements found that EEG from TCREs produced significantly better discrimination among movements of individual fingers (about 70% correct classification) than conventional disc electrodes (about 40%) [19].

TCREs are safe for stimulation [20, 21], and can be used for seizure control [22-26]. The stimulation can block epileptogenesis [27] and alter neurotransmitters to increase the effectiveness of anti-seizure drugs [28-30]. Stimulation experiments are underway to determine if transcranial focal stimulation via concentric ring electrodes is effective for modulating human brains.

Pain is a common medical problem but difficult to objectify as a personal experience of a sensation. Using TCREs both to selectively stimulate pain fibers and to record pain-related evoked potentials (PREPs) is one method of objectifying pain sensation [31-37]. Custom-made concentric stimulating electrodes can selectively stimulate pain afferents where conventional electrical stimulation with mono- or bi-polar stimulating electrodes failed. TCREs delivered paired electrical stimulations to the dorsal non-dominant hand. PREPs were recorded at Cz referenced to ear. For control participants, average PREP N1-P2 amplitude was significantly diminished by electroacupuncture. In another experiment control participants showed the expected habituation of PREP N1-P2 amplitude over time, but those with chronic low back pain showed an increase in PREP amplitude, presumably a physiological marker of central sensitization, the increased responsiveness to sensory information such as nociception.

TCREs on the skull under the skin may be an effective middle ground between implanted stimulation electrodes and the non-invasive but less effective transcranial stimulation. TCREs provide higher magnitude stimulation in gray and white matter than transcranial stimulation. Focused and unfocused stimulation on neurons have been studied in Macaque. Increased spatial precision with TCREs was demonstrated when stimulating rat motor cortex area for rear limb movement.

Conventional electrodes produced movement in both contralateral and ipsilateral limbs, but TCREs only produced contralateral limb movement [38].

Discussion covered practical considerations and design variations, including different numbers of rings and different spacing. TCREs sizes include 10mm, 6mm, 4mm, and even 3.5mm. TCREs use 10-20 paste, but work on using gels and possible dry electrode designs are being considered. Caps to hold TCREs were described, but need work for the smallest TCREs. Two disadvantages of TCREs are the need for a custom pre-amplifier from CREMedical and for precise scalp placement because higher spatial precision means steep attenuation over short distances. Laplacian transforms can be applied to EEG recorded from conventional disc electrodes, but 92 disc electrodes are required to obtain results similar to that provided by one TCRE. Publicly available sample data recorded from TCREs can be found at https://www.cs.colostate.edu/~anderson/res/eeg/tripolar/tripolar.zip.

Invasive brain computer interface technology: Open loop and closed loop decoding applications

Organizer: Christoph Kapeller (g.tec medical engineering GmbH, Austria)
Additional Presenters: Kyousuke Kamada, MD, PhD, (Megumino Hospital, Japan); Aysegul Gunduz, PhD, (University of Florida, USA); Peter Brunner, PhD, (Washington School of Medicine, St. Louis, USA); Kai Miller, MD, PhD, (Mayo Clinic Rochester, Minnesota, USA)

The workshop discussed state-of-the art BCI applications using open-loop and closed-loop decoding and neuromodulation. Implementation of these experimental setups in existing BCI platforms was also discussed.

Invasive electroencephalographic (iEEG) signals, such as electrocorticography (ECoG) or stereo EEG, contain information with high spatial and temporal resolution [39]. Several invasive BCIs have been realized over the past two decades. Closed-loop invasive BCIs have been used for control of prosthetic limbs [40] as well as avatars or cursors [41, 42]. Open-loop invasive BCIs have been used for decoding of speech [43-46], movements [47, 48] and vision [49, 50]. Establishing useful invasive BCI applications requires interdisciplinary efforts for the development of sensors and machine learning algorithms, with specialized efforts to make the resulting technology practical for a medical environment and matched to each individual's clinical indications. Further, the risk of implanting sensors has to be surpassed by the benefit that the BCI provides to meet the specific need of each patient [51].

Recent developments showed a transition from proof-of-concept demonstrations to clinical applications, including open-loop decoding for brain mapping [52-54] and BCI implants [55]. Such implants can provide ALS patients with a powerful BCI [42] and will be further investigated over the next years. The concept of open-loop electrical brain stimulation for neuromodulation has been widely used in presurgical brain mapping. Stimulating the somatosensory cortex can induce sensation in individual fingers [56], while stimulating the visual cortex causes illusory percepts like appearing faces or moving rainbows [57]. Open-loop deep brain stimulation (DBS) has been utilized for more than 40 years to manage

tremor [58]. More recently DBS has been used to treat Parkinson's disease, Tourette syndrome, dystonia, and depression [59]. Closed-loop stimulation based on iEEG signals improves the battery lifetime during the treatment of Tourette syndrome [60] and essential tremor [61]. Most of the aforementioned studies required the integration of sensors and amplifiers into signal processing platforms that are capable of real-time processing and synchronized with the patient's condition and/or stimulus presentation. Example BCI platforms in the workshop were BCI2000 [62] and the rapid prototyping platform g.HIsys in MATLAB/Simulink [63].

Riemannian Geometry Methods for EEG preprocessing, analysis and classification

Organizer: Louis Korczowski (Siopi.ai)
Additional Presenters: Marco Congedo (GIPSA-lab, CNRS, Université Grenoble- Alpes), Florian Yger (LAMSADE, CNRS, Univ. Paris-Dauphine, PSL Research Univ.), Sylvain Chevallier (LISV - UVSQ - Univ. Paris-Saclay), Pierre Clisson (Timeflux Research Group), Quentin Barthélemy (Foxstream)

Riemannian Geometry (RG) is a subject of growing interest within the BCI community. Machine learning methods based on RG have demonstrated robustness, accuracy and transfer learning capabilities for the classification of motor imagery [64], ERPs [65], SSVEPs [66], sleep stages [67], and other mental states [68]. This workshop provided an overview of RG, demonstrating its practical use for signal pre-processing, data analysis, mental state classification, and regression.

RG was first applied to BCI in 2010 [64]. Key articles highlighting different applications of RG include multi-class classification (e.g. minimum distance-to-mean (MDM) classifier) [69], transfer learning (e.g. Riemannian Procrustes Analysis) [70, 71], the first online BCI system using it (e.g. Brain Invaders) [65, 72], and milestone-like performance of RG methods in international competitions [73, 74]. Intrinsic properties of RG methods were discussed to explain their performances (e.g., simple parametrization of models, robustness induced by affine-invariant metrics) but also some drawbacks and how they can be managed (e.g. sensitivity to rank deficiency at high dimensionality) [75, 76]. Interestingly, RG can be used in combination with other effective methods such as common-spatial pattern and/or deep learning to outperform methods using Euclidean space alone, e.g. by projecting data in a tangent space [74].

The ecosystem of open-source libraries (that was scattered and scarce before) is now mature enough to improve several steps of the BCI system. For example, Riemannian methods outperforms Euclidean methods in accuracy and simplicity in use cases such as automatic artifact detection (e.g., Riemannian potato) [77, 78] or ERP classification (e.g. MDM with super covariance matrix). These performances are tested using the fair benchmarking approach [79] and are easy to replicate in online BCI thanks to libraries such as Timeflux [80].

Despite its performance advantages, publication data from https://www.dimensions.ai/ show that articles mentioning new contribution of "Riemannian Geometry" applied to BCI has remained in the range of 7 to 21 per year in the period 2016-2020 (mean citations : 27.71). For comparison, mention of

"common-spatial patterns" associated with BCI increased from 71 to 119 articles per year (mean citations: 20.75) and "deep learning" from 15 to 179 articles per year (mean citations : 11.67) in the same period.

We argue that the gap between the observed performance of RG applied to BCI and the proposal number of contributions in this field may be attributed to some combination of a perceived lack of easily accessible resources to make RG widely available to BCI research (e.g. 65.7% of respondents to the workshop questionnaire had never used RG before) and the lack of reproducible tools for benchmarking different methods while taking into consideration datasets heterogeneity (discussed at the previous BCI meeting workshop [81]).

This workshop was created to address these issues by increasing awareness of available resources for RG and encourage benchmarking with tools such as MOABB on a larger scale of datasets [79]. We encourage everyone to report benchmarking results. Further, we invite everyone to join us by using the open-source RG tools, and by contributing to the improvement of these tools either by providing feedback, or contributing to the open source project pyriemann. All the workshop resources are accessible, including slides, code tutorial, online demo, exhaustive workshop Q&A, and linked data: https://github.com/lkorczowski/BCI-2021-Riemannian-Geometry-workshop.

Open-source Python tools for BCIs

Organizer: Pierre Clisson (Timeflux Research Group)
Additional Presenters: Raphaëlle Bertrand-Lalo (Timeflux Research Group), Sylvain Chevallier (LISV, Université Paris-Saclay), Marco Congedo (GIPSA-lab, CNRS, Université Grenoble-Alpes)

Python started as a general-purpose programming language but has evolved into a tool of choice for the scientific community, quickly overtaking specialized languages such as R and MATLAB [82]. Several factors account for its success: Python is easy to learn, has a strong community, and benefits from a rich and efficient data science ecosystem.

This workshop had a two-fold objective: give an overview of the Python BCI landscape and provide hands-on instructions on a few chosen open-source tools.

As a foundation for the focus on practical BCI, we first reviewed the main BCI paradigms and the typical workflow of a BCI pipeline. We discussed common challenges for BCI applications: the need for precise synchronization of the EEG signal and the stimuli, the difficulty of obtaining good quality signals in real-life conditions, and the challenges of calibration.

Riemannian geometry (RG) for EEG-based BCI [65, 83] has produced state-of-the-art results in international competitions [76]. Machine-learning algorithms based on RG offer many advantages. They are computationally efficient and thus suitable for online applications. They usually converge to optimal results relatively quickly, reducing calibration duration (ongoing studies on transfer learning are attempting to remove this phase completely [70] [84]). Finally, they do not depend on the BCI paradigm and work equally well for ERP, SSVEP, and motor imagery tasks.

PyRiemann [85] is an actively maintained Python package for manipulating covariance matrices. It implements multiple data transformation techniques and classification methods. Workshop participants were guided through a Python notebook and instructed on using this library with concrete examples.

The RG framework includes multiple signal classification strategies and BCI researchers use many other algorithms, such as Logistic Regression, Regulated LDA, Support Vector Machines, and Neural Networks [86]. Valid comparisons between methods are essential. The Mother Of All BCI Benchmarks (MOABB) [79, 87] project offers comprehensive comparison tools that enable ranking new and existing algorithms with publicly available datasets, paving the way for reproducible research. We reviewed a practical example and explained the underlying code.

Timeflux (https://timeflux.io/) [80] is an open-source framework for building online BCIs. It is capable of acquiring, recording, and processing biosignals in real-time. It can also present precisely scheduled stimuli. It works hand-in-hand with PyRiemann and MOABB and rests on the shoulders of standard libraries such as Pandas [88], Scikit-learn [89], Lab Streaming Layer [90], and HDF5 [91]. It comes with a rich set of nodes and plugins for dynamic epoching, matrix manipulation, digital signal processing, machine learning, and other tools. It also provides a convenient JavaScript API for developing web interfaces. We reviewed the architectural principles of Timeflux and explained how to use it to design a P300 speller, finishing with a functional demo that runs in a web browser.

We only introduced the potential of the Python language for the BCI field. For instance, we only briefly described MNE [92], a full-fledged framework for offline analysis of EEG and MEG signals. This workshop provided a good starting point for further exploration. The presentation slides, notebooks, and code are publicly available [93].

Artificial Intelligence in Brain-Computer Interfacing

Organizer: Moritz Grosse-Wentrup (University of Vienna)
Additional Presenters: Tonio Ball (University of Freiburg), Aldo Faisal (Imperial College London), Gernot Müller-Putz (Graz University of Technology)

Artificial intelligence (AI) methods in general, and deep learning algorithms in particular, have revolutionized the field of machine learning [94]. Current AI systems outperform human experts in various cognitively challenging tasks [95, 96] and have enabled scientific insights that arguably could not have been obtained by human intelligence alone [97]. More recently, deep learning methods have been adapted to and developed for brain decoding and BCI systems [98, 99]. Building on a long history of discussions on the benefits of nonlinear decoding methods in BCI [100], this workshop discussed whether AI can outperform traditional BCI machine learning methods and which challenges should be addressed to realize the full potential of AI in BCI.

The consensus on the current performance of AI-BCI methods was that they perform essentially on par with the best non-deep decoding algorithms. However, a rigorous comparison of state-of-the-art Riemannian decoding methods [76, 101] with AI algorithms has yet to be done. The workshop participants concluded that a large-scale brain decoding challenge, e.g., hosted by a major AI or machine

learning conference, would be well suited for realizing a fair comparison of competing decoding architectures (e.g., https://beetl.ai/).

The workshop participants then considered which issues prevent, at least so far, AI methods from revolutionizing BCI systems in the same way they have already transformed other data-driven applications. The primary bottleneck identified in the discussion was the absence of large-scale datasets in the field of BCI. These datasets would ideally comprise thousands or even millions of BCI users from heterogeneous settings, i.e., including numerous experimental paradigms, recording setups, and user groups. While the workshop participants acknowledged the efforts of the BCI community to record large-scale datasets [102], they also noted that collecting datasets on a similar scale as those available in other scientific disciplines [103] is probably beyond the capabilities of the academic community. Consequently, the discussion shifted to the role of commercial BCI applications in recording and providing access to large-scale datasets. Several consumer EEG headsets have reached market readiness with the expectation of prompt deployment in passive BCI applications[104]. Comprehensive access to data recorded by these applications could provide the large-scale datasets required to realize the full potential of AI-BCI systems. In particular, the heterogeneous nature of such data, which stands in contrast to the homogeneous data typically recorded in academic settings, could be considered an advantage. The diversity of data might be leveraged to create feature representations that are user- as well as hardware-independent. Such feature representations would be essential to realize zero-training BCIs for commercial applications [105-107].

However, leveraging commercially recorded EEG datasets poses significant practical, legal, and ethical challenges. It is unclear what incentives companies would have to share their data publicly. Also, procedures would have to be developed that realize informed consent and honor data privacy regulations. The workshop participants considered an active engagement of the BCI community with industrial partners essential to make large-scale datasets a reality and realize the full potential of AI-BCI systems.

Adaptation in closed-loop BCIs

Organizer: Tetiana Aksenova (University Grenoble Alpes, CEA, LETI, CLINATEC)
Additional Presenters: Amy L. Orsborn (University of Washington), Martin Bogdan, Sophie Adama (Universität Leipzig), Blaise Yvert (U1205 Inserm, University Grenoble Alpes), José del R. Millán (University of Texas at Austin), Jean Faber (Universidade Federal de São Paulo)

BCI decoders calibrated in an open-loop, offline paradigm but then applied in close-loop, online paradigm show a significant drop in decoding performance. Adaptive algorithms in a close-loop session decrease this shortcoming by directly adjusting BCI parameters to incoming data. In addition, both the user and machine learn in a closed-loop BCI.

Closed-loop paradigms are often applied to BCIs that decode motor signals. Intracranial ECoG [108, 109] from a participant with tetraplegia was decoded with a fully adaptive decoder to operate a 4-limb exoskeleton. The decoder used an

adaptive Markov mixture of multilinear experts [110] to switch between independent decoders (experts) to interpret multiple degrees of freedom.

Closed-loop paradigms enable user/decoder co-adaptation to maximize performance through synergistic user-machine interactions between the two learners [e.g., 111]. However, learning trajectory models are needed to optimize these co-adaptive systems. A new game-theoretic model of co-adaptation [112] provides a framework to analyze system equilibria and predicts learning trajectories, but requires validation.

The balance of decoder vs patient adaptation is important. EEG-based motor BCIs illustrate the pros and cons of extensive machine-learning adaptation. Non-supervised context-aware algorithms can rapidly adapt so users can use a language model-based speller [113] without a calibration phase [114, 115]. However, this does not promote user learning—EEG patterns for BCI commands actually became less separable with practice rather than improving [115]. True mutual learning, where decoder and user learn from each other, seems to require slow decoder adaptation to promote improved EEG features [116] as seen in several longitudinal studies [117].

Mutual learning implies cortical plasticity and the BCI use as a neurorehabilitation tool specifically designed to support plasticity (i.e., user learning). A clinical trial in patients with severe hand plegia from stroke compared the effect of BCI-operated vs random functional electrical stimulation. Only the BCI group had significant and clinically important functional improvement and a significant increase of functional connectivity in the damaged sensorimotor hemisphere [118]. Regulation of the magnitude of the required EEG response was critical to keep the patient's attention high and promote recovery.

Hybrid BCIs (HBCIs) integrate brain and non-brain data sources with different classifiers schemes (serial, parallel, mixed) to achieve better results [119]. Thus, neuroplasticity can happen in multiple dimensions and temporal scales. Different learning times are associated with different physiological systems such as autonomic learning (heart/breath adaptation) [120, 121], motor learning (agency and control refinement) [122, 123], central learning (cortical adaptations) [124], and cognitive learning (embodiment, ownership and spatial perception) [125]. HBCIs therefore present a more complex challenge for balancing classifier adaptation rate vs. neural plasticity.

Adaptive BCIs also exist for non-motor applications. The hybrid Adaptive Decision Making system was designed for a patient with complete locked-in syndrome (CLIS) and uses multiple EEG features (Granger causality, the imaginary part of the coherency, and multiscale sample entropy) to increase the probability of correctly evaluating consciousness level [126]. Caregiver observations regarding the patient's state were input into the machine learning system to personalised consciousness level estimation. An adaptive speech BCI application illustrates the risk of audio contamination of neuronal activity recordings [127].

Group discussion placed a priority on developing better understanding of co-adaptation from both theoretical and experimental viewpoints to optimize BCI training and user benefit.

Optimising BCI performance by integrating information on the user's internal state

Organizer: Sebastian Halder (University of Essex)
Additional Presenters: Philipp Ziebell, University of Würzburg), Angela Riccio (Fondazione Santa Lucia), Yiyuan Han (University of Essex)

Ideally, a BCI could detect the physical and mental state of the user and adapt accordingly to allow optimal BCI control for both unimpaired and motor impaired end-users. This adaptation could (1) determine when to start, pause or stop a BCI session, (2) adapt parameters of the BCI session such as trial length, stimulus and feedback modality or (3) switch between BCI and other assistive technology types. User-centered design (UCD) is critical to optimize BCI control in this manner [128]. In general terms, an assistive technology should enable a person with a disability to overcome barriers in daily life, education, work, or leisure [129]. This can only be achieved if the needs and requirements of the user are investigated [130, 131]. Regarding BCI design, the cognitive [132-134] and physical [135, 136] characteristics of end-users need to be considered [132, 133]. Based on this knowledge, we can implement a system that adapts to the internal state of the user.

The UCD evaluation process is built around metrics to determine effectiveness (accuracy in percent of correct responses), efficiency (information transfer rate in bits/min and subjective workload) and satisfaction (via visual analogue scale, questionnaire, or user interview) [137, 138]. These metrics should also inform earlier stage BCI development before end-user evaluation [139, 140]. Further factors should be considered when designing the BCI paradigm, for instance, the design of tasks, feedback, instructions, and signal processing [86, 141-143]. Performance may improve via engaging task design (e.g., a “Star Wars Mission” task) and exploring different stimulus modalities (such as auditory and tactile) and better understanding of the mechanisms underlying training with a BCI [140, 144].

User characteristics ranging from physiological (e.g., the amplitude of the sensorimotor rhythm during rest [145]) to psychological (e.g., the ability to concentrate [132, 146]) can influence performance in varying degrees. For example, a user with a traumatic brain injury may be in a minimally conscious state with only transient windows of consciousness [147, 148]. Identifying such windows is an undeniable prerequisite to BCI control [149]. Evaluation of the efficacy of such measures and any new measures that will be developed can be accomplished during pharmacologically induced loss of consciousness such as the Wada test [150]. More subtle influences on BCI control may arise due to mood and motivation, fatigue and workload or whether the user is experiencing pain, which can be detected using integrative features such as phase-based connectivity [151-153]. Ideally, the BCI could adapt to all changes in the users’ state. Doing this efficiently requires knowledge of features in the EEG (or other signals) that reflect the state of the user.

Many challenges must be resolved before the full potential of the state of the user can be reliably used to optimize BCI performance. The main challenge comes from the variety of states that need to be decoded, each requiring the identification

of signal features that reflect these states, and integrating real-time identification of the states into the BCI design and usage environment.

BCIs for Specific Populations or Applications

The design of effective BCIs for children

Organizers: James J.S. Norton (National Center for Adaptive Neurotechnologies), Disha Gupta (National Center for Adaptive Neurotechnologies), Eli Kinney-Lang (University of Calgary)
Additional Presenters: Kim Adams (University of Alberta), Tom Chau (University of Toronto), Erica Floreani (University of Calgary), Kathleen M. Friel (Burke Neurological Institute), Dion Kelly (University of Calgary), Adam Kirton (University of Calgary), Ilyas Sadybekov (University of Calgary), Corinne Tuck (Glenrose Rehabilitation Hospital-I CAN Centre)

BCIs have the potential to enhance, restore, or replace function in children with neurodevelopmental disorders, neurodegenerative disorders, and severe motor disabilities caused by stroke, spinal cord injury, or other acquired injuries [154-157]. However, few studies have investigated BCIs for children [158-161] and these studies show conflicting results; it remains unclear whether children—especially those with neurological disabilities—can effectively use BCIs. Thus, this workshop was organized into three discussion panels that:

1. **Examined how BCIs can improve children's quality-of-life –**Children can use BCIs to [162] communicate, play games, and express themselves creatively. The greatest benefit BCIs offer children with motor disabilities is a sense of control, motivating children to engage more with BCIs and enabling them to practice repetitive tasks that lead to learning. Thus, the child's perception of a successful BCI may not match that of a researcher. For example, operating a BCI using a combination of brain activity and artifacts may improve the child's life and be considered a success from the child's perspective. Therefore, special consideration is needed to simultaneously engage children in activities that are educational, therapeutic, meet the goals of researchers, and are engaging for the children. Recommended strategies are gamification [163-167] and close interdisciplinary collaboration between diverse experts.
2. **Discussed the interfacing, signal-processing, and physiological challenges encountered during the design of BCIs for kids –** Developing BCIs for children presents unique signal acquisition, data analysis, and reporting challenges [154]. Signal acquisition hardware for pediatric BCIs needs to be more portable, lighter, more comfortable, and easier to use (e.g., faster setup, dry electrodes, robust to artifacts). Presently only a few signal analysis pipelines exist for pediatric BCIs [168, 169], due in part to differences in the EEG from children compared to adults [170]. For example, P300 timing varies more in children and BCIs may be more fatiguing for children. Improved and consistent reporting of demographic information and experimental details would allow for better cross-study analyses. Lastly, improved user interfaces are an area of critical need for pediatric BCIs.

3. Considered the use of BCIs for children as augmentative and alternative communication devices and for rehabilitation in clinical settings – The design of BCIs for communication and rehabilitation in children benefits from a patient-centered and neurologic deficit specific approach [161, 171]. For example, many children express an interest in using BCIs for gaming and social play. Collaborative and competitive interactions between family members, and especially siblings, are a critical social outlet for children with motor deficits that motivate them to use BCIs. Neurological deficits may be caused by damage to small areas of the brain that were acquired very early in life. Thus, the brain may reorganize and researchers should work with clinicians to consider neuroplasticity in the design of BCIs for children [172, 173]. In addition, working with clinicians and families will increase awareness of the potential of BCIs for children [174].

As members of the pediatric BCI community, we must put children first, understand what children want out of BCIs, and make it happen.

Non-invasive BCIs for people with cerebral palsy

Organizer: Jane E. Huggins (University of Michigan)
Additional Presenters: Katya Hill (University of Pittsburgh), Petra Karlsson (Cerebral Palsy Alliance, University of Sydney), Reinhold Scherer (University of Essex)

This workshop included extensive discussion about BCI design considerations for people with cerebral palsy (CP), the most common childhood physical disability [175]. CP is caused by injury or genetic abnormalities affecting the brain early in life leading to 15-19% without a communication method even with assistive technology [176-179]. However, BCIs that provide augmentative and alternative communication (AAC) for individuals with adult-onset impairments may unintentionally rely on skills that people with CP have not had an opportunity to learn.

Issues from the workshop *Design of Effective BCIs for Children* apply to children and adults with CP because of missed educational opportunities. Even those who have successful communication technology may need a BCI as age increases the severity of motor impairments. This makes BCI a competitive access option. For example, a participant with CP had similar communication rates on an AAC device with head-pointer access (1.33 words-per-minute, wpm) and BCI access (1.29 wpm).

Overall, BCI studies with people with CP show mixed results [162, 180, 181]. Some comparisons of BCI designs showed that SSVEP and SMR designs were preferred to the P300 design and had better performance [181]. Other comparisons of naïve users showed that some had significant SMR-BCI control (2 classes, 82±12%), others significant SSVEP-BCI control (4 classes, 43±7%), but few could use both and some could not use any BCI [182, 183].

Such results raise the specter that current BCI methods may not be appropriate for people with CP. If a person has no voluntary motor control, can they operate a

motor imagery BCI? Can people with limited access to schooling count flashes of a P300 BCI or perform mental arithmetic or spatial navigation?

EEG recordings are complicated in people with CP due to head shape variations or improper electrode cap fit [184, 185] as head asymmetry is reported among 40% of people with the most severe impairments from CP [186] and microcephaly at 30% [187] to 60% [188]. Abnormal neuroanatomy can also cause unusual localization of cortical function [189]. The impact on BCI is uncertain, but people with severe CP can benefit from individualized electrode locations [184, 190].

Extraneous movements, which are common [191], can also create EEG artifacts [e.g., [182]] and may make it difficult to focus on the BCI display. Further, gaze or visual impairments including ptosis (drooping) of the eye lid, nystagmus, and cerebral visual impairment (CVI) can lead to difficulty interpreting visual stimuli [192]for an SSVEP or P300 BCI device or visual feedback for an SMR BCI. Thus, special care is needed to understand how well the user can interpret visually presented information.

Indeed, user-centered design is important throughout BCI design and user training. Acclimation regimes may be needed with step-by-step introduction of individual BCI concepts. Family interactions, cooperation, and competition can increase motivation and engagement, which are essential for learning, but not a guarantor of good performance [193]. These factors are crucial as people with CP may have a long history of unsuccessful attempts to operate technology. Thus, the ideal BCI would be calibrated without the user following instructions, have intuitive operation and be inherently engaging. In addition, systems should build on familiar concepts, such as row-column scanning, to simplify the transition from calibration to end-use [183].

Ultimately, we need improved understanding of the effect of CP on EEG, user-centered design to match the BCI to the interest and needs of individual users, and user-tailored training paradigms. Finally, it is vital to recognize that for children with congenital disabilities, technology use and even communication itself, are skills that must be taught.

From Speech Decoding to Speech Neuroprostheses

Organizer: Christian Herff (Maastricht University) and Sergey Stavisky (University of California, Davis)
Additional Presenters: Jon Brumberg (Kansas University), Phil Kennedy (Neural Signals Inc.), Miguel Angrick (University of Bremen), Julia Berezutskaya (Radboud University), Qinwan Rabbani (Johns Hopkins University)

Despite impressive recent results in decoding speech from neural recordings, there remain many challenges to achieving a real-time, large-vocabulary BCI for restoring lost speech. In this workshop, five of these challenges, and potential solutions, were discussed.

First, existing speech decoding demonstrations have not yet achieved consistently intelligible outputs. Multiple groups presented new decoding architectures, including recurrent neural networks and GANs. Workshop

participants agreed that these modern machine learning approaches should benefit from additional data in future studies, and noted that all of the work presented used less than 20 minutes of neural recordings. Further, their performance did not saturate with training data quantity subsampled within these limited datasets.

A second challenge is how to obtain highly informative neural correlates about speech intent. Previous research almost exclusively relied on ECoG signals, which are not regularly used for long-term measurement. However, high-quality speech decoding and synthesis can also be achieved using penetrating microarrays implanted in the dorsal motor cortex [194], even though that area is not typically associated with speech production [195]. These Utah arrays have been used for multiple-year recordings in a number of participants and achieved high performance in, e.g., online decoding of attempted handwriting in people with tetraplegia [196] or speech perception decoding [197]. Alternatively, stereotactic EEG, which is very similar to Deep Brain Stimulation electrodes [198] that routinely remain implanted for decades, was proposed for high-quality speech synthesis. The neurotrophic electrode, an entirely different type of electrode with good long-term potential [199], was also proposed for speech neuroprosthesis [200].

Third, a functioning neuroprosthesis needs to generate or decode speech in or near real-time [45]. However, previous studies demonstrating speech synthesis [44, 201] or speech recognition [202, 203] from ECoG data have primarily (except for [204, 205]) been done offline on previously recorded overt or whispered speech. Approaches that process and decode intracranial EEG in real-time will provide direct feedback to the patient. This has been done using imagined speech processes [206], building on prior work such as [207]. Recent progress towards a low latency (250 ms) ECoG speech synthesis pipeline shows proof-of-concept open-loop results. A non-invasive EEG neuroprosthesis based on an artificial vocal tract model [207] provides auditory and visual feedback to the user and might therefore help train speech neuroprosthesis users and pilot online speech BCI methods.

Fourth, the field would benefit from better speech synthesis performance metrics. Recent works typically uses variants on measuring correlation between true and decoded audio (e.g. for spectral or pitch features), which are poor proxies for intelligibility. Workshop participants agreed that adopting subjective intelligibility metrics is important, but this may need to wait until decoding performance is good enough for these metrics to become relevant (or else they will suffer from floor effects).

Fifth, all presenters agreed that data sharing is key to accelerating progress. One recently shared large dataset of speech perception in fMRI, ECoG, and sEEG, along with the associated impressive reconstruction quality provides the public research community with a fully annotated dataset [208].

Brain-computer interfaces for the assessment of patients with disorders of consciousness

Organizer: Christoph Guger (g.tec Guger Technologies OG)
Additional Presenters: Damien Coyle, (Ulster University), Kyousuke Kamada, (Hokashin Group Megumino Hospital), Rossella Spataro, (University of Palermo), Jing Jin, (East China University), Steven Laureys, (Brain

Centre & GIGA Consciousness, Coma Science Group, University and University Hospital of of Liege, Belgium; International Disorders of Consciousness Institute, Hangzhou Normal University, China; CERVO Brain Research, U Laval)

Bedside evaluation to assess conscious awareness after coma requires inferences based on patients' motor responsiveness [209] with limited diagnostic precision and prognostic information, increasing the ethical difficulty of decisions on life-prolonging therapies. Technologies such as functional neuroimaging and BCIs provide objective tools for diagnostic, prognostic and therapeutic purposes [210]. About two thirds of patients clinically diagnosed with "unresponsive wakefulness syndrome (UWS)" (or "persistent vegetative state") may show residual brain activity in PET studies [211] and are hence actually in a minimally conscious state (MCS) with a better chance of recovery.

BCIs can help reduce the diagnostic and prognostic uncertainty of both acute and chronic disorders of consciousness [212, 213]. BCI should first be used to establish a reliable and reproducible response to a simple command. Then one can attempt functional communication with simple yes/no questions and eventually spelling or message creation [212, 213]. The mindBEAGLE (g.tec medical enginering GmbH) uses auditory P300, vibro-tactile P300 and motor imagery paradigms for these steps and rehabilitation protocols. Paradigms include a quick (2-8 minute) system calibration or patient assessment. Other BCI systems have also been designed for this purpose, including using auditory sensorimotor rhythm feedback for those with visual impairments [214, 215].

BCI assessment of DOC with locked-in and completely locked-in patients found 9 out of 12 patients could demonstrate command following by answering YES/NO questions [216]. Building on the pilot of 15 patients reported in [215], the workshop reported an update with 25 patients who each participated in 10, one-hour motor imagery BCI sessions. Of these, 5/9 UWS, 7/11 MCS, and 3/4 locked-in syndrome demonstrated significant capacity to modulate brain activity in stage I (assessment) and progressed to stage II/III (auditory feedback training and Q&A response). All participants in stage II/III responded significantly to YES/NO questions. Another study with unresponsive patients showed 3 out of 12 patients could successfully answer the YES/NO questions on some assessment days [217], showing that these patients have fluctuations in consciousness that can be detected by BCI systems.

BCIs can also help predict eventual recovery. Auditory P300 and vibro-tactile P300 provided a predictor of functional recovery for two patients with DOC. One patient did not show any auditory P300 or vibro-tactile P300 after three weeks and coma continued for more than 6 months. A second patient responded to auditory P300 and vibro-tactile P300 and after 6 months had recovered from coma and understood verbal commands. Such patients may benefit not only from BCI assessment, but also from BCI-based rehabilitation [218]. Longitudinal observation of 12 DOC patients showed that achieving mindBEAGLE classification accuracy of at least 50% predicts recovery of behavioural responsiveness (after six months) as measured by the coma-recovery scale revised (CRS-R) [219]. Moreover, 12 of 20 patients showed CRS-R score improvement after 10 sessions of a vibrotactile stimulation protocol [218].

BCI can also evaluate the effectiveness of other treatments for arousing DOC patients by analyzing EEG recorded during mental tasks before and after intervention. BCI methods have been used to assess the effectiveness of spinal cord stimulation and deep brain stimulation surgeries in arousing vegetative patients. Auditory, vibro-tactile, or motor imagery-based BCI systems have been used to assess 5 unresponsive patients and 3 vegetative patients in this on-going study.

BCIs are being cross-validated against neuroimaging techniques such as PET and fMRI [220]. The current challenge is to integrate BCIs with our increasing scientific understanding of recovery from severe brain injury to optimized the trajectory of clinical care after coma and improve the quality-of-life in disorders of consciousness and locked-in syndrome [221].

The promise of BCI-driven functional recovery after stroke: leveraging current evidence to define next steps

Organizer: A Nicole Dusang (Brown University/Providence VA Medical Center/ Massachusetts General Hospital)
Additional Presenters: Murat Akcakaya (University of Pittsburgh); Febo Cincotti (Sapienza University); Cuntai Guan (Nanyang Technological University); Christoph Guger (g.tec medical engineering GmbH); Kyousuke Kamada (Asahikawa Medical University); David Lin (Massachusetts General Hospital/ Providence VA Medical Center); Donatella Mattia (Fondazione Santa Lucia IRCCS); José del R. Millán (University of Texas at Austin); Ander Ramos-Murguialday (University of Tübingen / TECNALIA Research and Innovation); Vivek Prabhakaran (University of Wisconsin-Madison); and George F. Wittenberg (Pittsburgh VA Healthcare System / University of Pittsburgh)

Stroke is a leading cause of long-term disability worldwide, and 30–50% of stroke patients experience limited recovery. Rehabilitative EEG-BCIs are a promising neurotechnology for restoration of function after stroke. The hypothesis behind rehabilitative BCIs is that coupling neural activity with sensory feedback of limb movement induces cortical plasticity, improving functional recovery. This workshop featured twelve researchers developing rehabilitative EEG-BCIs for functional recovery from ten institutions around the globe. Presenters were split into two panels to consider how to translate this technology from the lab to the clinic. Randomized controlled trials (RCTs) have demonstrated the benefit of Rehabilitative EEG-BCIs, but employed diverse control methods, therapy doses, dosing intervals, and different types of neural dynamics and sensory feedback.

Panel 1 discussed optimal EEG-BCI support for stroke rehabilitation. Spatial neglect is an often overlooked deficit in stroke patients though it can significantly impact a patient's response to therapeutic intervention [222]. Technology is needed to objectively map neglect, quantify changes during recovery, and provide a rehabilitation platform to target spatial neglect. Although BCI addresses a gap in standard neurorehabilitation medicine [223], it still lacks an American Heart Association (AHA) class and evidence rating. BCIs empirically measure the signals of the damaged cortex and patients' functional disability during recovery. Rehabilitative EEG-BCIs restore the neural activity-functional output connection, supporting the retraining of neural activity. This is demonstrated by a RCT evaluating an EEG-BCI intervention for distal upper extremity function in a chronic

stroke population [224]. Results showed 64% of participants made significant gains in both primary and secondary outcome measures.

Panel 2 reflected on stakeholders' needs for translating this promising technology to a clinical environment. Though RCTs have demonstrated the therapeutic efficacy of rehabilitative EEG-BCIs, commercialization requires clear clinical and economic benefit and reliable function within the rigors and environment of long-term clinical use. BCI-FES systems must address both patients' and clinicians' needs [118]. Patients need an effective and engaging rehabilitation platform, while clinicians require a plug-n-play system with remote technical assistance and joint analysis. Unanswered questions remain along the spectrum of basic research to patient care [225]. The field has yet to determine the optimal neural modalities or features for rehabilitative EEG-BCIs, resulting in significant feature extraction variability in current EEG-BCI platforms. Additionally, past and current RCTs employed diverse outcome measures since no measure is clearly best for capturing recovery. Further, stroke is itself a heterogeneous condition and much remains unknown about the relationship between the type and location of damage and resulting deficits. The RecoveriX system (Guger Technologies), a certified medical product, analyzes motor imagery to trigger FES for upper and /or lower limbs. RecoveriX has shown effectiveness for spasticity reduction and movement restoration in upper and lower limbs [226, 227].

Convincing clinicians, patients, and payers that Rehabilitative BCIs are a worthy technology for investment was felt to require a large, multi-site, randomized control trial study, incorporating methods to minimize, or scientifically account for, heterogeneity between technology and control populations at various sites. Ideally, it will also address knowledge gaps such as long-term effects, dose-response curves, patient stratification, control features, and a comprehensive outcome evaluation.

Towards the decoding of neural information for motor control: present and future approaches

Organizer: Gernot Müller-Putz (Graz University of Technology)
Additional Presenters: Andrea I. Sburlea (Graz University of Technology), Valeria Mondini (Graz University of Technology), Damien Coyle (Ulster University), Cuntai Guan (NTU Singapore), Tonio Ball (University of Freiburg)

For people with a cervical spinal cord injury (SCI) from trauma or disease, upper extremity function is often reduced or lost, resulting in dependency on a caregiver or family member for most daily activities. BCI researchers have for decades worked to derive motor commands directly from brain activity to bypass the interrupted spinal cord pathways and establish direct control of a neuroprosthetics device [228] or robotic arm/exoskeleton [229]. Implantable BCI approaches have produced many advances [230, 231], however, in recent years, non-invasive approaches have moved beyond proof of concepts [232-234] and made major steps towards full arm control. This workshop focused on state-of-the-art approaches to non-invasive neural control of movement.

Non-invasive detection of multiple types of hand movements have been reported, including for people with cervical SCI [235, 236]. Analysis of movement-

related cortical potentials (MRCP) can detect and decode single hand movements [237] or movement attempts (e.g., hand open vs. hand close) or even different grasps (e.g., palmar vs. lateral grasp) [238, 239].

Understanding the neural and behavioral mechanisms involved in grasping is important for successful decoding. Investigations included the relationship between the broad-band EEG representation of observing and executing a large variety of hand-object interactions and the muscle and kinematic representations associated with the grasping execution [240]. Object properties and grasp types can be decoded during the planning and execution of the movement. Properties of the objects could be decoded even during the observation stage, while the grasp type could be accurately decoded even during the object release stage [241].

While the decoding of arm/hand trajectories has mainly been shown in intracortical recordings, major steps in the non-invasive field have been demonstrated. Closed-loop continuous decoding of executed [242, 243] but also attempted arm movement [244] has been done from low frequency EEG. Movement parameters like position and velocity, necessary for decoding [245, 246] were presented. In particular, the contribution of non-directional movement-parameters (distance and speed) has been highlighted [247-249]. Also, the first evidence for online decoding of attempted continuous movement has been reported [244]. Eye movement artifacts present a special challenge for all non-invasive decoding studies. Participants must be permitted to use their gaze to follow the feedback, electroc-oculogram (EOG) signals must therefore be removed from the EEG online [250].

In addition to decode of low frequency EEG components, decoding of executed and imagined 3D reaching tasks have involved delta frequencies, but also alpha, low and high beta frequencies [251, 252]. These studies include decoding of 3D lower limb movements that could be important for gait rehabilitation [253].

In the area of motor imagery and stroke rehabilitation, deep learning methods and convolutional neural networks (CNN) have been used for participant specific [254, 255], participant-independent [256], and adaptive classifiers [257]. CNNs have also been used in assistive robot control with online adaptive motor classification [258].

Beyond the pure application of CNNs for decoding [98], the internal data representation and the effects of hidden unit activations provide possible insights into what the units of such networks learn and the possible hierarchical organization of spectral features [259]. These first insights may open a new way of understanding brain processes.

Biomimetic approaches to restore somatosensation

Organizer: Robert Gaunt (University of Pittsburgh)
Additional Presenters: Sliman Bensmaia (University of Chicago), Karthik Kumaravelu (Duke University), Alberto Mazzoni (Scuola Superiore Sant'Anna), Emily Gracyzk (Case Western Reserve University), Luke Bashford (California Institute of Technology), Chris Hughes (University of Pittsburgh)

Rapid advances in BCI capabilities to decode and restore upper limb motor functions [260] often ignore the accompanying sensory losses. Strategies to restore somatosensation include intracortical microstimulation [261, 262], cortical

epidural stimulation [263-265], peripheral nerve stimulation [266-268] and spinal cord stimulation [269]. Regardless of approach, it is difficult to select stimulus parameters that improve the quality of conscious percepts and maximize functional capabilities. This workshop explored the idea of using biomimicry as a framework to create stimulus trains. Biomimetic stimulation leverages knowledge of intact somatosensory neurophysiology with the intuition that stimulation parameters that evoke patterns of neural activity that match normal patterns will improve perception and function.

Decades of work characterizing skin mechanoreceptor responses in the hand during object manipulation [270] were integrated into TouchSim to accurately simulate primary afferent responses to a mechanical input [271]. The simulated population-level activity resembles the spatiotemporal dynamics of somatosensory neurons in the cortex during the same mechanical stimuli [272], with large transient signals at contact onset and offset [270, 273]. However, simply replacing recorded or simulated spikes with stimulation pulses does not replicate the sensation. Additional computations are required to address anatomical complexities and electrical stimulation biophysics. A simulation platform using genetic algorithms and finite element models of the cortex, populated with realistic neurons, was developed to address these complexities [274]. Critically, the stimulus trains created through simulation more faithfully represented the desired cortical activity than stimulus trains designed using standard methods.

The utility of this computational tool and the principles of biomimicry were tested in peripheral nerve stimulation experiments in amputees. As a baseline, linear stimulation encoding schemes that did not capture important features of natural neural coding were effectively used by participants [266]. Similarly, event-based stimulation encoding that mimicked the natural onset-offset dynamics of primary afferents was also effective [275]. However, in a direct comparison, TouchSim was used to create multiple stimulation trains that were increasingly biomimetic. The most natural sensations were obtained with the stimulus trains that maximized biomimicry [276]. In other experiments, early work suggested that a particular biomimetic train could improve naturalness [268]. Upon repetition, and despite considerable effort to combine modeled fascicle recruitment with biomimetic and non-biomimetic stimulation trains, just two of five participants reported more natural sensation using biomimetic trains, highlighting the limitations of single-subject studies of perception.

Two different aspects of biomimicry were explored in human intracortical BCIs. Motor imagery and actual movement evoke similar brain activity. To explore this concept for somatosensation, neural activity patterns were recorded in somatosensory cortex and the supramarginal gyrus during imagined sensations [277]. Different imagined sensations were encoded stably in the somatosensory cortex, suggesting that imagined sensation could guide stimulus train design, even in people left insensate from their injury. Finally, in a direct test of biomimetic principles, intracortical stimulus trains using fixed amplitudes and frequencies were compared to trains with stimulation amplitudes modulated by cortical activity patterns recorded from non-human primates [273]. The participant frequently rated the biomimetic trains as more natural, especially when the overall intensity was matched.

In summary, biomimicry is a principled and likely fruitful approach to create stimulation trains to restore somatosensation. Simulation and modelling tools can help design these trains, which have outperformed less realistic trains in both the peripheral and central nervous systems. Nevertheless, considerable development is still necessary, and these results must be validated in larger numbers of participants.

Expanding BCI Usability and Availability

Toward an international consensus on user characterization and BCI outcomes in settings of daily living

Organizers: Mariska Vansteensel (UMC Utrecht) and Nataliya Kosmyna (Massachusetts Institute of Technology)
Additional Presenters: Andrew Geronimo (Department of Neurosurgery, Penn State College of Medicine, Hershey, PA, USA), Katya Hill (AAC-BCI iNNOVATION LAB, University of Pittsburgh, Pittsburgh, PA, USA), Theresa Vaughan (National Center for Adaptive Neurotechnologies, Stratton VA Medical Center, Albany, NY, USA)

BCI research is growing fast, and implantable and non-invasive communication-BCIs are being introduced to people with significant motor disability for independent use in daily living situations [e.g., 42, 278, 279-285], allowing end-users to participate in research and development experiments and provide critical input into iterative user-centered design [286]. Such studies are crucial for the development of usable communication-BCIs and for their eventual widespread implementation to resolve the communication problems of people with diseases such as amyotrophic lateral sclerosis. However, most studies include only limited numbers of participants. Since the target user population for communication-BCIs is relatively small [287], large studies may not actually be possible. For translation of communication-BCIs to practical use, it is therefore essential to compare results across studies and in this way learn about environmental and participant/user characteristics affecting BCI performance [e.g., 288, 289, 290] and the different usability perspectives of users, caregivers and other stakeholders. Such comparison will strongly benefit from standardized reporting about users/participants and their environment, and from the use of similar metrics to assess BCI performance and outcome [291]. This workshop was designed to initiate a consensus list of reporting recommendations, specifically directed at the use of communication-BCIs in the daily life settings of people with significant motor disability. After brief presentations to introduce the topics of discussion [196, 292-301], workshop participants shared their experiences and built consensus in breakout rooms. Key outcomes of these discussions include:

1. **Standardization is hard.** Standardization is a hard and complex task. Part of this complexity comes from the different focus areas of experiments designed by different disciplines.
2. **Age group matters.** Adult and pediatric BCI users need different training procedures and different primary outcome measures. But researchers need as much comparison as possible.

3. **Meeting users' end goals is paramount.** For any system to be introduced in their environment, end-users should be strongly involved in BCI design, goal setting, and outcome measure selection. Even existing standard metrics for reporting BCI system performance must be adapted to the goals of the end-user.
4. **Needs of primary users and their caregiver(s) may be different.** A BCI has multiple types of end-users and researchers must report on how well a BCI meets the needs and goals of both primary and secondary (e.g. caregivers) users.
5. **Different tasks produce different outcomes.** BCI outcome measures should consider the importance of each task to be conducted with the BCI, as well as the desired and accomplished frequency of conducting each task.
6. **Fatigue strongly affects BCI performance.** Both cognitive and physical fatigue need to be assessed and reported on.
7. **Medication can affect brain signals.** The effect of medication should not be underestimated, but medication use is seldom reported in papers.

As our next steps, we plan to engage in the bigger discussion about standardization, to collect more input from BCI researchers, and to use all collected information for a formal publication on reporting recommendations related to user characterization and outcome measures for the use-case of communication-BCIs in settings of daily living.

On the need of good practices and standards for Benchmarking Brain-Machine Interfaces

Organizer: Ricardo Chavarriaga (Zurich University Applied Sciences, ZHAW Switzerland)
Additional Presenters: Paul Sajda (Columbia University, USA), José Contreras-Vidal (IUCRC BRAIN, University of Houston, USA), Luigi Bianchi ("Tor Vergata" University of Rome, Italy), Zach McKinney (Scuola Superiore Sant'Anna, Italy), Laura Y. Cabrera (The Pennsylvania State University, USA)

Translating Brain-Machine Interface (BMI) systems onto real applications requires accepted, well-defined criteria to assess their effectiveness, usability, and safety. Benchmarking, specification, and performance evaluation are perceived as main priorities for standardization in the field [291, 302, 303]. This workshop discussed translational challenges, and ethical issues of BMI systems, as well as existing initiatives to address them.

The Future Neural Therapeutics technology roadmap [304] analyzes closed-loop neurotechnologies aimed at treating movement disorders and neurological diseases. This document summarizes the state of the art and identifies key technological challenges required to successfully develop a new generation of these technologies, including computational power, robustness and safety, usability and appropriate regulatory frameworks. As BMIs approach commercial availability, attention must be paid to concerns generated by the possibility of repurposing, misusing, or maliciously using consumer-oriented neurotechnology. These concerns include overstated claims on their efficacy or the influence of neurotechnology in markets related to employment or cognitive enhancement [305-307]. Moreover, widespread use of consumer-oriented technology can lead to

indiscriminate collection of neural data or user harm due to maladaptive processes triggered by neurostimulation devices.

The neuroethics subcommittee of the IEEE Brain Initiative focuses on the ethical and societal issues related to research and development of neurotechnologies They developed the IEEE Neuroethics Framework (https://brain.ieee.org/publications/ieee-neuroethics-framework/), a collective effort to evaluate the ethical, legal, social, and cultural issues that arise with the deployment of neurotechnologies and provide explicit guidance on how to address them. The framework is organized as a matrix that covers existing and emerging neurotechnologies for both current and foreseen applications. This framework is conceived as a living document that will evolve with the technology. Participation in this effort is open to interested participants.

Despite the large number of BMI publications, it is seldom possible to evaluate, verify or compare published results. Meta-analyses showed that a significant number of BCI publications lack necessary information [308, 309]. However, two standardization activities are addressing this issue. The IEEE Standards Working Group P2794: *Reporting Standard for in vivo Neural Interface Research* (RSNIR) (https://sagroups.ieee.org/2794/) aims to improve the transparency, interpretability, and replicability of neural interface research by specifying a set of technological and methodological characteristics to be reported in scientific literature and technical documentation.

They recently published a set of preliminary requirements for implantable neural interfaces [310] and are seeking broad community input and participation to ensure the Standard reflects the needs of a more diverse range of neuroscience and neurotechnology stakeholders, including device regulators, funding officers, clinicians, and end users. Information on providing such input can be found through the working group website. Another standardization project, IEEE P2731: *Standard for a Unified Terminology for Brain-Computer Interfaces (BCI)* (https://sagroups.ieee.org/2731/) aims at developing a comprehensive BCI lexicography and a functional model of BCI systems [311-313]. It is also working on identifying the required information to be stored in BCI files to enable efficient sharing of data and tools among stakeholders [314]. These activities can contribute to the development of standard experimental and usage protocols, benchmarking procedures, and increased interoperability of neurotechnology systems.

Overall, this workshop highlighted the need to continuously evaluate the state-of-the-art and the implications of neurotechnologies. This requires multi-stakeholder, anticipatory processes for developing appropriate tools -including ethical and technical guidelines, standards, and regulatory instruments- that allow translation of neurotechnologies for both consumer and medical applications [315-317].

Lessons from successfully implanted neurotechnology

Organizer: Erik Aarnoutse (Brain Center, University Medical Center Utrecht)
Additional Presenters: Fabien Sauter-Starace (CEA, LETI, Clinatec, University of Grenoble); Leigh Hochberg (Brown University; Massachusetts General Hospital; Providence VA Medical Center), RI Aysegul Gunduz (J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida)

Over the last 16 years, various clinical trials of implantable neurotechnology in humans have demonstrated successful applications. This technology has enabled users to move arms [318, 319], walk [108], and communicate [42] and has also alleviated disease symptoms [61]. Clinical trials require a great deal of effort but are an important and informative step along the route to wide availability of neurotechnology for users in need.

The route from design to clinical trial was illustrated by the Wimagine implant to operate an exoskeleton [108]. First, the medical needs of people with quadriplegia were combined with the neurosurgical requirements: no transcutaneous connection, no limit to battery lifetime and limited invasiveness. This created design choices of wireless data transmission, inductive charging, and epidural ECoG electrodes. Technical requirements were a trade-off between wishes and constraints. Animal studies assessed signal stability [320]. Regulatory compliance to the EU Medical Device Regulation meant proving compliance to ISO standards for quality management and standards for mechanical, electrical, and thermal safety, biocompatibility, and software. The clinical trial with bilateral implants has enrolled two patients so far [108]. Training was progressive by adding more complexity in the adaptive machine learning algorithm, from brain switch to 3D + pronation/supination [321]. The signal proved to be stable over months. The exoskeleton was only used in the laboratory.

The 17 years of BCI research with penetrating multi-electrode arrays produced many lessons [318]. Participants are colleagues, but also customers. They request new features (user needs), which are added to the design [196, 322]. The participants' motive is to advance science, they do not expect gain for themselves. However, the obligation of the field is to give users gain in daily life as soon as technology allows it [42, 285]. Neuroethics is important here. Hardware advances ease the technical constraints making neural data ever easier to gather and use.

With the entry of industry in this field, the question of the role of academia becomes more important, where academia is better equipped to ask fundamental (hypothesis based) questions of neuroscience. Development is important but is not easy to publish. Mainly, academia investigates (hardware agnostic) decoding principles.

A good example of the input of academic expertise is seen in the use of cortical ECoG recordings as part of essential tremor DBS therapy [61]. This cross-field input produced knowledge on biomarkers both for fundamental questions and treatment efficacy. Here, user needs for individualized therapy, reduction of side effects [323], and increased battery life were addressed. The research triggered a new hardware design that reduced stimulation artefacts.

So, academia provides design input (user needs, technical requirements, decoding principles) for future neurotechnology for home use. Academia seeks to create knowledge, optimize designs, and provide a foundation of information that can support translation of BCI to commercial availability. We have also identified barriers that must be overcome for home use (wireless link, power constraints, limits on the number of electrodes, portability, larger scale manufacturing). Overcoming these barriers requires more time and money than academia has, but

the generation of this knowledge by academic reduces the risk for industry and thus advances the likelihood that BCI will become widely, commercially available.

Next steps for practically useful BCI ethics

Organizer: Brendan Allison (UC San Diego)
Additional Presenters: Pim Haselager (Radboud University Nijmegen), Dr. Sonja Kleih-Dahms, (University of Würzburg), Donatella Mattia (Fondazione Santa Lucia, IRCCS)

This workshop was designed not for review or abstract academic discourse, but to develop practical next steps for BCI-related ethical issues. The organizers briefly presented examples of these issues [324-328] to promote discussion.

A public database of ethical use cases was proposed to raise awareness with an associated forum where people could share their perspectives on each case. The ethical use cases could also help professors and others who want to teach BCI ethics. Further discussion and development of ethical use cases would benefit from an ongoing collaborative effort, perhaps via online seminars, to develop a framework, assign people to develop different use cases, and create an online database. These efforts might be hosted by the BCI Society.

An immediate ethical concern is that research study participants do not usually keep the devices used in the study. Thus, people with disabilities may regain the ability to communicate or control a device with an experimental BCI, but then lose that ability when their study participation ends. Workshop contributors agreed that this is a serious and currently unresolved problem. Most funding sources do not support leaving devices with patients, nor providing ongoing technical support. However, several researchers include such considerations in their research plans. Possible next steps include raising awareness of this problem (such as through an online forum, survey, paper, or approaching journalists) and further engagement of funding organizations.

The rise of “Big BCI” through the recent initiation of BCI projects by high-profile companies creates its own set of ethical concerns. Workshop participants desired collaboration between the huge companies working on BCIs and the existing BCI community on efforts such as an online workshop or paper. This step was hoped to foster joint work on proposed ethical guidelines and regulatory issues.

Another concern comes from the many online articles and videos with misinformation about BCIs from different groups, including some manufacturers, neurofeedback practitioners, enthusiasts, and conspiracy theorists. Of course, such misinformation will continue indefinitely to some extent, but might be reduced through next steps such as publicly commenting on inaccuracies and producing and promoting high-quality information about BCIs. Indeed, some for-profit and non-profit entities do provide good BCI content. The ongoing increase in online BCI-related classes, conferences, workshops, competitions, and other activities has led to ample recorded material from reputable organizers and speakers that is usually available for free.

Many participants had seen online postings from, and/or been directly contacted by, people who believe that they are being involuntarily mind-controlled by a BCI or a similar device. A few participants reported trying to direct such

persons to appropriate mental health professionals, but without apparent success. Next steps at this time are not obvious aside from a possible paper or position statement with suggested guidelines, developed with mental health experts.

The workshop focused on specific, actionable next steps to raise awareness of ethical issues in BCI and further engage relevant groups through workshops, papers, online discussions and a database of use cases and surveys [329-331].

Brain-Computer Interfaces for Human Enhancement

Organizer: Davide Valeriani (Neurable Inc.)
Additional Presenters: Riccardo Poli (University of Essex), Maryam Shanechi (University of Southern California), Hasan Ayaz (Drexel University), Nataliya Kosmyna (MIT Media Lab), Yannick Roy (NeuroTechX), Marcello Ienca (ETH Zurich)

This workshop highlighted recent advances in BCI technologies that go beyond clinical applications and instead focus on augmenting human capabilities. The workshop brought together neuroscientists, engineers, neuro-ethicists, entrepreneurs and researchers at the cutting-edge of BCI development for human augmentation. Discussion focused on current trends and future prospects, as well as the critical role played by international communities such as NeuroTechX in educating and stimulating interest in BCI and neurotechnologies.

BCIs for cognitive human augmentation are intended to improve the process of acquiring knowledge and communicating with other individuals [332]. Passive BCIs can enhance individual decision-making in target detection by recognizing event-related potentials [333] or aggregating brain activity from multiple people [334]. Collaborative BCIs can also decode decision confidence from brain activity and use it to weigh individual opinions, leading to significant improvements in group performance in a variety of tasks [335-337]. These BCIs can also facilitate human-machine teaming in face recognition [338].

Combining brain recording (e.g., EEG, fNIRS) and stimulation (e.g., tDCS, TMS) improves processing speed [339] and spatial working memory [340], and introduces novel communication forms, such as brain-to-brain communication [341]. Moreover, it enables the development of BCIs capable of regulating abnormal mental states, with direct applications in the treatment of mental disorders [342, 343].

BCIs and other wearables support studying the brain in complex environments and diverse domains, a research field called neuroergonomics [344]. Advances in recording technologies, such as EEG and fNIRS, enable study in operational and realistic settings to monitor cognitive function, improve human-to-human communication, and enhance human-machine interaction [345]. Moreover, the integration of brain recordings with other physiological signals can provide biofeedback to users through audio, light, or haptic inputs, promoting performance, attention, and overall well-being [346]. These hybrid, multimodal BCIs will also help increase the reliability, accuracy, and commercial potential of non-invasive BCIs, which can be limited by the low signal-to-noise ratio of non-invasive neural recordings. Yet to implement multimodal BCIs we need to identify relationships

between modalities and develop new techniques to integrate neural recordings at different scales.

While neuroscience and neuro-engineering have shown that it is technically possible to develop BCIs that augment human capabilities in a variety of domains, neuro-ethicists are working to identify which applications are morally desirable [316]. Two main ethical principles should guide the development of BCIs for human augmentation: (1) cognitive liberty, which protects the rights of individuals to make free and competent decisions on using such devices, and (2) fair and equitable access to enhancement, which ensures they are available to everyone, regardless of race, gender or socioeconomic status. As with all biomedical devices, safety and data privacy are key pillars to make these devices ethically acceptable.

Overall, the workshop showcased the tremendous advantages of expanding BCIs from assistive devices to technologies for human enhancement, with a variety of potential applications. The most promising approaches seem to be the fusion of different physiological signals and integration with artificial intelligence, with a continuous awareness of the ethical challenges of enhancement applications.

Brain-Computer Interfaces for outside the lab: Neuroergonomics for human-computer interaction, education and sport

Organizers: Antonia Thelen (eemagine Medical Imaging Solutions GmbH, Berlin, Germany)
Additional Presenters: Fabien Lotte, (Inria Bordeaux Sud-Ouest); Camille Jeunet (CNRS, Bordeaux Neurocampus); Frédéric Dehais (ISAE-SUPAERO, Toulouse); Patrique Fiedler (TU Ilmenau, Ilmenau); Martijn Schreuder (ANT-Neuro, Enschede)

Traditionally, BCI research has been bound to the investigation of perceptual, cognitive and motor processes within stationary, hardware-intensive laboratory setups. While these studies provide intriguing real-time insights into such processes, the translation of these findings into real-world brain interactions is limited. The emergence of lightweight, high-density EEG solutions has permitted the extension of BCI applications into mobile setups within real-world situations. Use of high-density EEG enables the simultaneous utilization of different sensor configurations, providing greater adaptability with a single hardware setup.

This workshop focused on the efforts undertaken towards the instrumentalization of EEG and specifically BCI techniques within the field of neuroergonomics. The panel comprised experts who strove to provide methodological strategies to facilitate the transition of BCI applications into real-world and/or every-day settings. First, advances and current limitations of existing solutions were discussed. Second, an outlook upon possible new technological and methodological innovations was presented which could provide new avenues of interacting with the world by implementing systems with an explicit awareness of the concepts of embodied cognition. Embodied cognition, as described in [347], acknowledges that physical elements of the world are often integrated seamlessly into our cognitive processes in a way not easily captured by static diagrams with separate boxes for sensory inputs and physical outputs. Instead, cognition happens in conjunction and in parallel with the sensorimotor loops that provide

interactions with the world. Various neuroergonomics applications of BCI use outside the lab were also discussed, including evaluating 3D User Interfaces [348], Sport Science [349, 350] and Aviation [351].

Specifically, the robustness of signal processing methods used by BCI classifiers was discussed. How to apply such algorithms reliably across a large variety of application fields and how to make them cope with inter- and intra-individual variability is still a topic under investigation [352]. The contribution of state-of-the-art, lightweight, dry sensors resulting in varying signal-to-noise ratios and their impact upon such signal processing algorithms was highlighted [353, 354]. Moreover, the tradeoff between laboratory-based and real-world applications was discussed with regards to sensor application within these fundamentally different environments [350, 355]. Lastly, discussion focused on difficulties encountered when translating BCI-based interventions across different demographics, specifically differences in cognitive states and/or perceptual processes that were investigated within a research context or focused on clinical/therapeutic interventions.

Taken together, the workshop provided an overview of current advances made within the field of neuroergonomics.

Brain-Computer Interfaces for Art, Entertainment, and Domestic Applications

Organizer: Anton Nijholt (University of Twente)
Additional Presenters: Christoph Guger (g;tec medical engineering GmbH); Elisabeth Hildt (Illinois Institute of Technology); Erika Mondria (University of Art and Design); Ellen Pearlman (Massachusetts Institute of Technology); Stephanie Scott (Colorado State University); Aleksander Valjamae (Tallinn University)

BCI technology enables neurophysiological data from an individual user's affective and mental state to be used for online adaption of system and interaction methods [356]. Artistic, domestic, or entertainment use of such information shift the focus from efficiency to the importance of affect in social and playful interactions such as in family, community, playful, and artistically challenging situations. This workshop addressed the use of BCI for artistic, entertainment, educational, and health applications.

BCI has been used for many artistic applications [357-359]. In general, artistic projects reduce inhibitions and encourage people to engage with unfamiliar technologies such as BCI. Synergies of design, art, and research have shown interesting results which may also enrich clinical settings.

BCIs have been used for creative arts therapy [360, 361] as part of a conceptual framework bringing together several disciplines for researching the expansion of treatment modalities in the intersection of art, technology, and therapeutics. A recent insight is that a post-phenomenological approach towards human-technology interaction and technological artifacts in general will be useful when applied to BCI for therapy, art, and creative expression. In this approach user-specific needs for enabling self-expression are integrated in a transdisciplinary design perspective on meaningful and self-expressive

communication exploring brain activity underlying artistic creation and using neurofeedback research [362].

The BR41N.IO BCI Hackathon series, now in its 5th year [363, 364], provides opportunities for team-based development of new BCI applications within 24 hours. During the 2021 BCI & Neurotechnology Spring School, 321 developers, artists, programmers, and hackers participated in 38 teams and created many interesting and cutting-edge new applications or improved the signal processing of BCI data sets.

In neurotheatre and neurocinema research [365, 366], new media art and neurotechnologies allow for co-creation between actors, director, and audience to shape a performance by emotional experiences using BCI and other sensors and multisensory actuators. From a research perspective, neurotheatre can be seen as a novel integrative research environment for prototyping and exploring new social neuroscience paradigms, like collective decision making or shared affective experiences. From a societal perspective, the fusion of science, technology, and arts allows for so-called design fiction, a design practice aiming at exploring and criticizing possible futures by creating speculative, and often provocative, scenarios narrated through designed artifacts.

Affective brain-computer music [367, 368] Interface applications use affective BCIs for music-making and music listening. Given recent developments in direct-to-consumer devices (wearable BCIs, headphone sensors) and music streaming services these BCI applications aim at influencing the user's affective state (mood enhancement) by individualized music choices. Exaggerated claims about capabilities, increasing dependency on technology and limiting one's own capabilities, and privacy issues arising from long-term monitoring of a user's affective state are pitfalls related to a potential future, relatively widespread use of EEG-based affective brain-computer music interfaces in entertainment contexts [369].

A brain opera called "Noor" provides an example that combines these concepts through the use of artificial intelligence (AI). In "Noor", biometric variables, including BCI are integrated with natural language processing and machine learning. In the near-future, such integrated systems will be tasked with more responsibilities relating to many aspects of human congress, often with confusing legal oversight and minimal accountability, potentially leading to scenarios enforcing dystopic digital societies of control [370-372].

The workshop discussions revealed consensus about the benefit of the joint effort of art and science research for BCI research in general and the acceptance of BCI for the general public.

Conclusion

Together, these workshops provide foundational information, explore diverse applications for different populations, and further develop big picture ideas for new frontiers of BCI use. Many of these ideas will be further developed in the workshops of the planned in-person Ninth International Brain-Computer Interface Meeting, currently scheduled for June 7-10th, 2022 in the Sonian Forest, Brussels, Belgium.

Acknowledgements:

Overall Acknowledgements

The authors thank the National Institute on Deafness and other Communication Disorders (NIDCD) and the National Institute of Neurological Disorders and Stroke (NINDS) in the National Institutes of Health (NIH) of the United States, the National Science Foundation (NSF), and the Wellcome Foundation for support assisting student participation in the BCI Meeting. We also thank the Research Foundation Flanders for support of the Meeting. The opinions expressed are those of the authors and do not reflect the views of any funding agency that may have supported work presented at the BCI Meeting or in the individual workshops.

The organizers of the workshop thank their presenters and participants for their presentations and thoughtful discussion. And, of course, we thank the many and varied funding sources that supported the research presented in the workshops. The workshop organizers also thank the members of the Program Committee for the Eighth International Brain-Computer Interface Meeting: José del R. Millán, Chuck Anderson, Guy Cheron, Jennifer Collinger, Marc Van Hulle, Dean Krusienski, Steven Laureys, Marc Slutzky, and Mariska Vansteensel.

Individual Workshop Acknowledgements

The work presented in *Biomimetic approaches to restore somatosensation* was supported by the NIH (R01 NS095251, U01 NS098975, UH3NS107714, U01NS108922), the National Science Foundation (IOS 1150209), the Kimberley Clark Foundation, EU Grant FET 611687, the Swiss National Science Foundation National Competence Center in Research in Robotics, the Bertarelli Foundation, the T&C Chen Brain-Machine Interface Center, the USC Neurorestoration Center, DARPA (NC66001-15-C-4041, N66001-16-C4501), and the US Department of Veterans Affairs (C3819C).

The workshop *Brain-computer interfaces for the assessment of patients with disorders of consciousness* was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 841116 (RS).

The workshop *Brain-Computer Interfaces for Human Enhancement* was supported by the US DOD Bilateral Academic Research Initiative program (W911NF1810434). The funders had no role in workshop organization, decision to publish, or preparation of the manuscript.

The workshop *From Speech Decoding to Speech* was supported by NSF (1608140/1902395/2011595) and BMBF (01GQ1602) and as part of the NSF/NIH/BMBF Collaborative Research in Computational Neuroscience Program (CRCNS).

The workshop *On the need of good practices and standards for Benchmarking Brain-Machine Interfaces* was supported by the IEEE Brain Initiative, the IEEE Standards Association Industry Connections Program, the Confederation of Laboratories for Artificial Intelligence in Europe (CLAIRE), and National Science

Foundation Awards #1650536 I/UCRC for Building Reliable Advances and Innovation in Neurotechnology (IUCRC BRAIN Center) and PFI # 1827769.

The workshop *The design of effective BCIs for children* was supported by the National Institute of Biomedical Imaging and Bioengineering of the NIH (P41 EB018783), resources at the US Department of Veterans Affairs Stratton VA Medical Center, the New York State Spinal Cord Injury Board, the Alberta Children's Hospital Foundation, and the Alberta Children's Hospital Research Institute (ACHRI).

The workshop *Towards the decoding of neural information for motor control: present and future approaches* was partly supported by the European Research Council (ERC-CoG-2015 681231 'Feel Your Reach')

Conflict of Interest

CG is the owner and CEO of g.tec medical engineering GmbH.

DV is an employee of Neurable, Inc

GMP is on the board of directors of the BCI Society

JEH is a co-Editor-in-Chief of *Brain-Computer Interfaces*, on the board of directors of the BCI Society, and has a pending patent on a BCI application used in one of the referenced papers.

MV is on the board of directors of the BCI Society

RG is a member of the scientific advisory board for Braingrade GmbH.

Bibliography

[1] J. R. Wolpaw, N. Birbaumer, W. J. Heetderks, D. J. McFarland, P. H. Peckham, G. Schalk, E. Donchin, L. A. Quatrano, C. J. Robinson, and T. M. Vaughan, "Brain-computer interface technology: a review of the first international meeting," *IEEE transactions on rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society,* vol. 8, no. 2, pp. 164-173, 2000.

[2] T. M. Vaughan, W. J. Heetderks, L. J. Trejo, W. Z. Rymer, M. Weinrich, M. M. Moore, A. Kubler, B. H. Dobkin, N. Birbaumer, E. Donchin, E. W. Wolpaw, and J. R. Wolpaw, "Brain-computer interface technology: a review of the Second International Meeting," *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society,* vol. 11, no. 2, pp. 94-109, 2003.

[3] T. M. Vaughan, and J. R. Wolpaw, "The Third International Meeting on Brain-Computer Interface Technology: making a difference," *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society,* vol. 14, no. 2, pp. 126-127, 2006.

[4] T. M. Vaughan, and J. R. Wolpaw, "Special issue containing contributions from the Fourth International Brain-Computer Interface Meeting," *Journal of*

neural engineering, vol. 8, no. 2, pp. 020201-2560/8/2/020201. Epub 2011 Mar 24, 2011.

[5] J. E. Huggins, C. Guger, B. Allison, C. W. Anderson, A. Batista, A. M. Brouwer, C. Brunner, R. Chavarriaga, M. Fried-Oken, A. Gunduz, D. Gupta, A. Kübler, R. Leeb, F. Lotte, L. E. Miller, G. Müller-Putz, T. Rutkowski, M. Tangermann, and D. E. Thompson, "Workshops of the Fifth International Brain-Computer Interface Meeting: Defining the Future," *Brain-Computer Interface Journal,* vol. 1, no. 1, pp. 27-49, 2014.

[6] J. E. Huggins, and J. R. Wolpaw, "Papers from the fifth international brain-computer interface meeting. Preface," *Journal of neural engineering,* vol. 11, no. 3, pp. 030301-2560/11/3/030301. Epub 2014 May 19, 2014.

[7] J. J. Daly, and J. E. Huggins, "Brain-computer interface: current and emerging rehabilitation applications," *Archives of Physical Medicine and Rehabilitation,* vol. 96, no. 3 Suppl, pp. S1-7, 2015.

[8] J. E. Huggins, C. Guger, M. Ziat, T. O. Zander, D. Taylor, M. Tangermann, A. Soria-Frisch, J. Simeral, R. Scherer, R. Rupp, G. Ruffini, D. K. R. Robinson, N. F. Ramsey, A. Nijholt, G. Muller-Putz, D. J. McFarland, D. Mattia, B. J. Lance, P. J. Kindermans, I. Iturrate, C. Herff, D. Gupta, A. H. Do, J. L. Collinger, R. Chavarriaga, S. M. Chase, M. G. Bleichner, A. Batista, C. W. Anderson, and E. J. Aarnoutse, "Workshops of the Sixth International Brain-Computer Interface Meeting: brain-computer interfaces past, present, and future," *Brain Comput Interfaces (Abingdon),* vol. 4, no. 1-2, pp. 3-36, 2017.

[9] J. E. Huggins, G. Müller-Putz, and J. R. Wolpaw, "The Sixth International Brain-Computer Interface Meeting: Advances in Basic and Clinical Research," *Brain Comput Interfaces (Abingdon),* vol. 4, no. 1-2, pp. 1-2, 2017.

[10] J. E. Huggins, C. Guger, E. Aarnoutse, B. Allison, C. W. Anderson, S. Bedrick, W. Besio, R. Chavarriaga, J. L. Collinger, A. H. Do, C. Herff, M. Hohmann, M. Kinsella, K. Lee, F. Lotte, G. Muller-Putz, A. Nijholt, E. Pels, B. Peters, F. Putze, R. Rupp, G. Schalk, S. Scott, M. Tangermann, P. Tubig, and T. Zander, "Workshops of the Seventh International Brain-Computer Interface Meeting: Not Getting Lost in Translation," *Brain Comput Interfaces (Abingdon),* vol. 6, no. 3, pp. 71-101, 2019.

[11] J. E. Huggins, and M. W. Slutzky, "Articles from the Seventh International Brain-Computer Interface Meeting," *Brain-Computer Interfaces,* vol. 6, no. 4, pp. 103-105, 2019/10/02, 2019.

[12] W. G. Besio, K. Koka, R. Aakula, and W. Dai, "Tri-polar concentric ring electrode development for laplacian electroencephalography," *IEEE Trans Biomed Eng,* vol. 53, no. 5, pp. 926-33, May, 2006.

[13] K. Koka, and W. G. Besio, "Improvement of spatial selectivity and decrease of mutual information of tri-polar concentric ring electrodes," *J Neurosci Methods,* vol. 165, no. 2, pp. 216-22, Sep, 2007.

[14] X. Liu, O. Makeyev, and W. Besio, "Improved spatial resolution of electroencephalogram using tripolar concentric ring electrode sensors," *J Sensors*, 2020.

[15] W. G. Besio, I. E. Martínez-Juárez, O. Makeyev, J. N. Gaitanis, A. S. Blum, R. S. Fisher, and A. V. Medvedev, “High-Frequency Oscillations Recorded on the Scalp of Patients With Epilepsy Using Tripolar Concentric Ring Electrodes,” *IEEE J Transl Eng Health Med,* vol. 2, pp. 2000111, 2014.
[16] C. Toole, I. Martinez-Juarez, J. Gaitanis, A. Blum, S. Sunderam, L. Ding, J. Dicecco, and W. Besio, “Source localization of high frequency activity in tripolar electroencephalography of epilepsy patients,” *Epilepsy and Behavior,* vol. 101, pp. 106519, 2019.
[17] W. Besio, H. Cao, and P. Zhou, “Application of Tripolar Concentric Electrodes and Pre-Feature Selection Algorithm for Brain-Computer Interface,” *IEEE Trans Neural Systems & Rehab Eng,* vol. 16, no. 2, pp. 191-194, 2008.
[18] Y. Boudria, A. Feltane, and W. Besio, “Significant improvement in one dimensional cursor control using Laplacian electroencephalography over electroencephalography,” *J. Neural Engineering,* vol. 11, pp. 35014, 2014.
[19] S. I. Alzahrani, and C. W. Anderson, “A Comparison of Conventional and Tri-Polar EEG Electrodes for Decoding Real and Imaginary Finger Movements from One Hand,” *Int J Neural Syst*, pp. 2150036, Jul 10, 2021.
[20] W. Besio, V. Sharma, and J. Spaulding, “The effects of concentric ring electrode electrical stimulation on rat skin,” *Ann Biomed Eng,* vol. 38, no. 3, pp. 1111-8, Mar, 2010.
[21] G. Rogel-Salazar, H. Luna-Munguía, K. E. Stevens, and W. G. Besio, “Transcranial focal electrical stimulation via tripolar concentric ring electrodes does not modify the short- and long-term memory formation in rats evaluated in the novel object recognition test,” *Epilepsy Behav,* vol. 27, no. 1, pp. 154-8, Apr, 2013.
[22] W. G. Besio, K. Koka, and A. J. Cole, “Effects of noninvasive transcutaneous electrical stimulation via concentric ring electrodes on pilocarpine-induced status epilepticus in rats,” *Epilepsia,* vol. 48, no. 12, pp. 2273-9, Dec, 2007.
[23] W. G. Besio, O. Makeyev, A. Medvedev, and K. Gale, “Effects of transcranial focal electrical stimulation via tripolar concentric ring electrodes on pentylenetetrazole-induced seizures in rats,” *Epilepsy Res,* vol. 105, no. 1-2, pp. 42-51, Jul, 2013.
[24] O. Makeyev, X. Liu, H. Luna-Munguia, G. Rogel-Salazar, S. Mucio-Ramirez, Y. Liu, Y. L. Sun, S. M. Kay, and W. G. Besio, “Toward a noninvasive automatic seizure control system in rats with transcranial focal stimulations via tripolar concentric ring electrodes,” *IEEE Trans Neural Syst Rehabil Eng,* vol. 20, no. 4, pp. 422-31, Jul, 2012.
[25] O. Makeyev, H. Luna-Munguia, G. Rogel-Salazar, X. Liu, and W. G. Besio, “Noninvasive transcranial focal stimulation via tripolar concentric ring electrodes lessens behavioral seizure activity of recurrent pentylenetetrazole administrations in rats,” *IEEE Trans Neural Syst Rehabil Eng,* vol. 21, no. 3, pp. 383-90, May, 2013.
[26] W. G. Besio, X. Liu, L. Wang, A. V. Medvedev, and K. Koka, “Transcutaneous focal electrical stimulation via concentric ring electrodes

reduces synchrony induced by pentylenetetrazole in beta and gamma bands in rats," *Int J Neural Syst,* vol. 21, no. 2, pp. 139-49, Apr, 2011.

[27] A. Valdes-Cruz, B. Villasana-Salazar, B. Williams, D. Martinez-Vargas, V. M. Magdaleno-Madrigal, S. Almazan-Alvarado, and W. G. Besio, "Transcranial focal electrical stimulation via concentric ring electrodes in freely moving cats: Antiepileptogenic and postictal effects," *Exp Neurol,* vol. 320, pp. 113012, Oct, 2019.

[28] W. Besio, M. Cuellar-Herrera, H. Luna-Munguia, S. Orozco-Suarez, and L. Rocha, "Effects of transcranial focal electrical stimulation alone and associated with a sub-effective dose of diazepam on pilocarpine-induced status epilepticus and subsequent neuronal damage in rats," *Epilepsy Behav,* vol. 28, no. 3, pp. 432-6, Sep, 2013.

[29] C. E. Santana-Gomez, D. Alcantara-Gonzalez, H. Luna-Munguia, I. Banuelos-Cabrera, V. Magdaleno-Madrigal, R. Fernandez-Mas, W. Besio, and L. Rocha, "Transcranial focal electrical stimulation reduces the convulsive expression and amino acid release in the hippocampus during pilocarpine-induced status epilepticus in rats," *Epilepsy Behav,* vol. 49, pp. 33-9, Aug, 2015.

[30] D. Perez-Perez, J. L. Castaneda-Cabral, S. Orozco-Suarez, J. Sotelo, W. Besio, and L. Rocha, "Noninvasive transcranial focal stimulation affects the convulsive seizure-induced P-glycoprotein expression and function in rats," *Epilepsy Behav,* vol. 115, pp. 107659, Feb, 2021.

[31] N. Hansen, A. K. Kahn, D. Zeller, Z. Katsarava, C. Sommer, and N. Uceyler, "Amplitudes of Pain-Related Evoked Potentials Are Useful to Detect Small Fiber Involvement in Painful Mixed Fiber Neuropathies in Addition to Quantitative Sensory Testing - An Electrophysiological Study," *Front Neurol,* vol. 6, pp. 244, 2015.

[32] K. J. Oh, S. H. Kim, Y. H. Lee, J. H. Kim, H. S. Jung, T. J. Park, J. Park, and J. M. Shinn, "Pain-related evoked potential in healthy adults," *Ann Rehabil Med,* vol. 39, no. 1, pp. 108-15, Feb, 2015.

[33] O. S. Ozgul, C. Maier, E. K. Enax-Krumova, J. Vollert, M. Fischer, M. Tegenthoff, and O. Hoffken, "High test-retest-reliability of pain-related evoked potentials (PREP) in healthy subjects," *Neurosci Lett,* vol. 647, pp. 110-116, Apr 24, 2017.

[34] A. Papagianni, G. Siedler, C. Sommer, and N. Uceyler, "Capsaicin 8% patch reversibly reduces A-delta fiber evoked potential amplitudes," *Pain Rep,* vol. 3, no. 2, pp. e644, Mar, 2018.

[35] N. Uceyler, A. K. Kahn, D. Kramer, D. Zeller, J. Casanova-Molla, C. Wanner, F. Weidemann, Z. Katsarava, and C. Sommer, "Impaired small fiber conduction in patients with Fabry disease: a neurophysiological case-control study," *BMC Neurol,* vol. 13, pp. 47, May 24, 2013.

[36] N. Uceyler, D. Zeller, A. K. Kahn, S. Kewenig, S. Kittel-Schneider, A. Schmid, J. Casanova-Molla, K. Reiners, and C. Sommer, "Small fibre pathology in patients with fibromyalgia syndrome," *Brain,* vol. 136, no. Pt 6, pp. 1857-67, Jun, 2013.

[37] Z. Katsarava, I. Ayzenberg, F. Sack, V. Limmroth, H. C. Diener, and H. Kaube, "A novel method of eliciting pain-related potentials by

transcutaneous electrical stimulation," *Headache,* vol. 46, no. 10, pp. 1511-7, Nov-Dec, 2006.
[38] A. Khatoun, B. Asamoah, and M. Mc Laughlin, "Investigating the Feasibility of Epicranial Cortical Stimulation Using Concentric-Ring Electrodes: A Novel Minimally Invasive Neuromodulation Method," *Front Neurosci,* vol. 13, pp. 773, 2019.
[39] K. J. Miller, D. Hermes, and N. P. Staff, "The current state of electrocorticography-based brain-computer interfaces," *Neurosurg Focus,* vol. 49, no. 1, pp. E2, Jul, 2020.
[40] W. Wang, J. L. Collinger, A. D. Degenhart, E. C. Tyler-Kabara, A. B. Schwartz, D. W. Moran, D. J. Weber, B. Wodlinger, R. K. Vinjamuri, R. C. Ashmore, J. W. Kelly, and M. L. Boninger, "An electrocorticographic brain interface in an individual with tetraplegia," *PloS one,* vol. 8, no. 2, pp. e55344, 2013.
[41] P. Brunner, A. L. Ritaccio, J. F. Emrich, H. Bischof, and G. Schalk, "Rapid Communication with a "P300" Matrix Speller Using Electrocorticographic Signals (ECoG)," *Frontiers in neuroscience,* vol. 5, pp. 5, 2011.
[42] M. J. Vansteensel, E. G. M. Pels, M. G. Bleichner, M. P. Branco, T. Denison, Z. V. Freudenburg, P. Gosselaar, S. Leinders, T. H. Ottens, M. A. Van Den Boom, P. C. Van Rijen, E. J. Aarnoutse, and N. F. Ramsey, "Fully Implanted Brain-Computer Interface in a Locked-In Patient with ALS," *N Engl J Med,* vol. 375, no. 21, pp. 2060-2066, 11, 2016.
[43] F. Lotte, J. S. Brumberg, P. Brunner, A. Gunduz, A. L. Ritaccio, C. Guan, and G. Schalk, "Electrocorticographic representations of segmental features in continuous speech," *Frontiers in human neuroscience,* vol. 9, pp. 97, 2015.
[44] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature,* vol. 568, no. 7753, pp. 493-498, 2019.
[45] Q. Rabbani, G. Milsap, and N. E. Crone, "The Potential for a Speech Brain–Computer Interface Using Chronic Electrocorticography," *Neurotherapeutics,* vol. 16, no. 1, pp. 144-165, 2019.
[46] K. Dijkstra, P. Brunner, A. Gunduz, W. Coon, A. L. Ritaccio, J. Farquhar, and G. Schalk, "Identifying the Attended Speaker Using Electrocorticographic (ECoG) Signals," *Brain computer interfaces (Abingdon, England),* vol. 2, no. 4, pp. 161-173, 2015.
[47] K. J. Miller, G. Schalk, E. E. Fetz, M. den Nijs, J. G. Ojemann, and R. P. Rao, "Cortical activity during motor execution, motor imagery, and imagery-based online feedback," *Proc Natl Acad Sci U S A,* vol. 107, no. 9, pp. 4430-5, Mar, 2010.
[48] K. J. Miller, E. C. Leuthardt, G. Schalk, R. P. Rao, N. R. Anderson, D. W. Moran, J. W. Miller, and J. G. Ojemann, "Spectral changes in cortical surface potentials during motor movement," *The Journal of neuroscience : the official journal of the Society for Neuroscience,* vol. 27, no. 9, pp. 2424-2432, 2007.
[49] C. Kapeller, H. Ogawa, G. Schalk, N. Kunii, W. G. Coon, J. Scharinger, C. Guger, and K. Kamada, "Real-time detection and discrimination of visual

perception using electrocorticographic signals," *J Neural Eng,* vol. 15, no. 3, pp. 036001, Jun, 2018.

[50] K. J. Miller, G. Schalk, D. Hermes, J. G. Ojemann, and R. P. Rao, "Spontaneous Decoding of the Timing and Content of Human Object Perception from Cortical Surface Recordings Reveals Complementary Information in the Event-Related Potential and Broadband Spectral Change," *PLoS Comput Biol,* vol. 12, no. 1, pp. e1004660, Jan, 2016.

[51] E. C. Leuthardt, D. W. Moran, and T. R. Mullen, "Defining Surgical Terminology and Risk for Brain Computer Interface Technologies," *Front Neurosci,* vol. 15, pp. 599549, 2021.

[52] H. Ogawa, K. Kamada, C. Kapeller, R. Prueckl, F. Takeuchi, S. Hiroshima, R. Anei, and C. Guger, "Clinical Impact and Implication of Real-Time Oscillation Analysis for Language Mapping," *World Neurosurg,* vol. 97, pp. 123-131, Jan, 2017.

[53] C. Kapeller, M. Korostenskaja, R. Prueckl, P. C. Chen, K. H. Lee, M. Westerveld, C. M. Salinas, J. C. Cook, J. E. Baumgartner, and C. Guger, "CortiQ-based Real-Time Functional Mapping for Epilepsy Surgery," *J Clin Neurophysiol,* vol. 32, no. 3, pp. e12-22, Jun, 2015.

[54] H. Ogawa, K. Kamada, C. Kapeller, S. Hiroshima, R. Prueckl, and C. Guger, "Rapid and minimum invasive functional brain mapping by real-time visualization of high gamma activity during awake craniotomy," *World Neurosurg,* vol. 82, no. 5, pp. 912 e1-10, Nov, 2014.

[55] F. Kohler, C. A. Gkogkidis, C. Bentler, X. Wang, M. Gierthmuehlen, J. Fischer, C. Stolle, L. M. Reindl, J. Rickert, T. Stieglitz, T. Ball, and M. Schuettler, "Closed-loop interaction with the cerebral cortex: a review of wireless implant technology," *Brain-Computer Interfaces,* vol. 4, no. 3, pp. 146-154, 2017/07/03, 2017.

[56] D. R. Kramer, K. Lamorie-Foote, M. Barbaro, M. B. Lee, T. Peng, A. Gogia, G. Nune, C. Y. Liu, S. S. Kellis, and B. Lee, "Utility and lower limits of frequency detection in surface electrode stimulation for somatosensory brain-computer interface in humans," *Neurosurg Focus,* vol. 48, no. 2, pp. E2, Feb 1, 2020.

[57] G. Schalk, C. Kapeller, C. Guger, H. Ogawa, S. Hiroshima, R. Lafer-Sousa, Z. M. Saygin, K. Kamada, and N. Kanwisher, "Facephenes and rainbows: Causal evidence for functional and anatomical specificity of face and color processing in the human brain," *Proc Natl Acad Sci U S A,* vol. 114, no. 46, pp. 12285-12290, 11, 2017.

[58] J. Brice, and L. McLellan, "Suppression of intention tremor by contingent deep-brain stimulation," *Lancet,* vol. 1, no. 8180, pp. 1221-2, Jun 7, 1980.

[59] A. M. Lozano, and N. Lipsman, "Probing and regulating dysfunctional circuits using deep brain stimulation," *Neuron,* vol. 77, no. 3, pp. 406-24, Feb 6, 2013.

[60] R. Molina, M. S. Okun, J. B. Shute, E. Opri, P. J. Rossi, D. Martinez-Ramirez, K. D. Foote, and A. Gunduz, "Report of a patient undergoing chronic responsive deep brain stimulation for Tourette syndrome: proof of concept," *J Neurosurg,* vol. 129, no. 2, pp. 308-314, Aug, 2018.

[61] E. Opri, S. Cernera, R. Molina, R. S. Eisinger, J. N. Cagle, L. Almeida, T. Denison, M. S. Okun, K. D. Foote, and A. Gunduz, “Chronic embedded cortico-thalamic closed-loop deep brain stimulation for the treatment of essential tremor,” *Sci Transl Med,* vol. 12, no. 572, Dec 2, 2020.
[62] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, “BCI2000: a general-purpose brain-computer interface (BCI) system,” *IEEE transactions on bio-medical engineering,* vol. 51, no. 6, pp. 1034-1043, 2004.
[63] C. Guger, A. Schlogl, C. Neuper, D. Walterspacher, T. Strein, and G. Pfurtscheller, “Rapid prototyping of an EEG-based brain-computer interface (BCI),” *IEEE Trans Neural Syst Rehabil Eng,* vol. 9, no. 1, pp. 49-58, Mar, 2001.
[64] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Riemannian geometry applied to BCI classification," *Latent Variable Analysis and Signal Separation*, Lecture Notes in Computer Science Z. V. Vigneron V., Moreau E., Gribonval R., Vincent E., ed., pp. 629-636, Berlin, Heidelberg: Springer, 2010.
[65] A. Barachant, and M. Congedo, “A Plug & Play P300 BCI Using Information Geometry,” *arXiv,* vol. 1409.0107, 2014.
[66] E. K. Kalunga, S. Chevallier, and Q. Barthelemy, “Using Riemannian geometry for SSVEP-based Brain Computer Interface,” *arXiv:1501.03227 [cs, stat]*, 2015/02/24/, 2015.
[67] Y. Li, K. M. Wong, and H. D. Bruin, “Electroencephalogram signals classification for sleepstate decision - A riemannian geometry approach,” *IET Signal Processing,* vol. 6, no. 4, pp. 288-299, 2012/06//, 2012.
[68] C. Simar, A.-M. Cebolla, G. Chartier, M. Petieau, G. Bontempi, A. Berthoz, and G. Cheron, “Hyperscanning EEG and Classification Based on Riemannian Geometry for Festive and Violent Mental State Discrimination,” *Frontiers in Neuroscience,* vol. 14, pp. 1225, 2020, 2020.
[69] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, “Multiclass Brain–Computer Interface Classification by Riemannian Geometry,” *IEEE Transactions on Biomedical Engineering,* vol. 59, no. 4, pp. 920-928, 2012/04//, 2012.
[70] E. K. Kalunga, S. Chevallier, and Q. Barthélemy, "Transfer Learning for SSVEP-based BCI Using Riemannian Similarities Between Users." pp. 1685-1689.
[71] P. L. C. Rodrigues, C. Jutten, and M. Congedo, “Riemannian Procrustes Analysis: Transfer Learning for Brain–Computer Interfaces,” *IEEE Transactions on Biomedical Engineering,* vol. 66, no. 8, pp. 2390-2401, 2019/08//, 2019.
[72] L. Korczowski, M. Congedo, and C. Jutten, "Single-trial classification of multi-user P300-based Brain-Computer Interface using Riemannian geometry." pp. 1769-1772.
[73] A. Barachant. "MEG decoding using Riemannian Geometry and Unsupervised classification," http://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=024F0362243B7F26EF3D5B09D93C3BA3?doi=10.1.1.713.5131.

[74] A. Barachant, and R. Cycon, “Pushing the limits of BCI accuracy: Winning solution of the Grasp & Lift EEG challenge,” in 6th International Brain-Computer Interface Meeting, Pacific Grove, California, USA, 2016, pp. Paper 73.
[75] I. Horev, F. Yger, and M. Sugiyama, “Geometry-aware principal component analysis for symmetric positive definite matrices,” *Machine Learning,* vol. 106, no. 4, pp. 493-522, 2017/04/01/, 2017.
[76] M. Congedo, A. Barachant, and R. Bhatia, “Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review,” *Brain-Computer Interfaces,* vol. 4, no. 3, pp. 155-174, 2017/07/03, 2017.
[77] A. Barachant, A. Andreev, and M. Congedo, "The Riemannian Potato: an automatic and adaptive artifact detection method for online experiments using Riemannian geometry." pp. 19-20.
[78] Q. Barthélemy, L. Mayaud, D. Ojeda, and M. Congedo, “The Riemannian Potato Field: A Tool for Online Signal Quality Index of EEG,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 27, no. 2, pp. 244-255, 2019/02//, 2019.
[79] V. Jayaram, and A. Barachant, “MOABB: trustworthy algorithm benchmarking for BCIs,” *J Neural Eng,* vol. 15, no. 6, pp. 066011, Dec, 2018.
[80] P. Clisson, R. Bertrand-Lalo, M. Congedo, G. Victor-Thomas, and J. Chatel-Goldman, “Timeflux: an open-source framework for the acquisition and near real-time processing of signal streams,” in BCI 2019 - 8th International Brain-Computer Interface Conference, Graz, Austria, 2019.
[81] F. Lotte, C. Jeunet, R. Chavarriaga, L. Bougrain, D. E. Thompson, R. Scherer, M. R. Mowla, A. Kübler, M. Grosse-Wentrup, K. Dijkstra, and N. Dayan, “Turning negative into positives! Exploiting ‘negative’ results in Brain–Machine Interface (BMI) research,” *Brain-Computer Interfaces,* vol. 6, no. 4, pp. 178-189, 2019/10/02/, 2019.
[82] kaggle. "2018 Kaggle Machine Learning & Data Science Survey," 10 July 2021, 2021; https://kaggle.com/kaggle/kaggle-survey–2018.
[83] M. Congedo, A. Barachant, and A. Andreev, “A New Generation of Brain-Computer Interface Based on Riemannian Geometry
” *arXiv.org,* vol. ArXiv:1310.8115 [Cs, Math], 2013.
[84] S. Khazem, S. Chevallier, Q. Barthélemy, K. Haroun, and C. Noûs, "Minimizing Subject-dependent Calibration for BCI with Riemannian Transfer Learning." pp. 523-526.
[85] PyRiemann/PyRiemann. "pyRiemann," 2021; https://github.com/pyRiemann/pyRiemann.
[86] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, “A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update,” *J Neural Eng,* vol. 15, no. 3, pp. 031005, Jun, 2018.
[87] NeuroTechX. "NeuroTechX/Moabb," 2021; https://github.com/NeuroTechX/moabb.
[88] J. Reback, jbrockmendel, W. McKinney, J. Van den Bossche, T. Augspurger, P. Cloud, S. Hawkins, gfyoung, Sinhrks, M. Roeschke, A. Klein,

T. Petersen, J. Tratner, C. She, W. Ayd, H. Patrick, S. Naveh, M. Garcia, J. Schendel, A. Hayden, D. Saxton, M. E. Gorelli, R. Shadrach, V. Jancauskas, A. McMaster, F. Li, P. Battiston, S. Seabold, attack68, and K. Dong. "Pandas-Dev/Pandas: Pandas 1.3.0. v1.3.0, Zenodo," 2021.

[89] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research,* vol. 12, no. 85, pp. 2825-2830, 2021.

[90] S. C. f. C. Neuroscience. "Sccn/Labstreaminglayer," 2021; https://github.com/sccn/labstreaminglayer.

[91] T. H. Group. "The HDF5® Library & File Format," 2021; https://www.hdfgroup.org/solutions/hdf5/.

[92] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hamalainen, “MEG and EEG data analysis with MNE-Python,” *Front Neurosci,* vol. 7, pp. 267, Dec 26, 2013.

[93] Timeflux. "Timeflux/Workshops."

[94] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM,* vol. 60, no. 6, pp. 84-90, 2017, 2017.

[95] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, “Human-level control through deep reinforcement learning,” *Nature,* vol. 518, no. 7540, pp. 529-33, Feb 26, 2015.

[96] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, “Mastering the game of Go with deep neural networks and tree search,” *Nature,* vol. 529, no. 7587, pp. 484-9, Jan 28, 2016.

[97] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis, “Improved protein structure prediction using potentials from deep learning,” *Nature,* vol. 577, no. 7792, pp. 706-710, Jan, 2020.

[98] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, “Deep learning with convolutional neural networks for EEG decoding and visualization,” *Hum Brain Mapp,* vol. 38, no. 11, pp. 5391-5420, 11, 2017.

[99] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, “EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces,” *J Neural Eng,* vol. 15, no. 5, pp. 056013, Oct, 2018.

[100] K. R. Muller, C. W. Anderson, and G. E. Birch, “Linear and nonlinear methods for brain-computer interfaces,” *IEEE Trans Neural Syst Rehabil Eng,* vol. 11, no. 2, pp. 165-9, Jun, 2003.
[101] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, “Classification of covariance matrices using a Riemannian-based kernel for BCI applications,” *Neurocomputing,* vol. 112, pp. 172-178, 2013/07/18/, 2013.
[102] C. Jeunet, C. Benaroch, F. Cabestaing, R. Chavarriaga, E. Colamarino, M.-C. Corsi, D. Coyle, F. De Vico Fallani, S. Enriquez-Geppert, P. Figueirédo, M. Grosse-Wentrup, S. Kleih, S. Kober, A. Kübler, F. Lotte, E. Maby, D. Mattia, J. Mattout, G. R. Müller-Putz, S. Perdikis, L. Pillette, A. Riccio, S. Rimbert, A. Roc, R. N. Roy, R. Scherer, P. Seguin, H. Si-Mohammed, T. Tanaka, M. Tangermann, L. Tonin, A. Vourvopoulos, A. Vuckovic, G. Wood, and S. Wriessnegger, “A User-Centred Approach to Unlock the Potential of Non-Invasive BCIs: An Unprecedented International Translational Effort,” in CHIST-ERA, virtual, 2020.
[103] J. Deng, W. Dong, R. Socher, L. Li, L. Kai, and F.-F. Li, "ImageNet: A large-scale hierarchical image database." pp. 248-255.
[104] T. O. Zander, and C. Kothe, “Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general,” *Journal of neural engineering,* vol. 8, no. 2, pp. 025005-2560/8/2/025005. Epub 2011 Mar 24, 2011.
[105] V. Jayaram, M. Alamgir, Y. Altun, B. Schölkopf, and M. Grosse-Wentrup, “Transfer learning in brain-computer interfaces,” *IEEE Computational Intelligence Magazine,* vol. 11, no. 1, pp. 20-31, 2016.
[106] P. Zanini, M. Congedo, C. Jutten, S. Said, and Y. Berthoumieu, “Transfer Learning: A Riemannian Geometry Framework With Applications to Brain–Computer Interfaces,” *IEEE Transactions on Biomedical Engineering,* vol. 65, no. 5, pp. 1107-1116, 2018.
[107] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, “Learning Invariant Representations From EEG via Adversarial Inference,” *IEEE Access,* vol. 8, pp. 27074-27085, 2020.
[108] A. L. Benabid, T. Costecalde, A. Eliseyev, G. Charvet, A. Verney, S. Karakas, M. Foerster, A. Lambert, B. Moriniere, N. Abroug, M. C. Schaeffer, A. Moly, F. Sauter-Starace, D. Ratel, C. Moro, N. Torres-Martinez, L. Langar, M. Oddoux, M. Polosan, S. Pezzani, V. Auboiroux, T. Aksenova, C. Mestais, and S. Chabardes, “An exoskeleton controlled by an epidural wireless brain-machine interface in a tetraplegic patient: a proof-of-concept demonstration,” *Lancet Neurol,* vol. 18, no. 12, pp. 1112-1122, Dec, 2019.
[109] A. Eliseyev, V. Auboiroux, T. Costecalde, L. Langar, G. Charvet, C. Mestais, T. Aksenova, and A. L. Benabid, “Recursive Exponentially Weighted N-way Partial Least Squares Regression with Recursive-Validation of Hyper-Parameters in Brain-Computer Interface Applications,” *Sci Rep,* vol. 7, no. 1, pp. 16281, Nov 24, 2017.
[110] M. C. Schaeffer, and T. Aksenova, “Switching Markov decoders for asynchronous trajectory reconstruction from ECoG signals in monkeys for BCI applications,” *J Physiol Paris,* vol. 110, no. 4 Pt A, pp. 348-360, Nov, 2016.

[111] A. L. Orsborn, H. G. Moorman, S. A. Overduin, M. M. Shanechi, D. F. Dimitrov, and J. M. Carmena, “Closed-loop decoder adaptation shapes neural plasticity for skillful neuroprosthetic control,” *Neuron,* vol. 82, no. 6, pp. 1380-93, Jun 18, 2014.
[112] M. M. Madduri, S. A. Burden, and A. L. Orsborn, “A Game-Theoretic Model for Co-Adaptive Brain-Machine Interfaces,” in 10th International IEEE/EMBS Conference on Neural Engineering (NER), virtual, 2021, pp. 327-330.
[113] S. Perdikis, R. Leeb, J. Williamson, A. Ramsay, M. Tavella, L. Desideri, E. J. Hoogerwerf, A. Al-Khodairy, R. Murray-Smith, and J. D. Millan, “Clinical evaluation of BrainTree, a motor imagery hybrid BCI speller,” *J Neural Eng,* vol. 11, no. 3, pp. 036003, Jun, 2014.
[114] S. Perdikis, R. Leeb, R. Chavarriaga, and J. D. R. Millan, “Context-Aware Learning for Generative Models,” *IEEE Trans Neural Netw Learn Syst,* vol. 32, no. 8, pp. 3471-3483, Aug, 2021.
[115] S. Perdikis, R. Leeb, and J. D. Millan, “Context-aware adaptive spelling in motor imagery BCI,” *J Neural Eng,* vol. 13, no. 3, pp. 036018, Jun, 2016.
[116] S. Perdikis, and J. d. R. Millán, “Brain-machine interfaces: A tale of two learners. ,” *IEEE Systems, Man, and Cybernetics Magazine,* vol. 6, no. 3, pp. 12–19, 2020.
[117] S. Perdikis, L. Tonin, S. Saeedi, C. Schneider, and J. D. R. Millan, “The Cybathlon BCI race: Successful longitudinal mutual learning with two tetraplegic users,” *PLoS Biol,* vol. 16, no. 5, pp. e2003787, May, 2018.
[118] A. Biasiucci, R. Leeb, I. Iturrate, S. Perdikis, A. Al-Khodairy, T. Corbet, A. Schnider, T. Schmidlin, H. Zhang, M. Bassolino, D. Viceic, P. Vuadens, A. G. Guggisberg, and J. D. R. Millán, “Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke,” *Nat Commun,* vol. 9, no. 1, pp. 2421, 06, 2018.
[119] S. Amiri, R. Fazel-Rezai, and V. Asadpour, “A Review of Hybrid Brain-Computer Interface Systems,,” *Advances in Human-Computer Interaction,* vol. 8, pp. 187024, 2013.
[120] K. C. Armel, and V. S. Ramachandran, “Projecting sensations to external objects: evidence from skin conductance response,” *Proc Biol Sci,* vol. 270, no. 1523, pp. 1499-506, Jul 22, 2003.
[121] H. D. Critchley, “Electrodermal responses: what happens in the brain,” *Neuroscientist,* vol. 8, no. 2, pp. 132-42, Apr, 2002.
[122] F. Marini, C. F. Tagliabue, A. V. Sposito, A. Hernandez-Arieta, P. Brugger, N. Estevez, and A. Maravita, “Crossmodal representation of a functional robotic hand arises after extensive training in healthy participants,” *Neuropsychologia,* vol. 53, pp. 178-86, Jan, 2014.
[123] S. Shokur, S. Gallo, R. C. Moioli, A. R. C. Donati, E. Morya, H. Bleuler, and M. A. L. Nicolelis, “Assimilation of virtual legs and perception of floor texture by complete paraplegic patients receiving artificial tactile feedback,” *Sci Rep,* vol. 6, pp. 32293, Sep 19, 2016.
[124] P. Voss, M. E. Thomas, J. M. Cisneros-Franco, and E. de Villers-Sidani, “Dynamic Brains and the Changing Rules of Neuroplasticity: Implications for Learning and Recovery,” *Front Psychol,* vol. 8, pp. 1657, 2017.

[125] F. de Vignemont, “Embodiment, ownership and disownership,” *Conscious Cogn,* vol. 20, no. 1, pp. 82-93, Mar, 2011.
[126] V. S. Adama, S.-J. Wu, N. Nicolaou, and M. Bogdan, “Extendable Hybrid approach to detect consciousness states in a CLIS patient using machine learning.,” in 10th EUROSIM Congress, Logroño, Spain, 2019.
[127] P. Roussel, G. L. Godais, F. Bocquelet, M. Palma, J. Hongjie, S. Zhang, A. L. Giraud, P. Megevand, K. Miller, J. Gehrig, C. Kell, P. Kahane, S. Chabardes, and B. Yvert, “Observation and assessment of acoustic contamination of electrophysiological brain signals during speech production and sound perception,” *J Neural Eng,* vol. 17, no. 5, pp. 056028, Oct 15, 2020.
[128] A. Kubler, E. M. Holz, A. Riccio, C. Zickler, T. Kaufmann, S. C. Kleih, P. Staiger-Salzer, L. Desideri, E. J. Hoogerwerf, and D. Mattia, “The user-centered design as novel perspective for evaluating the usability of BCI-controlled applications,” *PloS one,* vol. 9, no. 12, pp. e112392, 2014.
[129] R. Andrich, N.-E. Mathiassen, E.-J. Hoogerwerf, and G. J. Gelderblom, “Service delivery systems for assistive technology in Europe: An AAATE/EASTIN position paper,” *Technology and Disability,* vol. 25, no. 3, pp. 127-146, 2013/01/01/, 2013.
[130] G. Liberati, A. Pizzimenti, L. Simione, A. Riccio, F. Schettini, M. Inghilleri, D. Mattia, and F. Cincotti, “Developing brain-computer interfaces from a user-centered perspective: Assessing the needs of persons with amyotrophic lateral sclerosis, caregivers, and professionals,” *Appl Ergon,* vol. 50, pp. 139-46, Sep, 2015.
[131] A. Riccio, F. Pichiorri, F. Schettini, J. Toppi, M. Risetti, R. Formisano, M. Molinari, L. Astolfi, F. Cincotti, and D. Mattia, “Interfacing brain with computer to improve communication and rehabilitation after brain damage,” *Progress in brain research,* vol. 228, pp. 357-387, 2016.
[132] A. Riccio, L. Simione, F. Schettini, A. Pizzimenti, M. Inghilleri, M. O. Belardinelli, D. Mattia, and F. Cincotti, “Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis,” *Frontiers in human neuroscience,* vol. 7, pp. 732, 2013.
[133] A. Riccio, F. Schettini, L. Simione, A. Pizzimenti, M. Inghilleri, M. Olivetti-Belardinelli, D. Mattia, and F. Cincotti, “On the Relationship Between Attention Processing and P300-Based Brain Computer Interface Control in Amyotrophic Lateral Sclerosis,” *Front Hum Neurosci,* vol. 12, pp. 165, 2018.
[134] M. Schreuder, A. Riccio, M. Risetti, S. Dähne, A. Ramsay, J. Williamson, D. Mattia, and M. Tangermann, “User-centered design in brain-computer interfaces-a case study,” *Artificial Intelligence in Medicine,* vol. 59, no. 2, pp. 71-80, 2013/10//, 2013.
[135] F. Schettini, A. Riccio, L. Simione, G. Liberati, M. Caruso, V. Frasca, B. Calabrese, M. Mecella, A. Pizzimenti, M. Inghilleri, D. Mattia, and F. Cincotti, “Assistive device with conventional, alternative, and brain-computer interface inputs to enhance interaction with the environment for people with amyotrophic lateral sclerosis: a feasibility and usability study,” *Archives of Physical Medicine and Rehabilitation,* vol. 96, no. 3 Suppl, pp. S46-53, 2015/03//, 2015.

[136] A. Riccio, E. M. Holz, P. Aricò, F. Leotta, F. Aloise, L. Desideri, M. Rimondini, A. Kübler, D. Mattia, and F. Cincotti, “Hybrid P300-based brain-computer interface to improve usability for people with severe motor disability: electromyographic signals for error correction during a spelling task,” *Archives of Physical Medicine and Rehabilitation,* vol. 96, no. 3 Suppl, pp. S54-61, 2015/03//, 2015.
[137] A. Riccio, F. Leotta, L. Bianchi, F. Aloise, C. Zickler, E. J. Hoogerwerf, A. Kübler, D. Mattia, and F. Cincotti, “Workload measurement in a communication application operated through a P300-based brain-computer interface,” *Journal of Neural Engineering,* vol. 8, no. 2, pp. 025028, 2011/04//, 2011.
[138] A. Kübler, F. Nijboer, and S. Kleih, “Hearing the needs of clinical users,” *Handbook of Clinical Neurology,* vol. 168, pp. 353-368, 2020, 2020.
[139] M. Eidel, and A. Kübler, “Wheelchair Control in a Virtual Environment by Healthy Participants Using a P300-BCI Based on Tactile Stimulation: Training Effects and Usability,” *Frontiers in Human Neuroscience,* vol. 14, pp. 265, 2020, 2020.
[140] P. Ziebell, J. Stümpfig, M. Eidel, S. C. Kleih, A. Kübler, M. E. Latoschik, and S. Halder, “Stimulus modality influences session-to-session transfer of training effects in auditory and tactile streaming-based P300 brain-computer interfaces,” *Scientific Reports,* vol. 10, no. 1, pp. 11873, 2020/07/17/, 2020.
[141] R. Chavarriaga, M. Fried-Oken, S. Kleih, F. Lotte, and R. Scherer, “Heading for new shores! Overcoming pitfalls in BCI design,” *Brain Computer Interfaces (Abingdon, England),* vol. 4, no. 1-2, pp. 60-73, 2017, 2017.
[142] J. D. Cunha, S. Perdikis, S. Halder, and R. Scherer, “Post-Adaptation Effects in a Motor Imagery Brain-Computer Interface Online Coadaptive Paradigm,” *IEEE Access,* vol. 9, pp. 41688-41703, 2021, 2021.
[143] A. Roc, L. Pillette, J. Mladenovic, C. Benaroch, B. N'Kaoua, C. Jeunet, and F. Lotte, “A review of user training methods in brain computer interfaces based on mental tasks,” *Journal of Neural Engineering*, 2020/11/12/, 2020.
[144] S. Halder, T. Leinfelder, S. M. Schulz, and A. Kübler, “Neural mechanisms of training an auditory event-related potential task in a brain-computer interface context,” *Human Brain Mapping,* vol. 40, no. 8, pp. 2399-2412, 2019/06/01/, 2019.
[145] B. Blankertz, C. Sannelli, S. Halder, E. M. Hammer, A. Kübler, K.-R. Müller, G. Curio, and T. Dickhaus, “Neurophysiological predictor of SMR-based BCI performance,” *NeuroImage,* vol. 51, no. 4, pp. 1303-1309, 2010/07/15/, 2010.
[146] A. Hammer, S. Vielhaber, A. Rodriguez-Fornells, B. Mohammadi, and T. F. Munte, “A neurophysiological analysis of working memory in amyotrophic lateral sclerosis,” *Brain research,* vol. 1421, pp. 90-99, 2011.
[147] R. Formisano, M. D'Ippolito, M. Risetti, A. Riccio, C. F. Caravasso, S. Catani, F. Rizza, A. Forcina, and M. G. Buzzi, “Vegetative state, minimally conscious state, akinetic mutism and Parkinsonism as a continuum of recovery from disorders of consciousness: an exploratory and preliminary study,” *Functional Neurology,* vol. 26, no. 1, pp. 15-24, 2011/03//Jan-undefined, 2011.

[148] J. T. Giacino, J. J. Fins, S. Laureys, and N. D. Schiff, “Disorders of consciousness after acquired brain injury: the state of the science,” *Nature Reviews. Neurology,* vol. 10, no. 2, pp. 99-114, 2014/02//, 2014.
[149] A. S. Nilsen, B. Juel, B. Thürer, and J. F. Storm, *Proposed EEG measures of consciousness: a systematic, comparative review*, PsyArXiv, 2020.
[150] S. Halder, B. E. Juel, A. S. Nilsen, L. V. Raghavan, and J. F. Storm, “Changes in measures of consciousness during anaesthesia of one hemisphere (Wada test),” *NeuroImage,* vol. 226, pp. 117566, 2021/02/01/, 2021.
[151] I. Kathner, S. C. Wriessnegger, G. R. Müller-Putz, A. Kubler, and S. Halder, “Effects of mental workload and fatigue on the P300, alpha and theta band power during operation of an ERP (P300) brain-computer interface,” *Biological psychology,* vol. 102, pp. 118-129, 2014.
[152] S. C. Kleih, F. Nijboer, S. Halder, and A. Kübler, “Motivation modulates the P300 amplitude during brain-computer interface use,” *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology,* vol. 121, no. 7, pp. 1023-1031, 2010/07//, 2010.
[153] Y. Han, E. Valentini, and S. Halder, “Prediction of tonic pain using support vector machines with phase-based connectivity features,” 2021, 2021.
[154] S. Orlandi, S. C. House, P. Karlsson, R. Saab, and T. Chau, “Brain-Computer Interfaces for Children With Complex Communication Needs and Limited Mobility: A Systematic Review,” *Front Hum Neurosci,* vol. 15, pp. 643294, 2021.
[155] E. Kinney-Lang, B. Auyeung, and J. Escudero, “Expanding the (kaleido)scope: exploring current literature trends for translating electroencephalography (EEG) based brain-computer interfaces for motor rehabilitation in children,” *J Neural Eng,* vol. 13, no. 6, pp. 061002, Dec, 2016.
[156] M. Fouillen, E. Maby, L. Le Carrer, V. Herbillon, and J. Mattout, “Erp-based BCI Training for children with ADHD: Motivations and Trial Design,” *GBCIC*, 2017.
[157] E. Mikołajewska, and D. Mikołajewski, “The prospects of brain — computer interface applications in children,” *Open Medicine,* vol. 9, no. 1, pp. 74-79, 2014.
[158] C. G. Lim, T. S. Lee, C. Guan, D. S. Fung, Y. Zhao, S. S. Teng, H. Zhang, and K. R. Krishnan, “A brain-computer interface based attention training program for treating attention deficit hyperactivity disorder,” *PloS one,* vol. 7, no. 10, pp. e46692, 2012.
[159] J. J. S. Norton, J. Mullins, B. E. Alitz, and T. Bretl, “The performance of 9-11-year-old children using an SSVEP-based BCI for target selection,” *J Neural Eng,* vol. 15, no. 5, pp. 056012, Oct, 2018.
[160] S. C. House, "Evaluation of a Motor Imagery Electroencephalography Brain-computer Interface as an Access Technology for Children (Doctoral dissertation)," University of Toronto, 2018.
[161] Z. Jadavji, J. Zhang, B. Paffrath, E. Zewdie, and A. Kirton, “Can Children With Perinatal Stroke Use a Simple Brain Computer Interface?,” *Stroke,* vol. 52, no. 7, pp. 2363-2370, Jul, 2021.

[162] E. Kinney-Lang, D. Kelly, E. D. Floreani, Z. Jadavji, D. Rowley, E. T. Zewdie, J. R. Anaraki, H. Bahari, K. Beckers, K. Castelane, L. Crawford, S. House, C. A. Rauh, A. Michaud, M. Mussi, J. Silver, C. Tuck, K. Adams, J. Andersen, T. Chau, and A. Kirton, “Advancing Brain-Computer Interface Applications for Severely Disabled Children Through a Multidisciplinary National Network: Summary of the Inaugural Pediatric BCI Canada Meeting,” *Frontiers in Human Neuroscience,* vol. 14, no. 530, 2020-December-03, 2020.
[163] D. Kelly, Z. Jadavji, E. Zewdie, E. Mitchell, K. Summerfield, A. Kirton, and E. Kinney-Lang, "A Child’s Right to Play: Results from the Brain-Computer Interface Game Jam 2019 (Calgary Competition)." pp. 6099-6102.
[164] E. Kinney-Lang, S. Murji, D. Kelly, B. Paffrath, E. Zewdie, and A. Kirton, “Designing a flexible tool for rapid implementation of brain-computer interfaces (BCI) in game development,” *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2020, pp. 6078-6081, Jul, 2020.
[165] J. L. Mullins, “SSVEP-based BCI performance in children,” 2015.
[166] K. Park, T. Kihl, S. Park, M.-J. Kim, and J. Chang, “Fairy tale directed game-based training system for children with ADHD using BCI and motion sensing technologies,” *Behaviour & Information Technology,* vol. 38, no. 6, pp. 564-577, 2019/06/03, 2019.
[167] E. V. Friedrich, N. Suttie, A. Sivanathan, T. Lim, S. Louchart, and J. A. Pineda, “Brain-computer interface game applications for combined neurofeedback and biofeedback treatment for children on the autism spectrum,” *Front Neuroeng,* vol. 7, pp. 21, 2014.
[168] E. Kinney-Lang, L. Spyrou, A. Ebied, R. F. M. Chin, and J. Escudero, “Tensor-driven extraction of developmental features from varying paediatric EEG datasets,” *J Neural Eng,* vol. 15, no. 4, pp. 046024, Aug, 2018.
[169] E. Kinney-Lang, A. Ebied, and J. Escudero, "Building a Tensor Framework for the Analysis and Classification of Steady-State Visual Evoked Potentials in Children." pp. 296-300.
[170] C. Forest, G. Beraldo, R. Mancin, E. Menegatti, and A. Suppiej, "Maturational aspects of visual P300 in children: a research window for pediatric Brain Computer Interface (BCI)[*]." pp. 451-455.
[171] E. Kinney-Lang, A. Ebied, B. Auyeung, R. F. M. Chin, and J. Escudero, “Introducing the Joint EEG-Development Inference (JEDI) Model: A Multi-Way, Data Fusion Approach for Estimating Paediatric Developmental Scores via EEG,” *IEEE Trans Neural Syst Rehabil Eng,* vol. 27, no. 3, pp. 348-357, Mar, 2019.
[172] E. Kinney-lang, and J. Escudero, "Programming for Pediatrics: A literature review of brain-computer interfaces for neurorehabilitation in children."
[173] A. Kirton, M. J. Metzler, B. T. Craig, A. Hilderley, M. Dunbar, A. Giuffre, J. Wrightson, E. Zewdie, and H. L. Carlson, “Perinatal stroke: mapping and modulating developmental plasticity,” *Nat Rev Neurol,* vol. 17, no. 7, pp. 415-432, Jul, 2021.
[174] S. Letourneau, E. T. Zewdie, Z. Jadavji, J. Andersen, L. M. Burkholder, and A. Kirton, “Clinician awareness of brain computer interfaces: a Canadian national survey,” *J Neuroeng Rehabil,* vol. 17, no. 1, pp. 2, Jan 6, 2020.

[175] S. o. C. P. i. Europe, “Surveillance of cerebral palsy in Europe: a collaboration of cerebral palsy surveys and registers. Surveillance of Cerebral Palsy in Europe (SCPE),” *Dev Med Child Neurol,* vol. 42, no. 12, pp. 816-24, Dec, 2000.
[176] J. Parkes, N. Hill, M. J. Platt, and C. Donnelly, “Oromotor dysfunction and communication impairments in children with cerebral palsy: a register study,” *Developmental medicine and child neurology,* vol. 52, no. 12, pp. 1113-1119, 2010.
[177] C. Cans, P. Guillem, C. Arnaud, F. Baille, J. Chalmers, V. McManus, G. Cussen, J. Parkes, H. Dolk, B. Hagberg, G. Hagberg, S. Jarvis, A. Colver, A. Johnson, G. Surman, I. Krageloh-Mann, R. Michaelis, M. J. Platt, P. Pharoah, M. Topp, P. Udall, M. G. Torrioli, M. Miceli, and M. Wichers, “Prevalence and characteristics of children with cerebral palsy in Europe,” *Developmental medicine and child neurology,* vol. 44, no. 9, pp. 633-640, 2002.
[178] J. Kennes, P. Rosenbaum, S. E. Hanna, S. Walter, D. Russell, P. Raina, D. Bartlett, and B. Galuppi, “Health status of school-aged children with cerebral palsy: information from a population-based sample,” *Developmental medicine and child neurology,* vol. 44, no. 4, pp. 240-247, 2002.
[179] E. Wood, and P. Rosenbaum, “The gross motor function classification system for cerebral palsy: a study of reliability and stability over time,” *Dev Med Child Neurol,* vol. 42, no. 5, pp. 292-6, May, 2000.
[180] R. E. Alcaide-Aguirre, S. A. Warschausky, D. Brown, A. Aref, and J. E. Huggins, “Asynchronous brain-computer interface for cognitive assessment in people with cerebral palsy,” *J Neural Eng,* vol. 14, no. 6, pp. 066001, 12, 2017.
[181] I. Daly, M. Billinger, J. Laparra-Hernandez, F. Aloise, M. L. Garcia, J. Faller, R. Scherer, and G. Muller-Putz, “On the control of brain-computer interfaces by users with cerebral palsy,” *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology,* vol. 124, no. 9, pp. 1787-1797, 2013.
[182] I. Daly, R. Scherer, M. Billinger, and G. Müller-Putz, “FORCe: Fully Online and Automated Artifact Removal for Brain-Computer Interfacing,” *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society,* vol. 23, no. 5, pp. 725-736, 2015.
[183] R. Scherer, M. Billinger, J. Wagner, A. Schwarz, D. T. Hettich, E. Bolinger, M. Lloria Garcia, J. Navarro, and G. Muller-Putz, “Thought-based row-column scanning communication board for individuals with cerebral palsy,” *Annals of physical and rehabilitation medicine,* vol. 58, no. 1, pp. 14-22, 2015.
[184] J. Huggins, M. Garcia, S. Tou, P. Karlsson, and S. Warschausky, “Comparison of brain-computer interface and eye-gaze interface technology for access to an untimed vocabulary test by people with cerebral palsy,” in Australasian Academy of Cerebral Palsy and Developmental Medicine, Perth, Australia, 2020.

[185] S. Warschausky, P. Karlsson, and J. Huggins, "Brain-Computer or Eye-Gaze Interfaces: Technical Challenges and Promise."
[186] M. Kawakami, M. Liu, T. Otsuka, A. Wada, K. Uchikawa, A. Aoki, and Y. Otaka, “Asymmetric skull deformity in children with cerebral palsy: frequency and correlation with postural abnormalities and deformities,” *J Rehabil Med,* vol. 45, no. 2, pp. 149-53, Feb, 2013.
[187] I. Minciu, “Clinical correlations in cerebral palsy,” *Maedica,* vol. 7, no. 4, pp. 319-324, 2012.
[188] S. Venkateswaran, and M. I. Shevell, “Comorbidities and clinical determinants of outcome in children with spastic quadriplegic cerebral palsy,” *Dev Med Child Neurol,* vol. 50, no. 3, pp. 216-22, Mar, 2008.
[189] M. Marneweck, H. C. Kuo, A. R. P. Smorenburg, C. L. Ferre, V. H. Flamand, D. Gupta, J. B. Carmel, Y. Bleyenheuft, A. M. Gordon, and K. M. Friel, “The Relationship Between Hand Function and Overlapping Motor Representations of the Hands in the Contralesional Hemisphere in Unilateral Spastic Cerebral Palsy,” *Neurorehabil Neural Repair,* vol. 32, no. 1, pp. 62-72, 01, 2018.
[190] S. L. J. Tou, S. A. Warschausky, P. Karlsson, and J. E. Huggins, “Individualized Electrode Subset Improves the Calibration Accuracy of an EEG P300-based Brain-Computer Interface for People with Severe Cerebral Palsy ” *JNE*, in revision.
[191] K. Himmelmann, G. Hagberg, L. M. Wiklund, M. N. Eek, and P. Uvebrant, “Dyskinetic cerebral palsy: a population-based study of children born between 1991 and 1998,” *Developmental medicine and child neurology,* vol. 49, no. 4, pp. 246-251, 2007.
[192] S. S. Philip, and G. N. Dutton, “Identifying and characterising cerebral visual impairment in children: a review,” *Clin Exp Optom,* vol. 97, no. 3, pp. 196-208, May, 2014.
[193] R. Scherer, A. Schwarz, G. R. Müller-Putz, V. Pammer-Schindler, and M. L. Garcia, "Lets play Tic-Tac-Toe: A Brain-Computer Interface case study in cerebral palsy." pp. 003736-003741.
[194] G. Wilson, S. Stavisky, F. Willett, D. Avansino, J. Kelemen, L. Hochberg, J. Henderson, S. Druckmann, and K. Shenoy, “Decoding spoken English from intracortical electrode arrays in dorsal precentral gyrus,” *Journal of Neural Engineering,* vol. 17, no. 6, pp. 66007-66007, 2020.
[195] S. D. Stavisky, F. R. Willett, G. H. Wilson, B. A. Murphy, P. Rezaii, D. T. Avansino, W. D. Memberg, J. P. Miller, R. F. Kirsch, L. R. Hochberg, A. B. Ajiboye, S. Druckmann, K. V. Shenoy, and J. M. Henderson, “Neural ensemble dynamics in dorsal motor cortex during speech in people with paralysis,” *eLife,* vol. 8, no. e46015, 2019.
[196] F. R. Willett, D. T. Avansino, L. R. Hochberg, J. M. Henderson, and K. V. Shenoy, “High-performance brain-to-text communication via handwriting,” *Nature,* vol. 593, no. 7858, pp. 249-254, 2021.
[197] C. Heelan, J. Lee, R. O’Shea, L. Lynch, D. M. Brandman, W. Truccolo, and A. V. Nurmikko, “Decoding speech from spike-based neural population recordings in secondary auditory cortex of non-human primates,” *Communications Biology,* vol. 2, no. 1, pp. 466-466, 2019.

[198] C. Herff, D. J. Krusienski, and P. Kubben, “The Potential of Stereotactic-EEG for Brain-Computer Interfaces: Current Progress and Future Directions,” *Frontiers in Neuroscience,* vol. 14, no. February, pp. 1-8, 2020.
[199] M. Gearing, and P. Kennedy, “Histological Confirmation of Myelinated Neural Filaments Within the Tip of the Neurotrophic Electrode After a Decade of Neural Recordings,” *Frontiers in Human Neuroscience,* vol. 14, 2020.
[200] J. Bartels, D. Andreasen, P. Ehirim, H. Mao, S. Seibert, E. J. Wright, and P. Kennedy, “Neurotrophic electrode: method of assembly and implantation into human motor speech cortex,” *J Neurosci Methods,* vol. 174, no. 2, pp. 168-76, Sep, 2008.
[201] C. Herff, L. Diener, M. Angrick, E. Mugler, M. C. Tate, M. A. Goldrick, D. J. Krusienski, M. W. Slutzky, and T. Schultz, “Generating Natural, Intelligible Speech From Brain Activity in Motor, Premotor, and Inferior Frontal Cortices,” *Frontiers in Neuroscience,* vol. 13, no. November, pp. 1-11, 2019.
[202] J. G. Makin, D. A. Moses, and E. F. Chang, “Machine translation of cortical activity to text with an encoder–decoder framework,” *Nature Neuroscience,* vol. 23, no. 4, pp. 575-582, 2020.
[203] C. Herff, D. Heger, A. de Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, “Brain-to-text: decoding spoken phrases from phone representations in the brain,” *Frontiers in neuroscience,* vol. 9, pp. 217, 2015.
[204] D. A. Moses, M. K. Leonard, J. G. Makin, and E. F. Chang, “Real-time decoding of question-and-answer speech dialogue using human cortical activity,” *Nature Communications,* vol. 10, no. 1, pp. 3096-3096, 2019.
[205] D. A. Moses, S. L. Metzger, J. R. Liu, G. K. Anumanchipalli, J. G. Makin, P. F. Sun, J. Chartier, M. E. Dougherty, P. M. Liu, G. M. Abrams, A. Tu-Chan, K. Ganguly, and E. F. Chang, “Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria,” *N Engl J Med,* vol. 385, no. 3, pp. 217-227, 07 15, 2021.
[206] M. Angrick, M. Ottenhoff, L. Diener, D. Ivucic, G. Ivucic, S. Goulis, J. Saal, A. J. Colon, L. Wagner, D. J. Krusienski, P. L. Kubben, T. Schultz, and C. Herff, “Real-time Synthesis of Imagined Speech Processes from Minimally Invasive Recordings of Neural Activity,” no. 2, 2020.
[207] J. S. Brumberg, K. M. Pitt, and J. D. Burnison, “A Noninvasive Brain-Computer Interface for Real-Time Speech Synthesis: The Importance of Multimodal Feedback,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 26, no. 4, pp. 874-881, 2018.
[208] J. Berezutskaya, M. J. Vansteensel, E. J. Aarnoutse, Z. V. Freudenburg, G. Piantoni, M. P. Branco, and N. F. Ramsey, “Open multimodal iEEG-fMRI dataset from naturalistic stimulation with a short audiovisual film,” *bioRxiv*, pp. 2021.06.09.447733-2021.06.09.447733, 2021.
[209] C. Aubinet, H. Cassol, O. Bodart, L. R. D. Sanz, S. Wannez, C. Martial, A. Thibaut, G. Martens, M. Carriere, O. Gosseries, S. Laureys, and C. Chatelle, “Simplified evaluation of CONsciousness disorders (SECONDs) in individuals with severe brain injury: A validation study,” *Ann Phys Rehabil Med,* vol. 64, no. 5, pp. 101432, Jul 27, 2021.

[210] L. R. D. Sanz, A. Thibaut, B. L. Edlow, S. Laureys, and O. Gosseries, “Update on neuroimaging in disorders of consciousness,” *Curr Opin Neurol,* vol. 34, no. 4, pp. 488-496, Aug 1, 2021.
[211] A. Thibaut, R. Panda, J. Annen, L. R. D. Sanz, L. Naccache, C. Martial, C. Chatelle, C. Aubinet, E. A. C. Bonin, A. Barra, M. M. Briand, B. Cecconi, S. Wannez, J. Stender, S. Laureys, and O. Gosseries, “Preservation of Brain Activity in Unresponsive Patients Identifies MCS Star,” *Ann Neurol,* vol. 90, no. 1, pp. 89-100, Jul, 2021.
[212] J. Annen, S. Laureys, and O. Gosseries, “Brain-computer interfaces for consciousness assessment and communication in severely brain-injured patients,” *Handb Clin Neurol,* vol. 168, pp. 137-152, 2020.
[213] J. Annen, I. Mertel, R. Xu, C. Chatelle, D. Lesenfants, R. Ortner, E. A. C. Bonin, C. Guger, S. Laureys, and F. Muller, “Auditory and Somatosensory P3 Are Complementary for the Assessment of Patients with Disorders of Consciousness,” *Brain Sci,* vol. 10, no. 10, Oct 17, 2020.
[214] D. Coyle, J. Stow, K. McCreadie, J. McElligott, and A. Carroll, “Sensorimotor modulation assessment and brain-computer interface training in disorders of consciousness,” *Archives of Physical Medicine and Rehabilitation,* vol. 96, no. 3 Suppl, pp. S62-70, 2015.
[215] N. Dayan, A. D. Bigirimana, A. E. Mccann, J. Stow, J. Mcelligott, A. Carroll, and D. H. Coyle, "Towards Answering Questions in Disorders of Consciousness and Locked-In Syndrome with a SMR-BCI."
[216] C. Guger, R. Spataro, B. Z. Allison, A. Heilinger, R. Ortner, W. Cho, and V. La Bella, “Complete Locked-in and Locked-in Patients: Command Following Assessment and Communication with Vibro-Tactile P300 and Motor Imagery Brain-Computer Interface Tools,” *Front Neurosci,* vol. 11, pp. 251, 2017.
[217] C. Guger, R. Spataro, F. Pellas, B. Z. Allison, A. Heilinger, R. Ortner, W. Cho, R. Xu, V. La Bella, G. Edlinger, J. Annen, G. Mandalá, C. Chatelle, and S. Laureys, “Assessing Command-Following and Communication With Vibro-Tactile P300 Brain-Computer Interface Tools in Patients With Unresponsive Wakefulness Syndrome,” *Front Neurosci,* vol. 12, pp. 423, 2018.
[218] N. Murovec, A. Heilinger, R. Xu, R. Ortner, R. Spataro, V. La Bella, Y. Miao, J. Jin, C. Chatelle, S. Laureys, B. Z. Allison, and C. Guger, “Effects of a Vibro-Tactile P300 Based Brain-Computer Interface on the Coma Recovery Scale-Revised in Patients With Disorders of Consciousness,” *Front Neurosci,* vol. 14, pp. 294, 2020.
[219] R. Spataro, A. Heilinger, B. Allison, D. De Cicco, S. Marchese, C. Gregoretti, V. La Bella, and C. Guger, “Preserved somatosensory discrimination predicts consciousness recovery in unresponsive wakefulness syndrome,” *Clin Neurophysiol,* vol. 129, no. 6, pp. 1130-1136, Jun, 2018.
[220] J. Annen, S. Blandiaux, N. Lejeune, M. A. Bahri, A. Thibaut, W. Cho, C. Guger, C. Chatelle, and S. Laureys, “BCI Performance and Brain Metabolism Profile in Severely Brain-Injured Patients Without Response to Command at Bedside,” *Front Neurosci,* vol. 12, pp. 370, 2018.

[221] A. Thibaut, N. Schiff, J. Giacino, S. Laureys, and O. Gosseries, “Therapeutic interventions in patients with prolonged disorders of consciousness,” *Lancet Neurol,* vol. 18, no. 6, pp. 600-614, Jun, 2019.
[222] A. Bowen, K. McKenna, and R. C. Tallis, “Reasons for Variability in the Reported Rate of Occurrence of Unilateral Spatial Neglect After Stroke,” *Stroke,* vol. 30, no. 6, pp. 1196-1202, 1999.
[223] E. Raffin, and F. C. Hummel, “Restoring Motor Functions After Stroke: Multiple Approaches and Opportunities,” *Neuroscientist,* vol. 24, no. 4, 2018.
[224] A. B. Remsik, K. Dodd, W. Leroy, J. Thoma, T. Jacobson, J. D. Allen, H. Advani, R. Mohanty, M. McMillan, S. Rajan, M. Walczak, B. M. Young, Z. Nigogosyan, C. A. Rivera, M. Mazrooyisebdani, N. Tellapragada, L. M. Walton, K. Gjini, P. L. E. Van Kan, T. J. Kang, J. A. Sattin, V. A. Nair, D. F. Edwards, J. C. Williams, and V. Prabhakaran, “Behavioral outcomes following brain⇓computer interface intervention for upper extremity rehabilitation in stroke: A randomized controlled trial,” *Frontiers in Neuroscience,* vol. 12, no. NOV, 2018.
[225] M. Grosse-Wentrup, D. Mattia, and K. Oweiss, “Using brain-computer interfaces to induce neural plasticity and restore function,” *Journal of neural engineering,* vol. 8, no. 2, pp. 025004, 2011.
[226] M. Sebastián-Romagosa, W. Cho, R. Ortner, N. Murovec, T. Von Oertzen, K. Kamada, B. Z. Allison, and C. Guger, “Brain Computer Interface Treatment for Motor Rehabilitation of Upper Extremity of Stroke Patients—A Feasibility Study,” *Frontiers in Neuroscience,* vol. 0, pp. 1056-1056, 2020.
[227] N. Murovec, M. Sebastian-Romagosa, S. Dangl, W. Cho, R. Ortner, and C. Guger, “Preliminary Results of a Brain-Computer Interface System based on Functional Electrical Stimulation and Avatar Feedback for Lower Extremity Rehabilitation of Chronic Stroke Patients,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* vol. 2020-October, pp. 7-11, 2020.
[228] R. Rupp, M. Rohm, M. Schneiders, A. Kreilinger, and G. Muller-Putz, “Functional rehabilitation of the paralyzed upper extremity after spinal cord injury by noninvasive hybrid neuroprosthesis,” *Proceedings of the IEEE,* vol. 103, no. 6, pp. 954-968, 2015.
[229] G. Müller-Putz, and R. Rupp, "Neuroprosthetics and Brain-Computer Interfaces," *Spinal Cord Injury: A Guide for Clinicians and End Users*: Springer Nature, 2021.
[230] J. L. Collinger, B. Wodlinger, J. E. Downey, W. Wang, E. C. Tyler-Kabara, D. J. Weber, A. J. McMorland, M. Velliste, M. L. Boninger, and A. B. Schwartz, “High-performance neuroprosthetic control by an individual with tetraplegia,” *Lancet,* vol. 381, no. 9866, pp. 557-64, Feb, 2013.
[231] L. R. Hochberg, M. D. Serruya, G. M. Friehs, J. A. Mukand, M. Saleh, A. H. Caplan, A. Branner, D. Chen, R. D. Penn, and J. P. Donoghue, “Neuronal ensemble control of prosthetic devices by a human with tetraplegia.see comment,” *Nature,* vol. 442, no. 7099, pp. 164-171, 2006.
[232] G. Pfurtscheller, G. R. Muller, J. Pfurtscheller, H. J. Gerner, and R. Rupp, “'Thought'--control of functional electrical stimulation to restore hand grasp in a patient with tetraplegia,” *Neuroscience letters,* vol. 351, no. 1, pp. 33-36, 2003.

[233] G. R. Müller-Putz, R. Scherer, G. Pfurtscheller, and R. Rupp, “EEG-based neuroprosthesis control: a step towards clinical practice,” *Neuroscience letters,* vol. 382, no. 1-2, pp. 169-174, 2005.
[234] M. Rohm, M. Schneiders, C. Muller, A. Kreilinger, V. Kaiser, G. R. Müller-Putz, and R. Rupp, “Hybrid brain-computer interfaces and hybrid neuroprostheses for restoration of upper limb functions in individuals with high-level spinal cord injury,” *Artificial Intelligence in Medicine,* vol. 59, no. 2, pp. 133-142, 2013.
[235] G. R. Muller-Putz, R. Rupp, P. Ofner, J. Pereira, A. Pinegger, A. Schwarz, M. Zube, U. Eck, B. Hessing, and M. Schneiders, “Applying intuitive EEG-controlled grasp neuroprostheses in individuals with spinal cord injury: Preliminary results from the MoreGrasp clinical feasibility study,” *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2019, pp. 5949-5955, Jul, 2019.
[236] P. Ofner, A. Schwarz, J. Pereira, D. Wyss, R. Wildburger, and G. R. Müller-Putz, “Attempted Arm and Hand Movements can be Decoded from Low-Frequency EEG from Persons with Spinal Cord Injury,” *Sci Rep,* vol. 9, no. 1, pp. 7134, 05 09, 2019.
[237] P. Ofner, A. Schwarz, J. Pereira, and G. R. Müller-Putz, “Upper limb movements can be decoded from the time-domain of low-frequency EEG,” *PLoS One,* vol. 12, no. 8, pp. e0182578, 2017.
[238] A. Schwarz, P. Ofner, J. Pereira, A. I. Sburlea, and G. R. Müller-Putz, “Decoding natural reach-and-grasp actions from human EEG,” *J Neural Eng,* vol. 15, no. 1, pp. 016005, 02, 2018.
[239] A. Schwarz, J. Pereira, R. Kobler, and G. R. Muller-Putz, “Unimanual and Bimanual Reach-and-Grasp Actions Can Be Decoded From Human EEG,” *IEEE Trans Biomed Eng,* vol. 67, no. 6, pp. 1684-1695, 06, 2020.
[240] A. I. Sburlea, and G. R. Müller-Putz, “Exploring representations of human grasping in neural, muscle and kinematic signals,” *Sci Rep,* vol. 8, no. 1, pp. 16669, Nov, 2018.
[241] A. I. Sburlea, M. Wilding, and G. R. Müller-Putz, “Disentangling human grasping type from the object's intrinsic properties using low-frequency EEG signals,” *Neuroimage: Reports,* vol. 1, no. 2, pp. 100012, 2021/06/01/, 2021.
[242] V. Martinez-Cagigal, R. J. Kobler, V. Mondini, R. Hornero, and G. R. Muller-Putz, “Non-linear online low-frequency EEG decoding of arm movements during a pursuit tracking task,” *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2020, pp. 2981-2985, 07, 2020.
[243] V. Mondini, R. J. Kobler, A. I. Sburlea, and G. R. Müller-Putz, “Continuous low-frequency EEG decoding of arm movement for closed-loop, natural control of a robotic arm,” *J Neural Eng,* vol. 17, no. 4, pp. 046031, 08 11, 2020.
[244] G. R. Müller-Putz, V. Mondini, V. Martínez-Cagigal, R. J. Kobler, J. Pereira, C. L. Dias, L. Hehenberger, and A. I. Sburlea, "Decoding of continuous movement attempt in 2-dimensions from non-invasive low frequency brain signals." pp. 322-325.
[245] R. J. Kobler, A. I. Sburlea, and G. R. Müller-Putz, “Tuning characteristics of low-frequency EEG to positions and velocities in visuomotor and oculomotor tracking tasks,” *Sci Rep,* vol. 8, no. 1, pp. 17713, Dec, 2018.

[246] R. J. Kobler, E. Kolesnichenko, A. I. Sburlea, and G. R. Müller-Putz, “Distinct cortical networks for hand movement initiation and directional processing: An EEG study,” *Neuroimage,* vol. 220, pp. 117076, 10 15, 2020.
[247] J. Hammer, J. Fischer, J. Ruescher, A. Schulze-Bonhage, A. Aertsen, and T. Ball, “The role of ECoG magnitude and phase in decoding position, velocity, and acceleration during continuous motor behavior,” *Front Neurosci,* vol. 7, pp. 200, 2013.
[248] J. Hammer, T. Pistohl, J. Fischer, P. Kršek, M. Tomášek, P. Marusič, A. Schulze-Bonhage, A. Aertsen, and T. Ball, “Predominance of Movement Speed Over Direction in Neuronal Population Signals of Motor Cortex: Intracranial EEG Data and A Simple Explanatory Model,” *Cereb Cortex,* vol. 26, no. 6, pp. 2863-81, Jun, 2016.
[249] R. J. Kobler, A. I. Sburlea, V. Mondini, M. Hirata, and G. R. Müller-Putz, “Distance- and speed-informed kinematics decoding improves M/EEG based upper-limb movement decoder accuracy,” *J Neural Eng,* vol. 17, no. 5, pp. 056027, 11 04, 2020.
[250] R. J. Kobler, A. I. Sburlea, C. Lopes-Dias, A. Schwarz, M. Hirata, and G. R. Müller-Putz, “Corneo-retinal-dipole and eyelid-related eye artifacts can be corrected offline and online in electroencephalographic and magnetoencephalographic signals,” *Neuroimage,* vol. 218, pp. 117000, 09, 2020.
[251] A. Korik, R. Sosnik, N. Siddique, and D. Coyle, “Decoding Imagined 3D Hand Movement Trajectories From EEG: Evidence to Support the Use of Mu, Beta, and Low Gamma Oscillations,” *Front Neurosci,* vol. 12, pp. 130, 2018.
[252] A. Korik, R. Sosnik, N. Siddique, and D. Coyle, “Decoding Imagined 3D Arm Movement Trajectories From EEG to Control Two Virtual Arms-A Pilot Study,” *Front Neurorobot,* vol. 13, pp. 94, 2019.
[253] O. Lennon, M. Tonellato, A. Del Felice, R. Di Marco, C. Fingleton, A. Korik, E. Guanziroli, F. Molteni, C. Guger, R. Otner, and D. Coyle, “A Systematic Review Establishing the Current State-of-the-Art, the Limitations, and the DESIRED Checklist in Studies of Direct Neural Interfacing With Robotic Gait Devices in Stroke Rehabilitation,” *Front Neurosci,* vol. 14, pp. 578, 2020.
[254] R. Mane, T. Chouhan, and C. Guan, “BCI for stroke rehabilitation: motor and beyond,” *J Neural Eng,* vol. 17, no. 4, pp. 041001, 08 17, 2020.
[255] R. Mane, K. K. Ang, and C. Guan, "Brain-Computer Interface for Stroke Rehabilitation," *Handbook of Neuroengineering*, T. N.V., ed., Singapore: Springer, 2021.
[256] O. Y. Kwon, M. H. Lee, C. Guan, and S. W. Lee, “Subject-Independent Brain-Computer Interfaces Based on Deep Convolutional Neural Networks,” *IEEE Trans Neural Netw Learn Syst,* vol. 31, no. 10, pp. 3839-3852, 10, 2020.
[257] K. Zhang, N. Robinson, S. W. Lee, and C. Guan, “Adaptive transfer learning for EEG motor imagery classification with deep Convolutional Neural Network,” *Neural Netw,* vol. 136, pp. 1-10, Apr, 2021.
[258] D. Kuhner, L. D. J. Fiederer, J. Aldinger, F. Burget, M. Völker, R. T. Schirrmeister, C. Do, J. Bödecker, B. Nebel, T. Ball, and W. Burgard, “Deep

Learning Based BCI Control of a Robotic Service Assistant Using Intelligent Goal Formulation," 2018-03-26, 2018.
[259] K. G. Hartmann, R. T. Schirrmeister, and T. Ball, "Hierarchical Internal Representation of Spectral Features in Deep Convolutional Networks Trained for EEG Decoding."
[260] B. Wodlinger, J. E. Downey, E. C. Tyler-Kabara, A. B. Schwartz, M. L. Boninger, and J. L. Collinger, "Ten-dimensional anthropomorphic arm control in a human brain-machine interface: difficulties, solutions, and limitations," *Journal of neural engineering,* vol. 12, no. 1, pp. 016011-2560/12/1/016011. Epub 2014 Dec 16, 2015.
[261] S. N. Flesher, J. L. Collinger, S. T. Foldes, J. M. Weiss, J. E. Downey, E. C. Tyler-Kabara, S. J. Bensmaia, A. B. Schwartz, M. L. Boninger, and R. A. Gaunt, "Intracortical microstimulation of human somatosensory cortex," *Science translational medicine,* vol. 8, no. 361, pp. 361ra141 361ra141, 2016.
[262] M. A. Salas, L. Bashford, S. Kellis, M. Jafari, H. Jo, D. Kramer, K. Shanfield, K. Pejsa, B. Lee, C. Y. Liu, and R. A. Andersen, "Proprioceptive and cutaneous sensations in humans elicited by intracortical microstimulation," *eLife,* vol. 7, pp. e32904, 2018.
[263] S. V. Hiremath, E. C. Tyler-Kabara, J. J. Wheeler, D. W. Moran, R. A. Gaunt, J. L. Collinger, S. T. Foldes, D. J. Weber, W. Chen, M. L. Boninger, and W. Wang, "Human perception of electrical stimulation on the surface of somatosensory cortex," *PloS one,* vol. 12, no. 5, pp. e0176020, 2017.
[264] B. Lee, D. Kramer, M. A. Salas, S. Kellis, D. Brown, T. Dobreva, C. Klaes, C. Heck, C. Liu, and R. A. Andersen, "Engineering Artificial Somatosensation Through Cortical Stimulation in Humans," *Frontiers in Systems Neuroscience,* vol. 12, pp. 24, 2018.
[265] L. A. Johnson, J. D. Wander, D. Sarma, D. K. Su, E. E. Fetz, and J. G. Ojemann, "Direct electrical stimulation of the somatosensory cortex in humans using electrocorticography electrodes: a qualitative and quantitative report," *Journal of neural engineering,* vol. 10, no. 3, pp. 036021, 2013.
[266] S. Raspopovic, M. Capogrosso, F. M. Petrini, M. Bonizzato, J. Rigosa, G. D. Pino, J. Carpaneto, M. Controzzi, T. Boretius, E. Fernandez, G. Granata, C. M. Oddo, L. Citi, A. L. Ciancio, C. Cipriani, M. C. Carrozza, W. Jensen, E. Guglielmelli, T. Stieglitz, P. M. Rossini, and S. Micera, "Restoring natural sensory feedback in real-time bidirectional hand prostheses," *Science translational medicine,* vol. 6, no. 222, pp. 222ra19 222ra19, 2014.
[267] T. S. Davis, H. A. Wark, D. T. Hutchinson, D. J. Warren, K. O'Neill, T. Scheinblum, G. A. Clark, R. A. Normann, and B. Greger, "Restoring motor control and sensory feedback in people with upper extremity amputations using arrays of 96 microelectrodes implanted in the median and ulnar nerves," *J Neural Eng,* vol. 13, no. 3, pp. 036001, 06, 2016.
[268] D. W. Tan, M. A. Schiefer, M. W. Keith, J. R. Anderson, J. Tyler, and D. J. Tyler, "A neural interface provides long-term stable natural touch perception," *Science Translational Medicine,* vol. 6, no. 257, pp. 257ra138-257ra138, 2014.

[269] S. Chandrasekaran, A. C. Nanivadekar, G. McKernan, E. R. Helm, M. L. Boninger, J. L. Collinger, R. A. Gaunt, and L. E. Fisher, “Sensory restoration by epidural stimulation of the lateral spinal cord in upper-limb amputees,” *eLife,* vol. 9, 2020.
[270] R. S. Johansson, and J. R. Flanagan, “Coding and use of tactile signals from the fingertips in object manipulation tasks,” *Nat Rev Neurosci,* vol. 10, no. 5, pp. 345 359, 2009.
[271] H. P. Saal, B. P. Delhaye, B. C. Rayhaun, and S. J. Bensmaia, “Simulating tactile signals from the whole hand with millisecond precision,” *Proceedings of the National Academy of Sciences of the United States of America,* vol. 114, no. 28, pp. E5693 E5702, 2017.
[272] E. V. Okorokova, Q. He, and S. J. Bensmaia, “Biomimetic encoding model for restoring touch in bionic hands through a nerve interface,” *Journal of Neural Engineering,* vol. 15, no. 6, pp. 066033, 2018.
[273] T. Callier, A. K. Suresh, and S. J. Bensmaia, “Neural Coding of Contact Events in Somatosensory Cortex,” *Cerebral cortex (New York, NY : 1991)*, 2019.
[274] K. Kumaravelu, T. Tomlinson, T. Callier, J. Sombeck, S. J. Bensmaia, L. E. Miller, and W. M. Grill, “A comprehensive model-based framework for optimal design of biomimetic patterns of electrical stimulation for prosthetic sensation,” *Journal of Neural Engineering,* vol. 17, no. 4, pp. 046045, 2020.
[275] C. M. Oddo, S. Raspopovic, F. Artoni, A. Mazzoni, G. Spigler, F. Petrini, F. Giambattistelli, F. Vecchio, F. Miraglia, L. Zollo, G. D. Pino, D. Camboni, M. C. Carrozza, E. Guglielmelli, P. M. Rossini, U. Faraguna, and S. Micera, “Intraneural stimulation elicits discrimination of textural features by artificial fingertip in intact and amputee humans,” *eLife,* vol. 5, pp. e09148, 2016.
[276] G. Valle, A. Mazzoni, F. Iberite, E. D'Anna, I. Strauss, G. Granata, M. Controzzi, F. Clemente, G. Rognini, C. Cipriani, T. Stieglitz, F. M. Petrini, P. M. Rossini, and S. Micera, “Biomimetic Intraneural Sensory Feedback Enhances Sensation Naturalness, Tactile Sensitivity, and Manual Dexterity in a Bidirectional Prosthesis,” *Neuron,* vol. 100, no. 1, pp. 37 45.e7, 2018.
[277] L. Bashford, I. Rosenthal, S. Kellis, K. Pejsa, D. Kramer, B. Lee, C. Liu, and R. A. Andersen, “The neurophysiological representation of imagined somatosensory percepts in human cortex,” *Journal of Neuroscience,* vol. 41, no. 10, pp. JN-RM-2460-20, 2021.
[278] E. W. Sellers, T. M. Vaughan, and J. R. Wolpaw, “A brain-computer interface for long-term independent home use,” *Amyotrophic lateral sclerosis : official publication of the World Federation of Neurology Research Group on Motor Neuron Diseases,* vol. 11, no. 5, pp. 449-455, 2010.
[279] R. Leeb, S. Perdikis, L. Tonin, A. Biasiucci, M. Tavella, M. Creatura, A. Molina, A. Al-Khodairy, T. Carlson, and J. D. R. Millán, “Transferring brain-computer interfaces beyond the laboratory: successful application control for motor-disabled users,” *Artificial Intelligence in Medicine,* vol. 59, no. 2, pp. 121-132, 2013/10//, 2013.
[280] E. M. Holz, L. Loic Botrel, and A. Kübler, “Independent home use of Brain Painting improves quality of life of two artists in the locked-in state

diagnosed with amyotrophic lateral sclerosis," *Brain-Computer Interfaces,* vol. 2, no. 2-3, pp. 117-134, 2015.
[281] J. R. Wolpaw, R. S. Bedlack, D. J. Reda, R. J. Ringer, P. G. Banks, T. M. Vaughan, S. M. Heckman, L. M. McCane, C. S. Carmack, S. Winden, D. J. McFarland, E. W. Sellers, H. Shi, T. Paine, D. S. Higgins, A. C. Lo, H. S. Patwa, K. J. Hill, G. D. Huang, and R. L. Ruff, "Independent home use of a brain-computer interface by people with amyotrophic lateral sclerosis," *Neurology,* vol. 91, no. 3, pp. e258-e267, 2018.
[282] E. G. M. Pels, E. J. Aarnoutse, S. Leinders, Z. V. Freudenburg, M. P. Branco, B. H. van der Vijgh, T. J. Snijders, T. Denison, M. J. Vansteensel, and N. F. Ramsey, "Stability of a chronic implanted brain-computer interface in late-stage amyotrophic lateral sclerosis," *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology,* vol. 130, no. 10, pp. 1798-1803, 2019/10//, 2019.
[283] A. Geronimo, and Z. Simmons, "TeleBCI: remote user training, monitoring, and communication with an evoked-potential brain-computer interface," *Brain Comput Interfaces,* vol. 7, no. 3-4, pp. 57-69, 2020, 2020.
[284] T. J. Oxley, P. E. Yoo, G. S. Rind, S. M. Ronayne, C. M. S. Lee, C. Bird, V. Hampshire, R. P. Sharma, A. Morokoff, D. L. Williams, C. MacIsaac, M. E. Howard, L. Irving, I. Vrljic, C. Williams, S. E. John, F. Weissenborn, M. Dazenko, A. H. Balabanski, D. Friedenberg, A. N. Burkitt, Y. T. Wong, K. J. Drummond, P. Desmond, D. Weber, T. Denison, L. R. Hochberg, S. Mathers, T. J. O'Brien, C. N. May, J. Mocco, D. B. Grayden, B. C. V. Campbell, P. Mitchell, and N. L. Opie, "Motor neuroprosthesis implanted with neurointerventional surgery improves capacity for activities of daily living tasks in severe paralysis: first in-human experience," *Journal of Neurointerventional Surgery,* vol. 13, no. 2, pp. 102-108, 2021/02//, 2021.
[285] J. D. Simeral, T. Hosman, J. Saab, S. N. Flesher, M. Vilela, B. Franco, J. N. Kelemen, D. M. Brandman, J. G. Ciancibello, P. G. Rezaii, E. N. Eskandar, D. M. Rosler, K. V. Shenoy, J. M. Henderson, A. V. Nurmikko, and L. R. Hochberg, "Home Use of a Percutaneous Wireless Intracortical Brain-Computer Interface by Individuals With Tetraplegia," *IEEE transactions on bio-medical engineering,* vol. 68, no. 7, pp. 2313-2325, 2021/07//, 2021.
[286] A. Kübler, E. M. Holz, A. Riccio, C. Zickler, T. Kaufmann, S. C. Kleih, P. Staiger-Sälzer, L. Desideri, E.-J. Hoogerwerf, and D. Mattia, "The user-centered design as novel perspective for evaluating the usability of BCI-controlled applications," *PloS One,* vol. 9, no. 12, pp. e112392, 2014, 2014.
[287] E. G. M. Pels, E. J. Aarnoutse, N. F. Ramsey, and M. J. Vansteensel, "Estimated Prevalence of the Target Population for Brain-Computer Interface Neurotechnology in the Netherlands," *Neurorehabil Neural Repair,* vol. 31, no. 7, pp. 677-685, Jul, 2017.
[288] K. Linse, E. Aust, M. Joos, and A. Hermann, "Communication Matters-Pitfalls and Promise of Hightech Communication Devices in Palliative Care of Severely Physically Disabled Patients With Amyotrophic Lateral Sclerosis," *Frontiers in Neurology,* vol. 9, pp. 603, 2018, 2018.
[289] K. M. Pitt, and J. S. Brumberg, "Guidelines for Feature Matching Assessment of Brain-Computer Interfaces for Augmentative and Alternative

Communication," *American Journal of Speech-Language Pathology,* vol. 27, no. 3, pp. 950-964, 2018/08/06/, 2018.
[290] K. M. Pitt, and J. S. Brumberg, "Evaluating person-centered factors associated with brain-computer interface access to a commercial augmentative and alternative communication paradigm," *Assistive technology: the official journal of RESNA*, pp. 1-10, 2021/03/05/, 2021.
[291] "Standards Roadmap: Neurotechnologies for Brain-Machine Interfacing IEEE Standards Association," 2020.
[292] Z. V. Freudenburg, M. P. Branco, S. Leinders, B. H. van der Vijgh, E. G. M. Pels, T. Denison, L. H. van den Berg, K. J. Miller, E. J. Aarnoutse, N. F. Ramsey, and M. J. Vansteensel, "Sensorimotor ECoG Signal Features for BCI Control: A Comparison Between People With Locked-In Syndrome and Able-Bodied Controls," *Frontiers in Neuroscience,* vol. 13, pp. 1058, 2019, 2019.
[293] A. Geronimo, H. E. Stephens, S. J. Schiff, and Z. Simmons, "Acceptance of brain-computer interfaces in amyotrophic lateral sclerosis," *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration,* vol. 16, no. 3-4, pp. 258-264, 2015/06//, 2015.
[294] K. Hill, T. Kovacs, and S. Shin, "Reliability of brain-computer interface language sample transcription procedures," *J Rehabil Res Dev,* vol. 51, no. 4, pp. 579-90, 2014.
[295] K. Hill, T. Kovacs, and S. Shin, "Critical issues using brain-computer interfaces for augmentative and alternative communication," *Archives of Physical Medicine and Rehabilitation,* vol. 96, no. 3 Suppl, pp. S8-15, 2015.
[296] K. Hill, and B. Romich, "A rate index for augmentative and alternative communication," *International Journal Speech Technology,* vol. 5, pp. 57-64, 2002.
[297] J. A. Lenker, H. H. Koester, and R. O. Smith, "Toward a national system of assistive technology outcomes measurement," *Assistive technology: the official journal of RESNA,* vol. 33, no. 1, pp. 1-8, 2021/01/02/, 2021.
[298] D. J. McFarland, W. A. Sarnacki, and J. R. Wolpaw, "Brain-computer interface (BCI) operation: optimizing information transfer rates," *Biological Psychology,* vol. 63, no. 3, pp. 237-251, 2003/07//, 2003.
[299] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex," *PLoS biology,* vol. 10, no. 1, pp. e1001251, 2012/01//, 2012.
[300] S. Sadeghi, and A. Maleki, "Accurate estimation of information transfer rate based on symbol occurrence probability in brain-computer interfaces," *Biomed Signal Process Control,* vol. 54, no. 44, pp. 101607, 2019, 2019.
[301] R. A. Robbins, Z. Simmons, B. A. Bremer, S. M. Walsh, and S. Fischer, "Quality of life in ALS is maintained as physical function declines," *Neurology,* vol. 56, no. 4, pp. 442-444, 2001/02/27/, 2001.
[302] R. Chavarriaga, C. Carey, J. L. Contreras-Vidal, Z. Mckinney, and L. Bianchi, "Standardization of Neurotechnology for Brain-Machine Interfacing: State of the Art and Recommendations," *IEEE Open Journal of Engineering in Medicine and Biology,* vol. 2, pp. 71-73, 2021.

[303] A. Y. Paek, J. A. Brantley, A. Sujatha Ravindran, K. Nathan, Y. He, D. Eguren, J. G. Cruz-Garza, S. Nakagome, D. S. Wickramasuriya, J. Chang, M. Rashed-Al-Mahfuz, M. R. Amin, N. A. Bhagat, and J. L. Contreras-Vidal, "A Roadmap Towards Standards for Neurally Controlled End Effectors," *IEEE Open Journal of Engineering in Medicine and Biology,* vol. 2, pp. 84-90, 2021.
[304] I. B. Initiative, "Future Neural Therapeutics: Technology Roadmap White Paper Version 2," IEEE, 2020.
[305] A. Y. Paek, J. A. Brantley, B. J. Evans, and J. L. Contreras-Vidal, "Concerns in the Blurred Divisions Between Medical and Consumer Neurotechnology," *IEEE Systems Journal,* vol. 15, no. 2, pp. 3069-3080, 2021.
[306] A. Wexler, and P. B. Reiner, "Oversight of direct-to-consumer neurotechnologies," *Science,* vol. 363, no. 6424, pp. 234-235, 2019.
[307] A. Wexler, "Separating neuroethics from neurohype," *Nature Biotechnology,* vol. 37, no. 9, pp. 988-990, 2019.
[308] M. Marchetti, and K. Priftis, "Brain–computer interfaces in amyotrophic lateral sclerosis: A metanalysis," *Clinical Neurophysiology,* vol. 126, no. 6, pp. 1255-1263, 2015.
[309] P. Wierzgała, D. Zapała, G. M. Wojcik, and J. Masiak, "Most Popular Signal Processing Methods in Motor-Imagery BCI: A Review and Meta-Analysis," *Frontiers in Neuroinformatics,* vol. 12, 2018.
[310] C. Eiber, J. Delbeke, J. Cardoso, M. de Neeling, S. John, C. W. Lee, J. Skefos, A. Sun, D. Prodanov, and Z. McKinney, "Preliminary Minimum Reporting Requirements for In-Vivo Neural Interface Research: I. Implantable Neural Interfaces," *IEEE Open Journal of Engineering in Medicine and Biology,* vol. 2, pp. 74-83, 2021.
[311] C. Easttom, L. Bianchi, D. Valeriani, C. S. Nam, A. Hossaini, D. Zapala, A. Roman-Gonzalez, A. K. Singh, A. Antonietti, G. Sahonero-Alvarez, and P. Balachandran, "A Functional Model for Unifying Brain Computer Interface Terminology," *IEEE Open Journal of Engineering in Medicine and Biology,* vol. 2, pp. 91-96, 2021.
[312] C. Easttom, "BCI glossary and functional model by the IEEE P2731 working group," *Brain-Computer Interfaces*, pp. 1-3, 2021.
[313] D. Zapała, A. Hossaini, M. Kianpour, G. Sahonero-Alvarez, and A. Ayesh, "A functional BCI model by the P2731 working group: psychology," *Brain-Computer Interfaces*, pp. 1-10, 2021.
[314] L. Bianchi, A. Antonietti, G. Bajwa, R. Ferrante, M. Mahmud, and P. Balachandran, "A functional BCI model by the IEEE P2731 working group: data storage and sharing," *Brain-Computer Interfaces*, pp. 1-9, 2021.
[315] D. Eke, A. Bernard, J. G. Bjaalie, R. Chavarriaga, T. Hanakawa, A. J. Hannan, S. L. Hill, M. Martone, A. McMahon, O. Ruebel, E. Thiels, and F. Pestilli, "International Data Governance for Neuroscience," PsyArXiv, 2021.
[316] M. Ienca, P. Haselager, and E. J. Emanuel, "Brain leaks and consumer neurotechnology," *Nat Biotechnol,* vol. 36, no. 9, pp. 805-810, 09, 2018.
[317] H. Garden, D. Winickoff, N. M. Frahm, and S. Pfotenhauer, "Responsible innovation in neurotechnology enterprises," *OECD Science, Technology and Industry Working Papers,* vol. 05, 2019.

[318] L. R. Hochberg, M. D. Serruya, G. M. Friehs, J. A. Mukand, M. Saleh, A. H. Caplan, A. Branner, D. Chen, R. D. Penn, and J. P. Donoghue, “Neuronal ensemble control of prosthetic devices by a human with tetraplegia,” *Nature,* vol. 442, no. 7099, pp. 164-171, 2006.
[319] L. R. Hochberg, D. Bacher, B. Jarosiewicz, N. Y. Masse, J. D. Simeral, J. Vogel, S. Haddadin, J. Liu, S. S. Cash, P. van der Smagt, and J. P. Donoghue, “Reach and grasp by people with tetraplegia using a neurally controlled robotic arm,” *Nature,* vol. 485, no. 7398, pp. 372-375, 2012.
[320] F. Sauter-Starace, D. Ratel, C. Cretallaz, M. Foerster, A. Lambert, C. Gaude, T. Costecalde, S. Bonnet, G. Charvet, T. Aksenova, C. Mestais, A. L. Benabid, and N. Torres-Martinez, “Long-Term Sheep Implantation of WIMAGINE(®), a Wireless 64-Channel Electrocorticogram Recorder,” *Front Neurosci,* vol. 13, pp. 847, 2019.
[321] C. Larzabal, V. Auboiroux, S. Karakas, G. Charvet, A. L. Benabid, S. Chabardès, T. Costecalde, and S. Bonnet, “The Riemannian Spatial Pattern method: mapping and clustering movement imagery using Riemannian geometry,” *J Neural Eng*, Mar 26, 2021.
[322] D. M. Brandman, T. Hosman, J. Saab, M. C. Burkhart, B. E. Shanahan, J. G. Ciancibello, A. A. Sarma, D. J. Milstein, C. E. Vargas-Irwin, B. Franco, J. Kelemen, C. Blabe, B. A. Murphy, D. R. Young, F. R. Willett, C. Pandarinath, S. D. Stavisky, R. F. Kirsch, B. L. Walter, A. Bolu Ajiboye, S. S. Cash, E. N. Eskandar, J. P. Miller, J. A. Sweet, K. V. Shenoy, J. M. Henderson, B. Jarosiewicz, M. T. Harrison, J. D. Simeral, and L. R. Hochberg, “Rapid calibration of an intracortical brain-computer interface for people with tetraplegia,” *J Neural Eng,* vol. 15, no. 2, pp. 026007, Apr, 2018.
[323] A. Gunduz, E. Opri, R. Gilron, V. Kremen, G. Worrell, P. Starr, K. Leyde, and T. Denison, “Adding wisdom to 'smart' bioelectronic systems: a design framework for physiologic control including practical examples,” *Bioelectron Med (Lond),* vol. 2, no. 1, pp. 29-41, Mar, 2019.
[324] B. Z. Allison, "Toward ubiquitous BCIs," *Brain-Computer Interfaces*, The Frontiers Collection G. B., P. G. and A. B., eds., pp. 357-387, Berlin, Heidelberg: Springer, 2009.
[325] S. Burwell, M. Sample, and E. Racine, “Ethical aspects of brain computer interfaces: a scoping review,” *BMC Med Ethics,* vol. 18, no. 1, pp. 60, Nov 09, 2017.
[326] P. Haselager, G. Mecacci, and A. Wolkenstein, "Can BCIs Enlighten the Concept of Agency? A Plea for an Experimental Philosophy of Neurotechnology," *Clinical Neurotechnology meets Artificial Intelligence: Philosophical, Ethical, Legal and Social Implications*, O. Friedrich, A. Wolkenstein, C. Bublitz, R. J. Jox and E. Racine, eds., p. 55, Cham, Switzerland: Springer Nature, 2021.
[327] M. J. Schneider, J. J. Fins, and J. R. Wolpaw, "Ethical issues in BCI research," *Brain–Computer Interfaces. Principles and Practice*, J. R. Wolpaw and E. W. Wolpaw, eds., pp. 373-383, New York, New York, USA: Oxford University Press, 2012.
[328] R. Yuste, S. Goering, B. A. Y. Arcas, G. Bi, J. M. Carmena, A. Carter, J. J. Fins, P. Friesen, J. Gallant, J. E. Huggins, J. Illes, P. Kellmeyer, E. Klein, A.

Marblestone, C. Mitchell, E. Parens, M. Pham, A. Rubel, N. Sadato, L. S. Sullivan, M. Teicher, D. Wasserman, A. Wexler, M. Whittaker, and J. Wolpaw, “Four ethical priorities for neurotechnologies and AI,” *Nature,* vol. 551, no. 7679, pp. 159-163, 11 08, 2017.

[329] F. Nijboer, J. Clausen, B. Z. Allison, and P. Haselager, “The Asilomar Survey: Stakeholders' Opinions on Ethical Issues Related to Brain-Computer Interfacing,” *Neuroethics,* vol. 6, pp. 541-578, 2013.

[330] M. Pham, S. Goering, M. Sample, J. E. Huggins, and E. Klein, “Asilomar survey: researcher perspectives on ethical principles and guidelines for BCI research,” *Brain-Computer Interfaces,* vol. 5, no. 4, pp. 97-111, 2018.

[331] M. Sample, S. Sattler, S. Blain-Moraes, D. Rodríguez-Arias, and E. Racine, “Do Publics Share Experts’ Concerns about Brain–Computer Interfaces? A Trinational Survey on the Ethics of Neural Technology,” *Science, Technology, & Human Values,* vol. 45, no. 6, pp. 1242-1270, 2020.

[332] C. Cinel, D. Valeriani, and R. Poli, “Neurotechnologies for Human Cognitive Augmentation: Current State of the Art and Future Prospects,” *Front Hum Neurosci,* vol. 13, pp. 13, 2019.

[333] A. Matran-Fernandez, and R. Poli, “Brain-Computer Interfaces for Detection and Localization of Targets in Aerial Images,” *IEEE Trans Biomed Eng,* vol. 64, no. 4, pp. 959-969, 04, 2017.

[334] A. Matran-Fernandez, and R. Poli, “Towards the automated localisation of targets in rapid image-sifting by collaborative brain-computer interfaces,” *PLoS One,* vol. 12, no. 5, pp. e0178498, 2017.

[335] R. Poli, D. Valeriani, and C. Cinel, “Collaborative brain-computer interface for aiding decision-making,” *PLoS One,* vol. 9, no. 7, pp. e102693, 2014.

[336] D. Valeriani, C. Cinel, and R. Poli, “Group Augmentation in Realistic Visual-Search Decisions via a Hybrid Brain-Computer Interface,” *Sci Rep,* vol. 7, no. 1, pp. 7772, 08, 2017.

[337] D. Valeriani, R. Poli, and C. Cinel, “Enhancement of Group Perception via a Collaborative Brain-Computer Interface,” *IEEE Trans Biomed Eng,* vol. 64, no. 6, pp. 1238-1248, 06, 2017.

[338] D. Valeriani, and R. Poli, “Cyborg groups enhance face recognition in crowded environments,” *PLoS One,* vol. 14, no. 3, pp. e0212935, 2019.

[339] A. Curtin, H. Ayaz, Y. Tang, J. Sun, J. Wang, and S. Tong, “Enhancing neural efficiency of cognitive processing speed via training and neurostimulation: An fNIRS and TMS study,” *Neuroimage,* vol. 198, pp. 73-82, 09, 2019.

[340] R. McKendrick, B. Falcone, M. Scheldrup, and H. Ayaz, “Effects of Transcranial Direct Current Stimulation on Baseline and Slope of Prefrontal Cortex Hemodynamics During a Spatial Working Memory Task,” *Front Hum Neurosci,* vol. 14, pp. 64, 2020.

[341] R. P. Rao, A. Stocco, M. Bryan, D. Sarma, T. M. Youngquist, J. Wu, and C. S. Prat, “A direct brain-to-brain interface in humans,” *PLoS One,* vol. 9, no. 11, pp. e111332, 2014.

[342] Y. Yang, S. Qiao, O. G. Sani, J. I. Sedillo, B. Ferrentino, B. Pesaran, and M. M. Shanechi, “Modelling and prediction of the dynamic responses of large-

scale brain networks during direct electrical stimulation," *Nat Biomed Eng,* vol. 5, no. 4, pp. 324-345, 04, 2021.
[343] M. M. Shanechi, "Brain-machine interfaces from motor to mood," *Nat Neurosci,* vol. 22, no. 10, pp. 1554-1564, 10, 2019.
[344] F. Dehais, Karwowski, W., and H. Ayaz, "Brain at Work and in Everyday Life as the Next Frontier: Grand Field Challenges for Neuroergonomics," *Front Neuroergonomics*, 2020.
[345] *Neuroergonomics: The brain at work and in everyday life*, London: Academic Press, an imprint of Elsevier, 2018.
[346] N. Kosmyna, and P. Maes, "AttentivU: An EEG-Based Closed-Loop Biofeedback System for Real-Time Monitoring and Improvement of Engagement for Personalized Learning," *Sensors (Basel),* vol. 19, no. 23, Nov, 2019.
[347] J. Van Dijk, and C. Hummels, "Beyond Distributed Representation: Embodied Cognition Design Supporting Socio-Sensorimotor Couplings."
[348] J. Frey, M. Hachet, and F. Lotte, "EEG-based neuroergonomics for 3d user interfaces: opportunities and challenges," *Le travail humain,* vol. 80, no. 1, pp. 73-92, 2017.
[349] C. Jeunet, D. Hauw, and J. d. R. Millán, "Sport Psychology: Technologies Ahead," *Frontiers in Sports and Active Living,* vol. 2, no. 10, 2020-February-13, 2020.
[350] S. di Fronso, P. Fiedler, G. Tamburro, J. Haueisen, M. Bertollo, and S. Comani, "Dry EEG in Sports Sciences: A Fast and Reliable Tool to Assess Individual Alpha Peak Frequency Changes Induced by Physical Effort," *Front Neurosci,* vol. 13, pp. 982, 2019.
[351] D. E. Callan, and F. Dehais, "Chapter 9 - Neuroergonomics for Aviation," *Neuroergonomics*, H. Ayaz and F. Dehais, eds., pp. 55-58: Academic Press, 2019.
[352] S. H. Fairclough, and F. Lotte, "Grand Challenges in Neurotechnology and System Neuroergonomics," *Frontiers in Neuroergonomics,* vol. 1, no. 2, 2020-November-30, 2020.
[353] D. B. Stone, G. Tamburro, P. Fiedler, J. Haueisen, and S. Comani, "Automatic Removal of Physiological Artifacts in EEG: The Optimized Fingerprint Method for Sports Science Applications," *Frontiers in Human Neuroscience,* vol. 12, no. 96, 2018-March-21, 2018.
[354] M. I. Casso, C. Jeunet, and R. N. Roy, "Heading for motor imagery brain-computer interfaces (MI-BCIs) usable out-of-the-lab: Impact of dry electrode setup on classification accuracy." pp. 690-693.
[355] F. Dehais, A. Duprès, S. Blum, N. Drougard, S. Scannella, R. N. Roy, and F. Lotte, "Monitoring Pilot's Mental Workload Using ERPs and Spectral Power with a Six-Dry-Electrode EEG System in Real Flight Conditions," *Sensors (Basel),* vol. 19, no. 6, Mar 16, 2019.
[356] C. Mühl, D. Heylen, and A. Nijholt, "Affective brain-computer interfaces: neuroscientific approaches to affect detection
" *Oxford Handbook of Affective Computing*, R. Calvo, S. D'Mello, J. Gratch and A. Kappas, eds., pp. 217-237, New York, NY, USA: Oxford University Press, 2015.

[357] A. Nijholt, *Brain Art. Brain-Computer Interfaces and Artistic Expression*, London, UK: Springer, 2019.
[358] A. Wadeson, A. Nijholt, and C. S. Nam, “Artistic Brain-Computer Interfaces: Current State-of-Art of Control Mechanisms,” *Brain-Computer Interfaces,* vol. 2, no. 2-3, pp. 70-75, 2015.
[359] M. Prpa, and P. Pasquier, " Brain-Computer Interfaces in Contemporary Art: A State of the Art and Taxonomy," *Brain art. Brain-computer interfaces and artistic expression*, Human-computer interaction series A. Nijholt, ed., pp. 65-115, London (UK): Springer, 2019.
[360] J. L. King, *Art therapy, trauma, and neuroscience : theoretical and practical perspectives*: Routledge, 2016.
[361] S. M. Scott, and L. Gehrke, "Neurofeedback during creative expression as a therapeutic tool," *Mobile Brain–Body Imaging and the Neuroscience of Art, Innovation and Creativity*, Bio- and Neurosystems J. L. Contreras-Vidal, D. Robleto, J. G. Cruz-Garza, J. M. Azorin and C. S. C.S. Nam, eds., pp. 161-166, Cham, Switzerland: Springer Nature, 2019.
[362] S. M. Scott, C. Raftery, and C. Anderson, "Advancing the rehabilitative and therapeutic potentials of BCI and noninvasive systems," *Brain art: Brain-computer interfaces for artistic expression*, A. Nijholt, ed., pp. 327-354, Switzerland: Springer International Publishing, 2019.
[363] A. Valjamae, L. Evers, B. Z. Allison, J. Ongering, A. Riccio, I. Igardi, and D. Lamas, "The BrainHack Project."
[364] C. Guger, B. Allison, M. Walchshofer, and S. Brienbauer, "The BR4IN.IO Hackathons," *Brain art. Brain-computer interfaces and artistic expression*, Human-computer interaction series N. A, ed., pp. 447-473, London (UK):: Springer, 2019.
[365] A. Väljamäe. "TLÜ 13: Neurotheatre as a research tool by Aleksander Väljamäe," 2021; https://www.youtube.com/watch?v=L4xA0vU5XtY.
[366] R. Ramchurn, S. Martindale, M. Wilson, S. Benford, and A. Chamberlain, "Brain-Controlled Cinema," *Brain Art. Brain-computer interfaces and artistic expression*, A. Nijholt, ed., pp. 377-408, Cham, Switzerland: Springer, 2019.
[367] D. Rosenboom, and T. Mullen, "More Than One—Artistic Explorations with Multi-agent BCIs," *Brain Art. Brain-computer interfaces and artistic expression.*, A. Nijholt, ed., pp. 117-143, Cham, Switzerland: Springer, 2019.
[368] D. Williams, "Evaluating BCI for Musical Expression: Historical Approaches, Challenges and Benefits," *Brain art. Brain-computer interfaces and artistic expression*, A. Nijholt, ed., pp. 145-158., Cham, Switzerland: Springer, 2019.
[369] E. Hildt, “Affective Brain-Computer Music Interfaces-Drivers and Implications,” *Front Hum Neurosci,* vol. 15, pp. 711407, 2021.
[370] E. Pearlman, “Brain Opera: Exploring Surveillance in 360-degree Immersive Theatre,” *PAJ: A Journal of Performance and Art,* vol. 39, no. 2, pp. 79-85, 2017.
[371] E. Pearlman, “AI Comes of Age,” *PAJ: A Journal of Performance and Art,* vol. 42, no. 3, pp. 55-62, 2020.

[372] E. Pearlman, ellenluminescense@gmail.com, R. U. Ellen Pearlman, 2 Durbas Iela, Riga, Lativa, and C. MIT OpenDoc Lab, MA, U.S.A. Email . “Is There a Place in Human Consciousness Where Surveillance Cannot Go? Noor: A Brain Opera,” *Leonardo*, pp. 542-546, 2021.

Data Centrism and the Core of Data Science as a Scientific Discipline

Thilo Stadelmann, Tino Klamt and Philipp H. Merkt

Abstract Data science is one of the most significant developments in computing in the 21^{st} century. It is also described as a discipline in the making, drawing principles, methods and tools from established fields like computer science, statistics, science, business, politics, and any domain with adequate data. What are data science's underlying principles and techniques (models, methods) that are applicable across different use cases and fields of application? What novel aspect of science underlies this emerging discipline? We argue that it is *data centrism* – the reliance on data itself, in mindset, methods and products – that makes data science more than the sum of its parts, as this is not done in any other discipline.

Thilo Stadelmann
ZHAW Centre for Artificial Intelligence and ZHAW Data Science Laboratory, Winterthur, Switzerland
✉ stdm@zhaw.ch

Tino Klamt
University of Greifswald, Greifswald, Germany
✉ tino.klamt@stud.uni-greifswald.de

Philipp H. Merkt
Carl Remigius Medical School, Research Group Emergency Medicine, Idstein, Germany
✉ philipp.merkt@carl-remigius.de

ARCHIVES OF DATA SCIENCE, SERIES A
(ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING
Vol. -, No. -, -

DOI 10.5445/KSP/XXXXXXXX/XX
ISSN 2363-9881

1 Introduction

Data science has been defined previously as *"a unique blend of principles and methods from analytics, engineering, entrepreneurship and communication that aim at generating value from the data itself"* (Stadelmann et al, 2019a). A similar notion was conveyed by Stadelmann et al (2013) and later refined in (Stadelmann et al, 2019b) when by referring to the data scientist the authors actually defined the activity of doing data science as being determined by what is taken out of the contributing disciplines (see Figure 1).

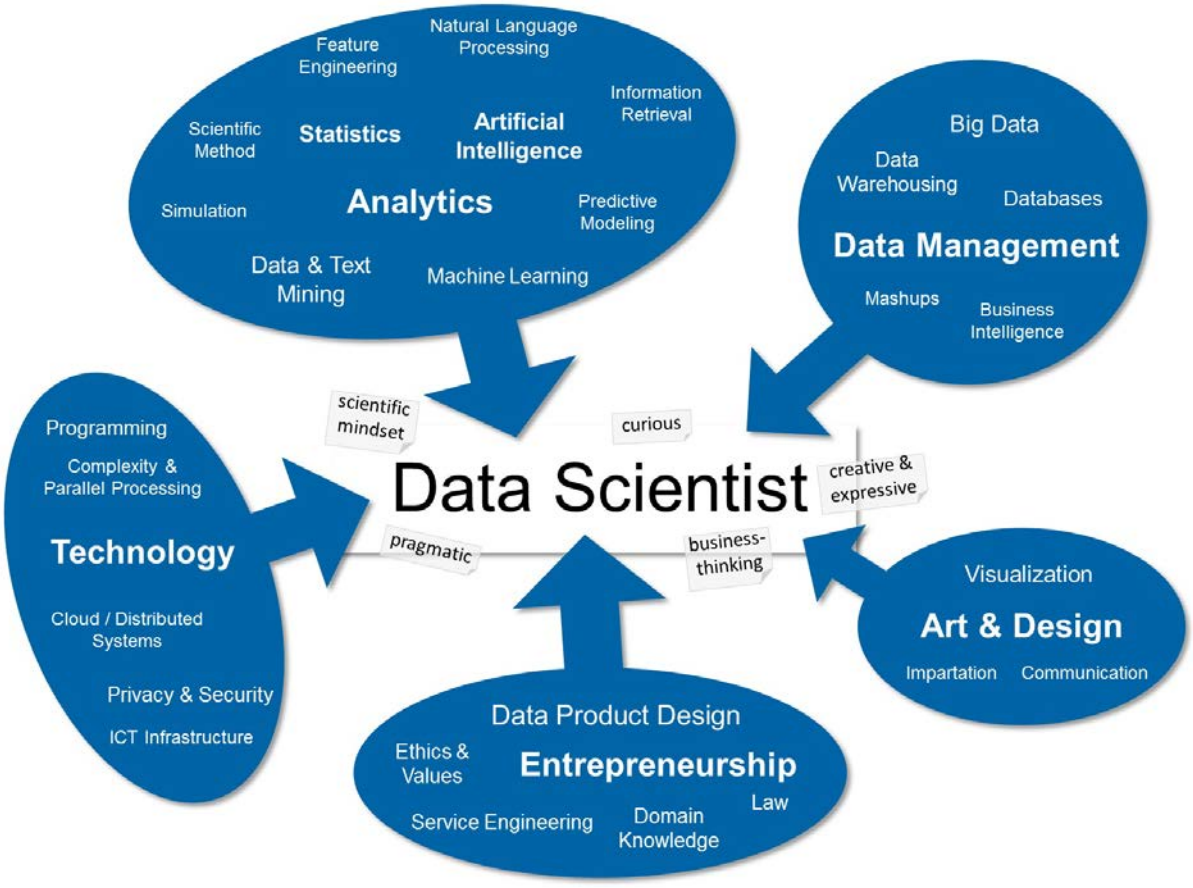


Fig. 1 The definition of a data scientist and, by implication, of the activity of doing data science, according to Stadelmann et al (2019b) (used with permission). In this paper, we argue that data science can *not* be defined merely as a unique cut of contributions from such contributing disciplines – it needs to have a scientific core of its own to warrant the designation of a discipline.

Now, several years after the main wave of the data science hype, one could ask heretically: "what remains of this 'discipline in the making' (Brodie, 2019b) if all there is in novelty is foremost a contribution to or from one of its constituting disciplines?" For example, when a data scientist develops a new analytical

method, it will foremost be a novelty in the field of statistics or machine learning, not specifically in data science. “No scientific discipline” would be the correct answer, if there wasn’t more than a selection of contributions from other fields – if there wasn’t more to data science than the sum of its parts (Denning, 2005). Data science needs to contribute theories of its own that must be falsifiable (Popper, 1961) to warrant the designation of a science.

In this paper, we argue that there needs to be a scientific core of data science that is (or: is going to become) unique to data science, i.e., that is not the core issue in one of the contributing disciplines. We introduce our proposal for this core in Section 2, followed by an example from medical data analysis practice in Section 3 to illustrate the point. We then discuss limitations of this proposal in Section 4, which might indicate that this view is only partial, and draw conclusions in Section 5.

2 Data centrism

Naturally, this disciplinary core of data science has to materialize in aspects that transcend what was taken out of the contributing disciplines. It needs to amount to more than the adoption of singular methods and tools by

(a) designating a unique object (or: phenomenon) of study (Denning, 2013) as well as by
(b) containing an overarching principle under which this study is performed (Denning, 2005).

Regarding (a), we agree with previous definitions like (Dhar, 2013; Luna-Reyes, 2018; Braschler et al, 2019) and others that the object of study in data science is the creation of value from data. With respect to (b), it is our view that the overarching principle is “data centrism”.

2.1 Data centrism and other disciplines

By data centrism we mean that data science, in contrast to the contributing disciplines, puts the highest value on data *itself*, by making the data itself central to the data-scientific mindset (source of inspiration), the conduct of doing data science (processes and methods) and its outcome (data products and predictions).

We believe this aspect to be the core of data science because it firmly differentiates data science from related fields, as is demonstrated by the following exemplary consideration of such related fields.

Machine learning revolves around *learning from data* (not data itself): principles and methods to gain general knowledge out of finite data (Samuel (1959) put the highest weight on the learning outcome itself in his famous definition and neglected the input entirely). Despite the efforts of Andrew Ng to teach the field otherwise (Ng, 2021), this is still mainly a model-centric endeavour, i.e., conferences, sub-fields and projects revolve around model architectures as the centre pieces. Then, suitable data to satisfy the needs of the predominantly supervised modeling approaches has to be delivered for machine learners to usually take up the work. It is arguably the influence of data science that un- and semi-supervised methods are increasingly researched and used in recent years: Unsupervised learning was for a long time mainly equated to clustering (Mitchell, 1997). The rise of unsupervised learning as, e.g., spearheaded by Meta's Yann LeCun (LeCun and Misra, 2021), coincides with the rise of data-driven companies like Meta's Facebook and their needs as addressed by data science.

Statistics is concerned with *quantifying data*: its distribution, variability, the certainty of predictions, etc. Data thereby is the main object of analysis, while models again are the center of thinking and acting as well as the main outcome (Breiman, 2001). Specifically, the main stream of statistics revolves around certain modeling assumptions (e.g., linearity, normal error distribution $(0, \sigma^2)$) to which the data has to comply in order to permit claims to be made.

Data management cares for proper *processing of data* in an efficient, reliable and accessible fashion. Again, data is the object under focus, while algebra provides the theoretical backdrop for modelling, machine learning may provide means for optimizing queries (Heitz and Stockinger, 2019) and tools may provide support for data integration (Stonebraker et al, 2013; Stadelmann et al, 2015). Data here (as before) is not the subject determining the course, but merely the object of study under the specific perspective of manageability.

Service engineering secures *value creation from data*: not just commercially, but for all stakeholders of the value chain, including providers and customers. It thus puts the pains and gains of all stakeholders at the center (Meierhofer et al, 2019), making data a natural resource rather than the centerpiece.

The list could be continued to include all major disciplines mentioned in Figure 1 as contributors to data science, but the pattern is already established, at least on an intuitive scale: These disciplines have data as an object of study

(to varying degrees). In contrast, data science has data at the centre, as the subject (or: the driving force), and methods are employed that expect everything from the data itself (e.g., structure, patterns, supervision, value). Specifically, data science is the science of studying the data *as is*: it doesn't impose assumptions on the quality or quantity of data before its methods can be applied, but seeks methods that can make the most out of the data *that is available*. This is what is implied in having "value-creation from [actual] data" as the focus of the discipline. It includes both the current data at hand, but also data that can realistically be produced by improved data acquisition and preprocessing methods.

2.2 The effects of data at the centre

The unique point of view upheld by data science, hence, and in contrast to any of the contributing disciplines, is the one that looks for supreme value *in* the data itself (and not just *out* of it, as one ingredient). The distinction is subtle, but crucial: "in" the data means that data is the main ingredient, the centerpiece, at the same time ultima ratio and conditio sine qua non. On the other hand, by "out" of data we mean that data is a mere resource in the pursuit of some further end. The difference can be likened to a private horse owner who sees value *in* a horse (e.g., relational value), in contrast to a farmer of old who saw value *out* of a horse (as a means to pull a plow). Let's exemplify how data science implements this principle with a couple of examples.

Empiricism is the driving force in data science: in contrast to pre-conceived models of reality, data science reinforces the *mindset* to establish theories out of the patterns that arise from potentially vast amounts of data (i.e., empirical evidence rather than human intuition) (Hey et al, 2009). The effect of this is that data science models tend to become complex and opaque, as they didn't originate in a simple human idea, but emerged in a data-driven way. Deep learning methods are a good example for this, and the recent trend to research and apply *explainable and trustworthy methods* (Samek et al, 2019; Amirian et al, 2021) can be seen as a direct reaction to the data science mindset: If the data itself is determining the model, the discipline responsible for this development, as a next step, has to provide methods that make this machine-conceived models again amenable to human intuition, decision and control.

Learning from less (e.g., less data with as little as possible human-provided interpretations/supervision) can guide the learning of decision-making functions out of mere observations and probably also should do so in order to avoid human-introduced biases (Glüge et al, 2020; Wehrli et al, 2021). While this naturally employs machine learning methods, it is the mindset of data science (seeking a solution that relies on data alone instead of human annotations) that prompts the selection of *un- and semi-supervised methods* and not vice versa. Additionally, such methods are also applied by data scientists to gain models (and out of them value) from obviously imperfect data sets. It is again the data science mindset that asks "what can be done to exploit the actual data best" rather than "who can bring me better data or labels to train my method" (Hollenstein et al, 2019; Simmler et al, 2021).

Data products are outcomes (digital services, physical products or anything in between) that have data at their core (Loukides, 2011) and not just as an ingredient. While again certain methods from the contributing disciplines are necessary conditions for them to function, a prime candidate being service engineering (Meierhofer et al, 2019), only by adding the data itself the sufficient conditions for value generation are met. Hence, they derive their added value from the added data.

2.3 Data centrism in the literature

This list in Section 2.2 could (and should) be extended as well to establish the pattern more strongly. However, intuitively, what the list resembles is the same mindset reinforced several times in the 2020-2021 issues of Andrew Ng's "The Batch" (DeepLearning.AI editorial team, 2021) of thinking data-centric rather than {model, user, customer, theory, application, ...}-centric. Similar arguments are provided for example by Della Corte and Della Corte (2021) and Gerdes (2021).

Putting data at the centre of *thinking* (i.e., assuming data necessary (Jeffreys and Jeffreys, 1988) for the realization of the expected added value, and data plus data science methods sufficient), has been already hinted at in Hey et al (2009) for applications in the sciences, and is of course discussed in contributing disciplines like machine learning (Ng, 2021; Ng et al, 2021). Data centrism has further been discussed (and partially been dismissed) as a guiding principle for physical computer network organisation (Shenker, 2003) and server design

(Siegl et al, 2016), database (Haas et al, 2011) and middleware development (Chen et al, 2008) as well as the build-up of whole embedded (Alvarez-Coello et al, 2021) and enterprise software architectures (Rajabi and Abade, 2012).

However, the furthering of the data-centric mindset as the core of a scientific discipline on a broader scale within the data-related community, with the *subsequent* consideration within the contributing disciplines to data science in recent years (Lau et al, 2018; Nwokeji et al, 2015; Ng et al, 2021), is arguably the effect and contribution of data science. This view is shared by Leonelli (2019) and Fekete et al (2021). However, while we are concerned here with a proper *delineation* of the fields of science and technology such as the ones identified by (Braschler et al, 2019) as being contributors to data science (cp. Figure 1), Leonelli presents a philosophical analysis of data-centric research, and Fekete and colleagues are concerned with data science teaching.

3 An example from practice

To illustrate the contrasting approaches in data science and related disciplines, an example is presented from resilience research. It is a prototypical example of a use case that could build on multiple highly different data sources, which would require different methodology to exploit them, leading to different research outcomes in terms of type and scope.

The example research is concerned with increasing the resilience of emergency workers from heterogeneous professional backgrounds such as fire brigade, rescue service, police, military and NGOs, to stressful situations. This comprises answering the two questions of (a) how to effectively and efficiently (i.e., practically possible for professionals in service) measure stress under realistic conditions, and (b) how to increase the resilience to such stress by interventions like individual trainings. The setup for this research in a first phase is as follows (with the prospect to scale up to larger samples in the next phase): over a period of 72 hours, a group of ca. 20 participants are cast into a series of role-playing scenarios belonging to a fictitious foreign catastrophe situation (Merkt and Wilk-Vollmann, 2021). In these scenarios, they face constantly increasing challenges of asymmetric threat (cp. Figure 2) while data is being recorded. Specifically, all radio traffic is recorded, physiological parameters are taken (heart rate; blood pressure; blood gas analysis for lactate, base excess, glucose; and neurophysiological biomarkers like cortisol and α-amylase), and

questionnaires for subjective assessment of the stress level are taken based on standardized interview settings.

Fig. 2 Example of a catastrophe scenario as used in the described resilience research (Merkt and Wilk-Vollmann, 2021): Role play is used to create realistic, stressful crisis situations; data is collected during and after the scenarios from the participants to reflect their stress level (picture shows one of the authors). Copyright © by Stefan Mikolon (used with permission).

Typical resilience research would focus on structured questionnaires as data sources to account for human factors in the dealing with stress (Merkt et al, 2020), evaluating them using a qualitative research approach based on Grounded Theory (Adolph et al, 2011). The advantage of these methods lies in the inductive development of categories and theories. This means that the heterogeneous and complex situations within catastrophe scenarios that cannot be standardized beforehand could be dealt with very individually. As part of the qualitative content analysis according to Mayring (2015), which is based on the Grounded Theory, the inductive theory formation is specified by a concrete methodological analysis process. The core of this process is the coding of individual statements, aiming at assigning the interview content to different categories. These categories, in turn, are validated as part of a reliability test on the basis of various statistical measures, after which an evaluation and interpretation takes place. This is the strength of the qualitative, social science approach, which is based on a formal, structured process of data acquisition.

However, when this resilience research project enters the next phase, it has to scale up to thousands of participants, not only in controlled settings of role-playing scenarios, but in emergency operations in practice. As there is simply

no way of getting structured, standardized questionnaire data from all subjects in practice, a data-centric approach rooted in data science is a valid alternative: Subjects are equipped with few easily manageable sensors and post-hoc stress analysis is attempted with the data that these deliver. Additionally, communication under stress reveals a lot about the communicators' stress level, so it is worthwhile to decode the radio communication using AI-based emotion recognition (Biondi et al, 2017). While qualitative methods might in principle deliver more meaningful results based on smaller samples, such methods are excluded by the use case. Only a data science approach with its mindset of "creating value from the actual data" can lead to any result, where "actual" data is the data either readily available or at least realistically producible.

4 Limitations

Focusing on a single aspect is necessary for any detailed study, and identifying the core of an emerging scientific discipline is no exception. We are convinced that data centrism as discussed above is of utmost importance to the scientific core of data science in the sense that it serves as a focal point in deciding what is data science and what is part of a contributing discipline. However, we do not see clearly enough yet if this is the scientific core itself, or some inner ring around it.

Particularly, the following duality illustrates that zooming in too much on data as a subject in data science rather than mere object of study can be misleading in the limit: Making data the "subject that determines the course" naturally assumes given data as the starting point of data science endeavours, and we have presented examples above that illustrate the importance of data science in working with the data *one has*, the given data, to subsequently research and apply methods that make the most of it rather than dismissing it.

However, already the (real) use case in Section 3 shows that also a data-centric approach rooted in data science has to take into consideration the source, acquisition and quality improvement of data. It will develop adequate methods for this distinct from data acquisition methods in, e.g., qualitative analysis. But this case shows that equating data centrism with "creating value out of *given* data" falls short of the scope of data science and the power of the data-centric paradigm: Data science does contain methods, data-centric methods, to improve on the data by getting more adequate data. Such methods for example

analyze the data at hand, realize shortcomings, and prompt users for specific improvements such as filling gaps in the coverage of the data set (guided by data, aimed at data – thus having data at the centre) or create new synthetic samples as in data augmentation (Shorten and Khoshgoftaar, 2019). We thus chose to refer to data science as the discipline dealing with *actual* data (cp. end of Section 3) rather than idealized data (idealization that happens, e.g., when assuming Gaussianty, as discussed by Li (2007)).

On a more fundamental level, having data science as a discipline that puts supreme value in actual data (rather than, e.g., human theories on the causes of this data) opens the door to all kind of problems inherited from this data: The data might be biased (Wehrli et al, 2021) and thus barely suitable to build models on it; it might, in the absence of any theory on its origin and requirements on its quality, give rise to models that find spurious patterns and consequently produce models of machine magical thinking (Diaconis, 2006). It might not find any value at all because the data, in combination with current methods, turns out to be insufficient to realize the added value. For all these – true, actual – risks of assumption-free data analysis, it is important to make data science not a replacement of other scientific disciplines, but an enrichment. If the more formal, less error-prone methods of statistics can be applied in a certain analysis, then this should be done; if causal analysis (Pearl, 2009) can be done and is important for the validity of the result, this should not be neglected. But if no other principle of analysis can be applied than data centrism, for practical or theoretical reasons, then it is important to have the best possible data science methods available. Mitchell (1997) proves that no learning is possible without assumptions; we argue that data science is home to those methods that deliberately work with the least possible amount of assumptions, which sometimes is the only viable route to take. Of course, such approaches can only detect correlations in the data and make no statements about causality (Cap, 2019). But while correlation is not causation, correlation often is enough (Brodie, 2019b,a; Stockinger et al, 2019). Hence, furthering data science as a data-centric discipline adds something unique to the quiver of scientific methodologies. The skilled hunter will carefully chose the appropriate arrow for each situation.

5 Conclusion

If the scientific core of data science is constituted of those aspects that put data at the core of thinking, acting and expectation, and if, next, methodology from other fields is assembled around this core as need arises, the following tentative list of novel areas of research (and the respective works therein) can arguably be seen as being genuine first-class citizens of the discipline of data science – the non-borrowed part of it:

Machine Learning Operations (MLOps): The discipline of machine learning could live well without taking care of operational issues for several decades (Mitchell, 1997). It is since the advent of data science and hence the data-centric paradigm that methods are created and community is formed to care for the development process including the operation of the complete data product pipeline, and the various feedbacks between them (Mäkinen et al, 2021).

Applied semi- and weakly-supervised learning: While the research of methods on how to learn from little supervision is core machine learning terrain inspired by findings in neuroscience (Zador, 2019), the application of such findings to data problems in industry, health, finance, retail, etc. is the domain and contribution of data science.

Data product design: The data product Loukides (2010) already appeared to be one of the outstanding contributions of data science in one of the first major courses on the subject (Howe, 2014).

Explainable Artificial Intelligence (XAI): Few other fields with a strong technical core have managed to incorporate overarching (societal) concerns into the discipline itself as well as data science has. Be it under the terms of explainable artificial intelligence, data ethics, {DataScience, AI}4Good or others, these developments wouldn't come out of the neighboring disciplines of AI or ethics without the mindset promoted by data science – data centrism. Only data centrism promotes methods that seek value from the data itself without deferring to humans for modeling decisions, which in turn creates the demand for new methods and frameworks for transparency, interpretation and ethical acting.

Future work will include a more thorough analysis of data centrism: Its origins and current traces, and if this confirms the view suggested here of data centrism being the scientific core of the discipline – and hence kingmaker of data science.

References

Adolph S, Hall W, Kruchten P (2011) Using grounded theory to study the experience of software development. Empirical Software Engineering 16(4):487–513, DOI 10.1007/s10664-010-9152-6

Alvarez-Coello D, Wilms D, Bekan A, Gómez JM (2021) Towards a data-centric architecture in the automotive industry. Procedia Computer Science 181:658–663, DOI 10.1016/j.procs.2021.01.215

Amirian M, Tuggener L, Chavarriaga R, Satyawan YP, Schilling FP, Schwenker F, Stadelmann T (2021) Two to trust: Automl for safe modelling and interpretable deep learning for robustness. In: Heintz F, Milano M, O'Sullivan B (eds) Trustworthy AI - Integrating Learning, Optimization and Reasoning, Springer International Publishing, Cham, pp 268–275, DOI 10.1007/978-3-030-73959-1_23

Biondi G, Franzoni V, Poggioni V (2017) A deep learning semantic approach to emotion recognition using the ibm watson bluemix alchemy language. In: International Conference on Computational Science and Its Applications, Springer, pp 718–729, DOI 10.1007/978-3-319-62398-6_51

Braschler M, Stadelmann T, Stockinger K (2019) Data science. In: Applied Data Science, Springer, pp 17–29, DOI 10.1007/978-3-030-11821-1_2

Breiman L (2001) Statistical modeling: the two cultures. with comments and a rejoinder by the author. Statist Sci 16(3):199–231, DOI 10.1214/ss/1009213726

Brodie ML (2019a) On developing data science. In: Applied Data Science, Springer International Publishing, Cham, pp 131–160, DOI 10.1007/978-3-030-11821-1_9

Brodie ML (2019b) What is data science? In: Applied Data Science, Springer, pp 101–130, DOI 10.1007/978-3-030-11821-1_8

Cap CH (2019) Risks and side effects of data science and data technology. In: Applied Data Science, Springer International Publishing, Cham, pp 79–95, DOI 10.1007/978-3-030-11821-1_6

Chen G, Li M, Kotz D (2008) Data-centric middleware for context-aware pervasive computing. Pervasive and mobile computing 4(2):216–253, DOI 10.1016/j.pmcj.2007.10.001

DeepLearningAI editorial team (2021) The batch. `https://www.deeplearning.ai/the-batch/`, [Online; accessed 23-June-2021]

Della Corte D, Della Corte KA (2021) The data-centric lab: a pharmaceutical perspective. In: Future of Information and Communication Conference, Springer, pp 1–15, DOI 10.1007/978-3-030-73103-8_1

Denning PJ (2005) Is computer science science? Communications of the ACM 48(4):27–31, DOI 10.1145/2447976.2447988

Denning PJ (2013) The science in computer science. Communications of the ACM 56(5):35–38, DOI 10.1145/2447976.2447988

Dhar V (2013) Data science and prediction. Communications of the ACM 56(12):64–73, DOI 10.1145/2500499

Diaconis P (2006) Theories of data analysis: From magical thinking through classical statistics. In: Exploring Data Tables, Trends, and Shapes, John Wiley & Sons, Ltd, chap 1, pp 1–36, DOI https://doi.org/10.1002/9781118150702.ch1

Fekete A, Kay J, Röhm U (2021) A data-centric computing curriculum for a data science major. In: Proceedings of the 52nd ACM Technical Symposium on Computer Science Education, pp 865–871, DOI 10.1145/3408877.3432457

Gerdes A (2021) A participatory data-centric approach to ai ethics by design. Applied Artificial Intelligence pp 1–19, DOI 10.1080/08839514.2021.2009222

Glüge S, Amirian M, Flumini D, Stadelmann T (2020) How (not) to measure bias in face recognition networks. In: IAPR Workshop on Artificial Neural Networks in Pattern Recognition, Springer, pp 125–137, DOI 10.1007/978-3-030-58309-5_10

Haas PJ, Maglio PP, Selinger PG, Tan WC (2011) Data is dead... without what-if models. Proceedings of the VLDB Endowment 4(12):1486–1489, DOI 10.14778/3402755.3402802

Heitz J, Stockinger K (2019) Join query optimization with deep reinforcement learning algorithms. arXiv preprint arXiv:191111689

Hey AJ, Tansley S, Tolle KM, et al (2009) The fourth paradigm: data-intensive scientific discovery, vol 1. Microsoft research Redmond, WA

Hollenstein L, Lichtensteiger L, Stadelmann T, Amirian M, Budde L, Meierhofer J, Füchslin RM, Friedli T (2019) Unsupervised learning and simulation for complexity management in business operations. In: Applied Data Science, Springer, pp 313–331, DOI 10.1007/978-3-030-11821-1_17

Howe B (2014) Introduction to data science. `https://www.classcentral.com/course/datasci-451`, [Online; accessed 02-February-2022]

Jeffreys H, Jeffreys BS (1988) §1.036: Necessary: Sufficient". In: Methods of Mathematical Physics, 3rd Edition, Cambridge University Press, pp 10–11

Lau FDH, Adams NM, Girolami MA, Butler LJ, Elshafie MZ (2018) The role of statistics in data-centric engineering. Statistics & Probability Letters 136:58–62, DOI 10.1016/j.spl.2018.02.035

LeCun Y, Misra I (2021) Self-supervised learning: The dark matter of intelligence. https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/, [Online; accessed 02-February-2022]

Leonelli S (2019) Data governance is key to interpretation: Reconceptualizing data in data science. Harvard Data Science Review 1(1), DOI 10.1162/99608f92.17405bb6

Li C (2007) Non-Gaussian, Non-stationary and Nonlinear Signal Processing Methods-with Applications to Speech Processing and Channel Estimation. Institut for Elektroniske Systemer, Aalborg Universitet

Loukides M (2010) What is data science? O'Reilly Media, Inc., URL https://www.oreilly.com/radar/what-is-data-science/, [Online; accessed 02-February-2022]

Loukides M (2011) The evolution of data products. O'Reilly Media, Inc., URL https://www.oreilly.com/radar/evolution-of-data-products/, [Online; accessed 02-February-2022]

Luna-Reyes LF (2018) The search for the data scientist: creating value from data. ACM SIGCAS Computers and Society 47(4):12–16, DOI 10.1145/3243141.3243145

Mäkinen S, Skogström H, Laaksonen E, Mikkonen T (2021) Who needs MLOps: What data scientists seek to accomplish and how can MLOps help? arXiv preprint arXiv:210308942

Mayring PAE (2015) Qualitative Inhaltsanalyse: Grundlagen und Techniken, 12. Auflage. Springer

Meierhofer J, Stadelmann T, Cieliebak M (2019) Data products. In: Applied Data Science, Springer, pp 47–61, DOI 10.1007/978-3-030-11821-1_4

Merkt PH, Wilk-Vollmann S (2021) Anspruchsvolle Übungslagen: Kommunikationsverhalten und Stressreaktionen. Rettungsdienst 1:14–18

Merkt PH, Wilk-Vollmann S, Wolz C (2020) Forschung in der Notfall- und Katastrophenmedizin. Taktik+Medizin 4:28–31

Mitchell TM (1997) Machine Learning. McGraw-Hill

Ng AY (2021) MLOps: From model-centric to data-centric AI. https://www.youtube.com/watch?v=06-AZXmwHjo

Ng AY, Aroyo L, Coleman C, Diamos G, Reddi VJ, Vanschoren J, Wu CJ, Zhou S (eds) (2021) Online Proceedings of the NeurIPS'21 Data-Centric AI Workshop. URL https://datacentricai.org/

Nwokeji JC, Clark T, Barn B, Kulkarni V, Anum SO (2015) A data-centric approach to change management. In: 2015 IEEE 19th International Enterprise Distributed Object Computing Conference, pp 185–190, DOI 10.1109/EDOC.2015.34

Pearl J (2009) Causality: Models, Reasoning, and Inference, 2nd edn. Cambridge University Press, Cambridge

Popper KR (1961) The Logic of Scientific Discovery. Basic Books, Inc., New York

Rajabi Z, Abade MN (2012) Data-centric enterprise architecture. International Journal of Information Engineering and Electronic Business 4(4):53, DOI 10.5815/ijieeb.2012.04.08

Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR (2019) Explainable AI: interpreting, explaining and visualizing deep learning, vol 11700. Springer Nature, DOI 10.1007/978-3-030-28954-6

Samuel AL (1959) Some studies in machine learning using the game of checkers. IBM Journal of research and development 3(3):210–229

Shenker S (2003) The data-centric revolution in networking. In: Proceedings 2003 VLDB Conference, Elsevier, p 15, DOI 10.1016/B978-012722442-8/50010-0

Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. Journal of Big Data 6(1):1–48, DOI 10.1186/s40537-019-0197-0

Siegl P, Buchty R, Berekovic M (2016) Data-centric computing frontiers: A survey on processing-in-memory. In: Proceedings of the Second International Symposium on Memory Systems, pp 295–308, DOI 10.1145/2989081.2989087

Simmler N, Sager P, Andermatt P, Chavarriaga R, Schilling FP, Rosenthal M, Stadelmann T (2021) A Survey of Un-, Weakly-, and Semi-Supervised Learning Methods for Noisy, Missing and Partial Labels in Industrial Vision Applications. In: 8th Swiss Conference on Data Science, IEEE, DOI 10.1109/SDS51136.2021.00012

Stadelmann T, Stockinger K, Braschler M, Cieliebak M, Baudinot G, Dürr O, Ruckstuhl A (2013) Applied data science in europe: Challenges for academia

in keeping up with a highly demanded topic. In: 9th European Computer Science Summit, Amsterdam, Niederlande, 8-9 October 2013

Stadelmann T, Cieliebak M, Stockinger K (2015) Toward automatic data curation for open data. ERCIM News 2015(100):32–33

Stadelmann T, Braschler M, Stockinger K (2019a) Introduction to applied data science. In: Braschler M, Stadelmann T, Stockinger K (eds) Applied Data Science, Springer, pp 3–16, DOI 10.1007/978-3-030-11821-1_1

Stadelmann T, Stockinger K, Bürki GH, Braschler M (2019b) Data scientists. In: Braschler M, Stadelmann T, Stockinger K (eds) Applied Data Science, Springer, pp 31–45, DOI 10.1007/978-3-030-11821-1_3

Stockinger K, Braschler M, Stadelmann T (2019) Lessons learned from challenging data science case studies. In: Applied Data Science, Springer International Publishing, Cham, pp 447–465, DOI 10.1007/978-3-030-11821-1_24

Stonebraker M, Bruckner D, Ilyas IF, Beskales G, Cherniack M, Zdonik SB, Pagan A, Xu S (2013) Data curation at scale: The data tamer system. In: Sixth Biennial Conference on Innovative Data Systems Research, CIDR 2013, Asilomar, CA, USA, January 6-9, 2013, Online Proceedings, www.cidrdb.org, URL `http://cidrdb.org/cidr2013/Papers/CIDR13_Paper28.pdf`

Wehrli S, Hertweck C, Amirian M, Glüge S, Stadelmann T (2021) Bias, awareness, and ignorance in deep-learning-based face recognition. AI and Ethics pp 1–14, DOI 10.1007/s43681-021-00108-6

Zador AM (2019) A critique of pure learning and what artificial neural networks can learn from animal brains. Nature communications 10(1):1–7, DOI 10.1038/s41467-019-11786-6

A Theory of Natural Intelligence

Christoph von der Malsburg[1,2,3*], Thilo Stadelmann[2,4†] and Benjamin F. Grewe[3†]

[1*]Frankfurt Institute for Advanced Studies, Frankfurt, Germany.
[2]Centre for Artificial Intelligence, Zurich University of Applied Sciences, Winterthur, Switzerland.
[3]Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland.
[4]European Centre for Living Technology, Venice, Italy.

*Corresponding author(s). E-mail(s): malsburg@fias.uni-frankfurt.de;
Contributing authors: stdm@zhaw.ch; bgrewe@ethz.ch;
†These authors contributed equally to this work.

Abstract

Introduction: In contrast to current AI technology, natural intelligence – the kind of autonomous intelligence that is realized in the brains of animals and humans to attain in their natural environment goals defined by a repertoire of innate behavioral schemata – is far superior in terms of learning speed, generalization capabilities, autonomy and creativity. How are these strengths, by what means are ideas and imagination produced in natural neural networks?
Methods: Reviewing the literature, we put forward the argument that both our natural environment and the brain are of low complexity, that is, require for their generation very little information and are consequently both highly structured. We further argue that the structures of brain and natural environment are closely related.
Results: We propose that the structural regularity of the brain takes the form of net fragments (self-organized network patterns) and that these serve as the powerful inductive bias that enables the brain to learn quickly, generalize from few examples and bridge the gap between abstractly defined general goals and concrete situations.
Conclusions: Our results have important bearings on open problems in artificial neural network research.

Keywords: Ontogenesis, emergence, structural regularity, net fragments, visual perception, scene representation, homeomorphic mapping, inductive bias, autonomous behavior

1 Introduction

There may be different kinds of intelligence. We here concentrate on the one that is epitomized in humans and animals. This kind of intelligence is often defined as the ability to successfully pursue general goals in varying contexts, goals such as feeding oneself, avoiding danger or creating offspring. The emphasis of our communication is on the neural mechanisms that generate this ability, our main point being that besides nature and nurture the process is dominated by a third generative factor, *emergence*. In this context, 'nature' refers to the influence of the genes and therewith to that of evolution, while 'nurture' to that of experience, instruction and education. We would like to maintain here that neither quantitatively nor qualitatively genes and experience alone can account for the structure of the nervous system nor the intelligence it supports, leaving a large gap to be closed by emergence.

On the quantitative side, as to 'nature', the human genome contains one gigabyte of information (3.3 billion nucleotides of DNA [1]) while one petabyte is required to describe the connectivity of the human brain[1]. In the case of humans, 'nurture' during the first years of life is provided for by an environment (the nursery, the family, toys, books etc.) that is deliberately kept simple and could be simulated in its visual aspects on the basis of a virtual reality program of a few gigabytes. Additionally, the rate at which humans absorb information into permanent memory is estimated [2] at only $1-2$ bits per second, signifying a couple of gigabits over a long lifetime. These amounts of information are to be compared to the petabyte needed to list all connections in the brain.

The qualitative side is the essence of the problem we want do address: how can intelligence, in terms of ideas, imaginations and insights surpass so much everything that has been 'programmed' into the genes, and how can it learn so fast and generalize so boldly beyond all the examples it has seen before?

To deal with the quantitative side of the problem one has to distinguish the raw amount of information needed to describe a structure from the minimal amount of information required to generate it. The latter, the bit length of the shortest algorithm that can generate the structure, is called Kolmogorov complexity [3] and may be smaller by many orders of magnitude than the amount of information required to describe the structure. An extreme example of low Kolmogorov complexity is illustrated in Figure 1. Obviously, nature and nurture need only gigabytes to construct, respectively instruct, the brain. A logical consequence of this efficiency is that the brain is totally dominated by structural regularity, so that instead of from all randomly possible connectivity

[1] 10^{14} synapses, each taking 33 bits to address one of the 10^{10} neurons of the brain.

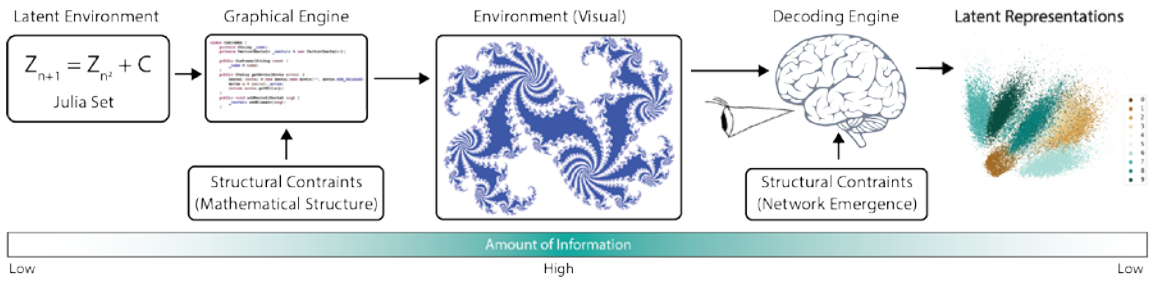


Fig. 1 Illustration of Kolmogorov complexity. Julia sets (middle panel) need, literally, infinite amounts of information to be described, but very little information suffices to generate them (left; recursive definition and mathematical grammar). Perception and efficient learning are possible by reducing the flood of sensory signals produced by the environment to an underlying low-complexity description (right).

patterns among its neurons nature and nurture only need to pick from a vastly smaller space of pre-structured patterns. A central thesis of our communication is that the *structural regularity* implied by this low Kolmogorov complexity *acts as the domain-specific inductive bias* that any system needs [4, 5] or [6, ch. 2.7] to be able to learn efficiently.

The remainder of this paper is organized as follows: In Section 2 we put forward the hypothesis that the Kolmogorov algorithm of the brain is network self-organization as studied extensively on the example of the ontogenetic development of retino-topic connections. In Section 3 we discuss a small number of cognitive sample processes that are in need to be understood and implemented. In Section 4 we try to make plausible how net fragments can serve as basis to solve these problems and in Section 5 we discuss the relevance of the perspective we are creating to open problems within the current field of AI.

2 Network Self-Organization as Kolmogorov Algorithm of the Brain

What is the type of mechanism, the concise Kolmogorov algorithm, by which the connectivity of the brain and hence the structural regularity is generated under genetic guidance? We suggest to adopt as paradigm the experimentally and theoretically well-studied mechanism of the ontogenesis of retinotopic connections: The axons growing out from the retinae of vertebrates reach their target structures (e.g., the optic tectum) in more or less random order, but after a relatively brief period they order themselves so as to establish a smooth mapping conserving geometry [7]. Of all the mechanisms that have been proposed to explain the process only one survived comparison to experiment, *network self-organization* [8, 9]. Its general idea is quite simple. An initial connectivity supports spontaneous activity. This activity acts back by synaptic plasticity to alter the network, and this loop, from connectivity to activity and back to connectivity, continues until a stationary state, an *attractor network*, is reached.

Therefore we propose that network self-organization, as displayed in the retino-tectal system, is the Kolmogorov algorithm generating the wiring of the brain. Sensory signals, as soon as they become available, participate in the mechanism, co-determining the attractor networks that are allowed to form. Attractor networks can be characterized by optimizing two properties: *sparsity* and *consistency*. A network is sparse if it has a small number of connections converging on or diverging from any neuron and connectivity is consistent if it supports high-order temporal correlations between sets of signals arriving at any given neuron. This consistency means that a network is dominated by sets of alternative signal pathways (of approximately equal conduction delay) between many pairs of source and target neurons [10].

As result of such network self-organization, the brain develops as an overlay of attractor networks ('*net fragments*') [11]. Each net fragment comprises a set of neurons and the connections among them. If a set of neurons is activated again and again for a sufficient total time its internal connectivity can converge towards an attractor state. There is positive feedback between the activity of the set and the structure of its connectivity. As large sets of neurons are very unlikely to occur more than once, only small sets will be given a chance to establish themselves as net fragments. Each neuron can be part of several net fragments.

Many systems of low Kolmogorov complexity and implied high regularity arise by emergence. Such systems are composed of building elements that interact by physical, chemical, mechanical etc. forces. Well-known examples are soap bubbles or crystals: Under appropriate conditions (e.g., low temperature in a liquid) large-scale ordered configurations arise in which the forces between the elements interlock such as to lend the configuration stability. In these, weak interactive forces between the building elements (e.g., molecules) can achieve large-scale stability only by interlocking in consistent configurations. In the brain, where quite a number of connections have to conspire (i.e., fire simultaneously) to activate a neuron, a vanishingly small subset of all possible connectivity patterns is singled out by their ability to dynamically self-stabilize as attractors of network self-organization.

After sufficient self-organization of the system larger sets of neurons can only be active as interlocking net fragments, each of which can only become active in the context of overlapping other fragments. This favors the activation of large *coherent nets*, that is, networks which, if given sufficient time, would be attractors under network self-organization. The term 'net' emphasizes composition of smaller fragments, although a net can itself be a fragment of larger nets.

In order not to be caught in local optima, network self-organization needs to start from an initial state that already establishes a coarse global structure from which it can proceed in a coarse-to-fine manner (for which a gradual tightening of inhibitory strength over the course of development [12, 13] may be the basis). This initial connectivity structure, set up by earlier ontogenetic processes which rely on genetically controlled emergence [14] establishes gross

connectivity between sensor organs, effector organs and the behavioral control circuits enabling animals to already function at the time of birth.

In the next sections we will give a sample of typical cognitive processes that are to be implemented and understood (Section 3), will explain how net fragments can serve to do so (Section 4) and how this framework supports efficient learning, generalization and autonomy (Section 5).

3 Cognitive Processes to be Implemented

What essential functions are at the basis of natural intelligence? A lioness stalking pray in the savanna has to integrate a complex array of factors into one coherent strategy in order to be successful. One little disturbing factor can throw off the whole situation. It may be that this complexity of natural situations, in distinction to the logical simplicity of classical AI accomplishments, is responsible for Moravec's paradoxon ("it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility" [15, p. 15]).

The organization of behavior within a given scene is based on a representation of that scene in the brain. *Scene representation*, a contested concept [16, 17], does not imply static and complete rendering of detail as in a photographic image but is rather to be seen as an organizational framework putting abstract interpretations of scene lay-out and scene elements in relation to each other and to potential actions and emotional responses. This framework supports quick flashes of attention which materialize detailed reconstructions of narrow sectors of the scene. Scene representations have to be built up by *perception*. Perception is difficult because sensory data are insufficient and ambiguous and contain in only entangled form the different factors (shape, color, material, motion etc.) that make up the scene. Perception is therefore to be seen as an active process that constructs a model of the scene that uniquely explains the sensory signals and their changes under motion.

According to ethologists, animal and human behavior is defined and controlled by a number of drives (such as to satiate hunger or avoid danger), each of which is laid down under genetic guidance in a schematic form [18, 19]. A *behavioral schema* can be activated by a sensory trigger feature, executes a behavioral response, evaluates the outcome and is modified by the experience. The basic behavioral machinery, which serves a function analogous to a computer user acting through the machine's operating system, is the fruit of evolutionary trial and error over many generations, and presumably is laid down in the style of business process models or Petri-nets in terms of relatively few appropriately connected neurons or neural pools. To integrate this basic machinery in a meaningful way into the flow of scene representations is, however, a very complex affair and is the basic goal of *learning*.

Even beyond the organization of behavior, there is a long tradition [20–22] or [23, pp. 147–172] of discussing schemata as basis for understanding

phenomena and define meaning. It therefore seems important to have a clear view how concrete instances can be related to abstract schemata.

Learning takes place inside tasks that are governed by the behavioral drives. The currently active drive decides which elements of the scene are relevant, focuses attention accordingly and curtails the scene representation to its needs. The drive, as originally defined and further developed by experience, can be seen as an abstract scene description that can serve to shape and interpret actual scenes as schema instantiations. This setting, *a behavioral schema-interpreted scene, serves to powerfully constrain the learning process.*

How can these functions be understood and implemented on the basis of net fragments?

4 Net Fragments as Implementation Medium

As we have argued, both our natural environment and our brain have very low Kolmogorov complexity (cf. Figure 1). We take computer graphics and virtual reality as models for the structure of our natural environment, and we take network self-organization, as studied on the example of the ontogenesis of retinotopy, as the mechanism by which the connectivity of the brain arises. We further note that for a system to efficiently learn it needs to have a strong bias towards its domain [4, 5] or [6, ch. 2.7]. As the human brain indeed learns very efficiently we feel encouraged to propose the hypothesis that the connectivity structures that result from network self-organization, together with the neural dynamics that governs their activation in the establishment of scene representations (see below) are the inductive bias, the *a priori* structure (compare [20]), that tunes the brain to the natural environment.

In the remainder of this section we will discuss how net fragments can serve to implement structures and processes, taking vision as sample modality.

4.1 Data Structure of Primary Visual Cortex

Primary visual cortex is populated with a collection of feature detector neurons with an abundance of short-range lateral excitatory connections between them [24]. Sensory signals coming from a point within the retina in response to visual input activate a subset of the feature neurons whose receptive fields cover that point and its immediate environment. Different local textures activate different such sets. Within some months of early experience network self-organization will re-arrange the excitatory connections within each of these sets and with neurons in the neighborhood. There are 100 times more neurons in primary visual cortex compared to the number of axons coming out of the retina [25], opening the way to sparse codes (as in [26]). Visual input first briefly activates an exuberance of neurons, most of which will then be silenced (by, e.g., balanced inhibition [27]) leaving only the small subset of those neurons active that can support each other by lateral connections inside net fragments (for a model of this process see [28]). (Membership in activated fragments is perhaps indicated by bursting activity [29, 30].) As result of early visual experience

texture patches (at the scale of the range of lateral connections) that dominate the statistics of the input will therefore become represented by net fragments.

This developing structure of the primary visual cortex resembles associative memory [31, 32], except that due to the short range of lateral connections it has the two-dimensional topological structure of the visual field and that its stored local states are defined on a statistical basis. The local net fragments can be compared to the codebook vectors of some image compression algorithms [33]. They can be considered as filters that interpret the actual visual input in terms of patterns previously experienced with statistical significance. They suppress redundancy and regularize responses, as is important, for instance, to extract stereo depth [34] or motion. The net fragments that respond to the surface of a coherent object overlap in terms of neurons and connections and thus form a coherent net, covering the object. Net fragments can thus be seen as implementation of the *Gestalt laws* [35], and the coherent nets they form as realization of the 'force fields' that that movement is speaking of. The coherence of a net covering the cortical region occupied by an object can serve as basis for figure-ground discrimination [36].

The example illustrates the power of net fragments as inductive bias. Local texture-representing net fragments as such could be replaced by the higher-level feature neurons of deep learning systems. However, due to neuron-wise overlap net fragments in distinction to those are exclusively activated when merging into a coherent field, a Gestalt. Net fragments and their dynamics thus naturally render the topological structure of the continuous surfaces that dominate our environment and allows them to be handled as a whole, as seen in the next subsection.

4.2 Invariant Object Representation

A concrete object can appear in the visual cortex in an infinitude of versions differing in position, size, orientation and other factors. In all these versions the object image gets represented, as just discussed, by coherent nets composed of local net fragments. To store and later recognize the object when it appears in the retina in transformed version it is necessary to lay down connections that permit to construct, in response to visual input, nets that represent views of the object independent of its position, orientation etc. In the human brain these invariant representations presumably are located in infero-temporal cortex [37]. There is psychophysical evidence [38] that for a large class of structured object types the visual system is able to construct such invariant representations out of shape primitives that are common to such objects. We propose to *see these shape primitives be represented as net fragments* which have the flexibility to adapt to the shape of actual objects in spite of metric deformations, depth rotation and of course position within object-centered coordinates. The identity and relative position of these shape-primitive-representing net fragments can then serve to identify the object type [38] and serve as basis for manipulation.

To enable such invariant responses to the position- etc. variant representation of objects in the primary cortices the proposal has been made [39–41] that there are rapidly switchable connections ('*shifter circuits*') between the primary visual cortices and invariant representations in infero-temporal cortex that can connect nets in those two areas in a structure-preserving way. In both areas the object is represented by a two-dimensional field of neighborhood-connected neurons. A mapping between them is called structure preserving ('homeomorphic') if it is smooth (connecting neighbors in one field to neighbors in the other) and connects only neurons of the same type.

Simple versions of invariant object recognition on the basis of shifter circuits have been demonstrated [41–43]. Shifter circuits are composed of net fragments and can be formed by network self-organization [44]. Active maps that connect variant images with their invariant representation as well as the movements and deformations of those maps constitute valuable information (as argued in the introduction of [41]), so that, for instance, the shape of an object rotating in front of the eyes can be deduced from the deformation of this map. The separation of visual object representation into external coordinates ('where') and internal structure ('what') is an important example of the disentanglement of sensory patterns into the factors they contain.

The example of invariant object representation again illustrates the power of self-organized net fragments as inductive bias. Different views onto the same object or surface are related by homeomorphy, and net fragments are a natural way to form homeomorphic mappings. Such mappings, seen as dynamic entities, can track and model the movements of objects and surfaces in the environment and their relations to the eye. They are an essential element needed to reconstruct and model in the brain the geometry, kinematics and dynamics of the natural environment.

It is tempting to see invariant visual object representation as a special case of the more general problem of representing the relationship between abstract schemata and instances they apply to. Assuming that this relationship has the character of a homeomorphic mapping (preserving types of entities and their relations) it is conceivable that the ensemble of schema, instance and mapping between them comes to be represented by a coherent net composed of previously established fragments, just as in the example of invariant object representation.

4.3 Net Fragments as Data Structure of the Mind

There is a broad consensus of seeing neurons as atoms of meaning [45]. As such, individual neurons may refer to entities on any level of complexity, but in doing so they act merely as labels, while beyond a low level of complexity they cannot render unambiguously the specific structure of what they refer to. To do this requires a *compositional data structure* (as convincingly argued in [46]). The lack of compositionality in artificial neural networks is referred to as the binding problem [40, 47].

We here argue that net fragments are the brain's compositional data structure and its solution to the binding problem. It is illustrated by the visual representation of objects in both the variant and the invariant versions. Individual feature neurons can, in response to visual input, fire stably only in the context of a net fragment they are part of (see Subsection 4.1 or [28]), and this net fragment can do so only when overlapping with other net fragments (as neurons only fire as part of a net fragment they are part of), so that the response to the input actually is that of a net spanning the whole object as currently pictured. This net is a one-time structure rendering the never-repeating way the object appears at any moment. It responds holistically, as result of a collective effect [48], just as the Gestalt psychologists [35] would have it, and it still renders the Gestalt in minute detail. A hierarchy of features of various complexity levels is represented by nested net fragments of different size.

A good composite data structure has to be able to exert effect on the basis of its structure and be productive in the sense of giving rise to analogous structures [46]. Our example of invariant visual object recognition illustrates this condition. The actual recognition takes place by the activation of a net forming a homeomorphic point-to-point mapping between the invariant and the variant representation. This net gets created by the activation of net fragments each of which connects a small region in the plane with the variant representation (primary visual cortex) with a corresponding small region in the invariant representation (infero-temporal cortex). These 'maplets' are activated by homeomorphy between the small regions they connect and they overlap such as to form a coherent global map between variant and invariant representations of the object, as demonstrated in [39, 40]. Consequently it takes just one exposure to a new object type and formation and storage of a model thereof in the invariant domain to recognize that type of object independently of transformation state. This explains the brain's ability [49] to recognize novel objects in altered position and pose after a single brief exposure. The representation of objects is compositional and productive, as requested by [46], in that the composite mappings can serve any object and represent the position, size and orientation of the variant object image, the invariant representation of an object can render a large number of variant versions thereof, and the net fragments in the two domains can be re-used for an infinitude of different objects.

Compositionality applies also to representing cognitive structure in terms of submodalities (in vision, for instance, texture, color, motion, form, size, position etc.). Whereas sensory signals contain submodalities in implicit form, specific submodality patterns can be represented separately within their own specialized cortical regions. Submodalities are basically independent of each other – object form, for instance, abstracting from position, size, surface texture or coloring. Concrete mental objects can be constructed by linking them together with the help of maps of connections as described above, in a process analogous to the way computer graphics creates visual output by mapping different sub-modalities to each other and into the virtual camera.

Mental objects thus constructed are to be seen as larger net fragments composed as mergers of pre-existing net fragments. In a sufficiently pre-trained brain such nets, once selected by input, are stable constructs that are attractors both in terms of the fast dynamics of neural activation and inactivation and the slow dynamics of network self-organization. Like in associative memory [31], active neurons are pushed by a number of simultaneously firing excitatory connections into a high-activity state, while silent neurons are reliably suppressed by converging inhibitory connections. Such network states can be characterized as of high consistency – consistency between different signals arriving on individual neurons and consistency between the set of currently active neurons and and their connectivity. Network self-organization works on a slower time-scale by performing something like a stochastic gradient descent of neural connections with a cost function, at each individual neuron, that favors binary dynamics with either a highly excited or deeply suppressed state.

4.4 Neural Dynamics: How a Trained Brain Perceives

Perception is difficult due to the paucity and ambiguity of sensory signals and because scene representations have to be spontaneously constructed such as to uniquely explain the sensory input. Given the speed with which our brain routinely performs the task, this construction cannot be based on sequential memory search. To this speed we offer the following explanation. The sensory signals in their great ambiguity reach and alert all net fragments that are compatible with them. Among these, some overlap and dynamically support each other while others are mutually inhibitory. Buried in this dynamics is (given, of course, sufficient previous experience) the comprehensive net that represents the scene. Due to its pervasive consistency of all connections this net prevails in the dynamic process, establishes itself and inhibits all incompatible net fragments. The activation of this net is due to a collective process [48] comparable to a phase transition [50] (like magnetization) instead of to sequential search.

5 Relevance to Open Problems

Grave limitations [51–54] of contemporary AI [55] have to do, first, with inability to generalize sufficiently beyond human-provided examples. We trace this inability to the lack, in current systems, of a sufficiently powerful inductive bias for learning. Inductive biases are specific to application domains [4–6]. We accordingly focus on what we call *natural intelligence* which is tuned to solving general problems in our natural environment.

So far, we have argued that our natural environment has low Kolmogorov complexity, interpreting today's virtual reality systems (which have low complexity) as sufficiently convincing approximation to that environment. We have further noted that the brain also is of low Kolmogorov complexity and have subscribed to the view that its connectivity structure arises by emergence realized by network self-organization. We have taken the brain's tremendous power

to learn and generalize from scant examples as indication that emerging connectivity structures (net fragments) are the data structure of the brain and constitute its inductive bias for learning.

As to learning, two stages have to be distinguished: First, a system has to develop the toolbox that is necessary to model the surrounding scene. Second, once it is in a position to model specific arrangements and processes it can learn to relate them in finer and finer detail to its set of behavioral schemata and the corresponding goals. For brains, the first stage is partly reached in pre-natal development under genetic guidance, partly by sensory-motor experimentation by the young individual. In the context of AI, this stage is modeled in the field of developmental robotics [56].

For brains, learning in the second stage is, by comparison to current AI technology, powerfully alleviated by two factors. First, during scene construction in interaction with and under the influence of a currently ruling behavioral schema the schema-relevant scene elements are labeled as such by their mapping to and from the schema. This goes a long way towards credit assignment during the evaluation of the ongoing experience and suppresses irrelevant detail. Second, the essential structure to be picked up from the current situation (object, motion pattern, etc.) is already modeled as part of the scene representation, not only in concrete detail but also on more abstract levels. It is therefore possible to tie together all essential elements of the situation – the relevant scene elements, their relative arrangement, their roles as defined in the behavioral schema – by strengthening or creating a small number of connections to fixate the experience. This fixation has to happen at an appropriately abstract level (the ability to find this level being a subject for an appropriate kind of meta-learning), so that the particular experience generalizes to analogous situations.

For AI systems, however, this generalization ability is still to be realized. The presented methods could therefore, if properly implemented, mitigate the above-mentioned problems of sample efficiency (including slow learning) and generalization in a principled and unified way, with the effect of leading to results that can approach common sense (compare with compartmentalised approaches in [57–59]).

A second set of weaknesses of present AI technology revolves around low level of autonomy. In typical applications rather narrow goals are formulated by humans, application-specific data are collected and human-tuned architectures and hyper-parameter settings are empirically determined [60]. This limits systems to specific applications and causes great expense, which is well illustrated by the enormous time and investment in terms of human effort necessary to develop autonomous vehicles. True autonomy requires a complete (in some sense) set of abstract goals and behavioral schemata together with the ability to (learn to) relate these schemata to concrete situations. The difficulty of this is due to the enormous distance in terms of abstraction between concrete scene elements and the representations of general goals. We suggest that this distance is bridged by homeomorphic relationships, and that these

homeomorphic relationships can be found with the help of composition of net fragments.

The superiority of human intelligence over that of animals is due to a very rich complement of culturally acquired schemata many of which are absorbed in verbal or symbolic form. We are born with a behavioral repertoire that is very similar in principle to that of a range of animal species, but soon new goals are acquired, grafted upon a small set of innate behavior patterns (such as wanting to please or imitate social partners) acting as gateways. It has been argued that higher intellectual abilities grow in the individual as layers of generalization by analogy, starting with the sensory-motor coordination structure acquired early in life [56, 61]. So far it hasn't been possible to model and artificially replicate that process. We suggest that the missing element is a potently pre-conditioned data structure and that network self-organization is providing this pre-conditioning in our brain.

6 Conclusion

A deep riddle of our existence is the question how the ideas and imaginations in our mind arise. Super-natural influences and exotic force fields or quantum processes are widely invoked. According to our proposal mental phenomena appear like mathematical structures, which are singled out by the condition of logical consistency and seem to be there even before being discovered by mathematicians.

Acknowledgements

This work was conducted during the first author's stay as visiting professor at the UZH/ETH Institute of Neuroinformatics and the ZHAW Centre for AI, financed by UZH/ETH. The authors are grateful for the catalytic effect brought about by the Mindfire Foundation and helpful discussions with Rodney Douglas.

References

[1] Consortium, T.I.H.G.M.: A physical map of the human genome. Nature **409**, 934–941 (2001). https://doi.org/10.1038/35057157

[2] Landauer, T.K.: How much do people remember? some estimates of the quantity of learned information in long-term memory. Cognitive Science **10**(4), 477–493 (1986). https://doi.org/10.1016/S0364-0213(86)80014-3

[3] Kolmogorov, A.: On tables of random numbers. Theoretical Computer Science **207 (2)**, 387–395 (1998). https://doi.org/10.1016/S0304-3975(98)00075-9

[4] Geman, S., Bienenstock, E., Doursat, R.: Neural networks and the bias/variance dilemma. Neural Computation **4**, 1–58 (1992)

[5] Wolpert, D.H.: The lack of a priori distinctions between learning algorithms. Neural Computation **8**, 1341–1390 (1996)

[6] Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)

[7] Goodhill, G.J.: Contributions of theoetical modeling to the understanding of neural map development. Neuron **56**, 301–311 (2007)

[8] Willshaw, D.J., von der Malsburg, C.: How patterned neural connections can be set up by self-organization. Proceedings of the Royal Society of London, Series B **194**, 431–445 (1976)

[9] Willshaw, D.J., von der Malsburg, C.: A marker induction mechanism for the establishment of ordered neural mappings: its application to the retinotectal problem. Phil. Trans. R. Soc. Lond. B **287**, 203–243 (1979)

[10] von der Malsburg, C., Bienenstock, E.: A neural network for the retrieval of superimposed connection patterns. Europhysics Letters **3**, 1243–1249 (1987)

[11] von der Malsburg, C.: Concerning the neural code. J. Cog. Sci. **19 (4)**, 511–550 (2018). https://doi.org/10.17791/jcs.2018.19.4.511

[12] Li, Y.-t., Ma, W.-p., Pan, C.-j., Zhang, L.I., Tao, H.W.: Broadening of cortical inhibition mediates developmental sharpening of orientation selectivity. Journal of Neuroscience **32**(12), 3981–3991 (2012). https://doi.org/10.1523/JNEUROSCI.5514-11.2012

[13] Lim, L., Mi, D., Llorca, A., Marı´n, O.: Development and functional diversification of cortical interneurons. Neuron **100**, 294–313 (2018)

[14] Waddington, C.H.: The Strategy of the Genes. Ruskin House, London, Great Britain (1957)

[15] Moravec, H.: Mind Children. Harvard University Press, Cambridge, Massachusetts (1988)

[16] Freeman, W.J., Skarda, C.A.: Representations: Who needs them? In: JL, M., Weinberger, N., Lynch, G. (eds.) Third Conference, Brain Organization and Memory: Cells, Systems and Circuits. Guilford Press, New York, Oxford (1990)

[17] O'Regan, J.K., Noë, A.: A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences **24**(5), 939–973 (2001). https://doi.org/10.1017/S0140525X01000115

[18] Shettleworth, S.: Cognition, Evolution, and Behavior (2nd Ed.). Oxford University Press, Oxford (2010)

[19] Kilmer, W.L., McCulloch, W.S., Blum, J.: A model of the vertebrate central command system. International Journal of Man-Machine Studies **1**, 279–309 (1969)

[20] Kant, I.: Critique of Pure Reason. Cambridge University Press, Cambridge, England (Original work published in 1781) (1781/1999)

[21] Piaget, J.: Langage et Pensée Chez L'enfant, p. 43. Delachaux et Niestlé, Neuchâtel (1923)

[22] Bartlett, F.C.: Remembering: A Study in Experimental and Social Psychology. Cambridge University Press, Cambridge, England (1932)

[23] Johnson, M.: The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason. University of Chicago Press, Chicago (1987)

[24] Kandel, E., Schwartz, J., Jessell, T., Siegelbaum, S., Hudspeth, A.: Principles of Neural Science, 5th Ed. McGraw-Hill, New York (2012)

[25] Leuba, G., Kraftsik, R.: Changes in volume, surface estimate, three-dimensional shape and total number of neurons of the human primary visual cortex from midgestation until old age. Anat Embryol **190**, 351–366 (1994). https://doi.org/10.1007/BF00187293

[26] Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive fields properties by learning a sparse code for natural images. Nature **381**, 607–609 (1996)

[27] Vogels, T., Sprekeler, H., Zenke, F., Clopath, C., Gerstner, W.: Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. Science **334**, 1569–73 (2011)

[28] Wansch, J.: An associative network based on balanced inhibition. Master's thesis, Goethe University, Frankfurt (January 2020)

[29] Payeur, A., Guerguiev, J., Zenke, F., Richards, B., Naud, R.: Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. Nat Neurosci **24(7)**, 1010–1019 (2021). https://doi.org/10.1038/s41593-021-00857-x

[30] Naud, R., Sprekeler, H.: Sparse bursts optimize information transmission in a multiplexed neural code. Proceedings of the National Academy of Sciences **115**(27), 6329–6338 (2018). https://doi.org/10.1073/pnas.1720995115

[31] Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences **79**(8), 2554–2558 (1982). https://doi.org/10.1073/pnas.79.8.2554

[32] Krotov, D., Hopfield, J.J.: Dense associative memory is robust to adversarial inputs. Neural Computation (12), 3151–3167 (2018)

[33] Taubman, D., Marcellin, M.W.: Jpeg-2000 Image Compression: Fundamentals, Standards and Practice. Kluwer Academic Publishers, Dordrecht (2002)

[34] Marr, D., Poggio, T.: Cooperative computation of stereo disparity. Science **194**, 283–287 (1976)

[35] Ellis, W.E. (ed.): A Source Book of Gestalt Psychology. Routledge & Kegan Paul, London (1950)

[36] Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22(8)**, 888–905 (2000)

[37] Rolls, E.T.: Learning invariant object and spatial view representations in the brain using slow unsupervised learning. Frontiers in Computational Neuroscience **15** (2021). https://doi.org/10.3389/fncom.2021.686239

[38] Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychol Rev. **94**, 115–147 (1987)

[39] Anderson, C.H., van Essen, D.C.: Shifter circuits: A computational strategy for dynamic aspects of visual processing. PNAS **84**, 6297–6301 (1987)

[40] von der Malsburg, C.: The correlation theory of brain function. Internal report, 81-2, Max-Planck-Institut für Biophysikalische Chemie, Postfach 2841, 3400 Göttingen, FRG (1981/1994). Reprinted in E. Domany, J.L. van Hemmen, and K.Schulten, editors, *Models of Neural Networks II*, chapter 2, pages 95–119. Springer, Berlin, 1994.

[41] Arathorn, D.W.: Map-Seeking Circuits in Visual Cognition – A Computational Mechanism for Biological and Machine Vision. Standford Univ. Press, Stanford, California (2002)

[42] Olshausen, B., CH, A., Van Essen, D.: A multiscale dynamic routing circuit for forming size- and position-invariant object representations. Journal of Computational Neuroscience **2**, 45–62 (1995)

[43] Wolfrum, P., Wolff, C., Lücke, J., von der Malsburg, C.: A recurrent

dynamic model for correspondence-based face recognition. Journal of Vision **8**, 1–18 (2008). doi:10.1167/8.7.34

[44] Fernandes, T., von der Malsburg, C.: Self-organization of control circuits for invariant fiber projections. Neural Computation **27**, 1005–1032 (2015). https://doi.org/10.1162/NECO_a_00725

[45] Quiroga, R., Reddy, L., Kreiman, G., Koch, C., Fried, I.: Invariant visual representation by single neurons in the human brain. Nature **435**, 1102–1107 (2005)

[46] Fodor, J.A., Pylyshyn, Z.W.: Connectionism and cognitive architecture: A critical analysis. Cognition **28**(1), 3–71 (1988). https://doi.org/10.1016/0010-0277(88)90031-5

[47] Roskies, A.L.: Introduction: The binding problem. Neuron **24**, 7–9 (1999)

[48] Fano, U.: A common mechanism of collective phenomena. Rev. Mod. Phys. **64**, 313–319 (1992). https://doi.org/10.1103/RevModPhys.64.313

[49] Biederman, I., Bar, M.: One-shot viewpoint invariance in matching novel objects. Vision Research **39**, 2885–2899 (1999)

[50] Stanley, H.: Introduction to Phase Transitions and Critical Phenomena. Clarendon Press, Oxford (1971)

[51] Marcus, G.: Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631 (2018)

[52] Shrestha, A., Mahmood, A.: Review of deep learning algorithms and architectures. IEEE Access **7**, 53040–53065 (2019). https://doi.org/10.1109/ACCESS.2019.2912200

[53] Zador, A.M.: A critique of pure learning and what artificial neural networks can learn from animal brains. Nature communications **10**(1), 1–7 (2019)

[54] Tuggener, L., Schmidhuber, J., Stadelmann, T.: Imagenet as a representative basis for deriving generally effective cnn architectures. arXiv preprint arXiv:2103.09108 (2021)

[55] Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks **61**, 85–117 (2015)

[56] Lee, M.: How to Grow a Robot: Developing Human-Friendly, Social AI. MIT Press, Boston, MA (2020)

[57] Botvinick, M., Ritter, S., Wang, J.X., Kurth-Nelson, Z., Blundell, C.,

Hassabis, D.: Reinforcement learning, fast and slow. Trends in Cognitive Sciences **23**(5), 408–422 (2019). https://doi.org/10.1016/j.tics.2019.02.006

[58] Sejnowski, T.J.: The unreasonable effectiveness of deep learning in artificial intelligence. Proceedings of the National Academy of Sciences **117**(48), 30033–30038 (2020). https://doi.org/10.1073/pnas.1907373117

[59] Zellers, R., Holtzman, A., Peters, M., Mottaghi, R., Kembhavi, A., Farhadi, A., Choi, Y.: PIGLeT: Language grounding through neuro-symbolic interaction in a 3D world. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 2040–2050. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.159

[60] Stadelmann, T., Amirian, M., Arabaci, I., Arnold, M., Duivesteijn, G.F., Elezi, I., Geiger, M., Lörwald, S., Meier, B.B., Rombach, K., Tuggener, L.: Deep learning in the wild. In: IAPR Workshop on Artificial Neural Networks in Pattern Recognition, pp. 17–38 (2018). Springer

[61] Lakoff, G., Nunez, R.E.: Where Mathematics Come From. How The Embodied Mind Brings Mathematics Into Being. Basic Books, New York, NY (2000)

ORIGINAL PAPER

Towards a Governance Framework for Brain Data

Marcello Ienca · Joseph J. Fins · Ralf J. Jox · Fabrice Jotterand · Silja Voeneky · Roberto Andorno · Tonio Ball · Claude Castelluccia · Ricardo Chavarriaga · Hervé Chneiweiss · Agata Ferretti · Orsolya Friedrich · Samia Hurst · Grischa Merkel · Fruzsina Molnár-Gábor · Jean-Marc Rickli · James Scheibner · Effy Vayena · Rafael Yuste · Philipp Kellmeyer

Received: 17 September 2021 / Accepted: 23 May 2022

Abstract The increasing availability of brain data within and outside the biomedical field, combined with the application of artificial intelligence (AI) to brain data analysis, poses a challenge for ethics and governance. We identify distinctive ethical implications of brain data acquisition and processing, and outline a multi-level governance framework. This framework is aimed at maximizing the benefits of facilitated brain data collection and further processing for science and medicine whilst minimizing risks and preventing harmful use. The framework consists of four primary areas of regulatory intervention: binding regulation, ethics and soft law, responsible innovation, and human rights.

Keywords Brain data · Data governance · Neurodata · Neurotechnology · Regulation

Introduction

Human brain data are becoming a sought-after commodity in an increasing number of contexts and activities. Until a few years ago their acquisition and

M. Ienca (✉)
College of Humanities, EPFL, Lausanne, Switzerland
e-mail: marcello.ienca@epfl.ch

M. Ienca · A. Ferretti · E. Vayena
Department of Health Sciences and Technologies, ETH Zurich, Zurich, Switzerland

J. J. Fins
New York Presbyterian Hospital and Weill Cornell Medical College, New York, NY, USA

R. J. Jox
Institute of Humanities in Medicine, Lausanne University Hospital, Lausanne, Switzerland

F. Jotterand
Medical College of Wisconsin, Milwaukee, WI, USA

F. Jotterand
Institute of Biomedical Ethics, University of Basel, Basel, Switzerland

S. Voeneky · R. Chavarriaga
Department of International Law and Ethics of Law, Law Faculty, University of Freiburg, Freiburg, Germany

R. Andorno
Faculty of Law and Institute for Biomedical Ethics, University of Zurich, Zurich, Switzerland

T. Ball
Neuromedical AI Lab, Department of Neurosurgery, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg im Breisgau, Germany

C. Castelluccia
Privatics Group, INRIA, Paris, France

R. Chavarriaga
ZHAW School of Engineering, IEEE Standards Association; Confederation of Laboratories for AI Research in Europe (CLAIRE), Winterthur, Switzerland

Published online: 03 June 2022

analysis were limited to the clinical field and biomedical, psychological or behavioral research. Today, brain data are also increasingly being used in employment, education, and military contexts, as well as for personal use through an increasing number of consumer-grade neurotechnological devices.

In the consumer space, information technology companies are developing devices and applications that leverage brain data for consumer purposes such as cognitive monitoring, neurofeedback, device control or other forms of brain-computer interfacing. For example, between 2017 and 2021 Facebook worked on a brain-computer interface (BCI) research program aimed at building a wearable BCI that enables users to type by simply imagining speech. Microsoft is working in parallel on non-invasive interactive BCIs for the general population while a whole ecosystem of neurotechnology companies such as *Neuralink, Emotiv* and *Kernel* is rapidly emerging. Consumer neurotechnology, e-learning, digital phenotyping, affective computing, psychographics and neuromarketing are some of the domains of application that leverage brain data as a commodity [1, 2].

In the educational and work setting, attempts have been made to collect and process brain data for purposes such as improving learning and redesigning workflows. For example, last year, in China, primary school children were enrolled in a trial where electroencephalography (EEG) data were recorded during cognitive tasks to assess their attention spans [3]. Also in China, government-backed workplace surveillance projects are deploying personal neurotechnologies to detect changes in brain activity among factory employees on the production line. These neurotechnologies are intended to monitor productivity and adjust the pace of production accordingly [4].

Finally, military uses of neurotechnologies and the associated acquisition of brain data have increased in quantity and variety. One example is the "Next-generation Nonsurgical Neurotechnology Program" (N^3), a $104 million effort launched in 2019 by the United States Defense Advanced Research Projects Agency (DARPA) with the aim of developing non-invasive, portable and bidirectional BCIs for service members [5]. Several other nations have military research programs that involve brain data [6].

These novel uses of brain data add to the already extensive use of these data in clinical medicine and biomedical research. In these fields, electrophysiology and neuroimaging datasets have steadily grown in volume, variety and analytic complexity [7, 8]. Data repurposing, a frequent occurrence in digital health and digital phenotyping, also permits cross-domain data transfer, blurring the lines between biomedical and non-medical data uses.

An Ethical and Policy Challenge

The increasing availability of brain data inside and outside the biomedical and health-care domain raises challenges for regulation and governance. On the one

H. Chneiweiss
Centre de Recherche Neuroscience Paris Seine, CNRS; UNESCO Chair of Bioethics, Paris, France

O. Friedrich
Institute for Philosophy, FernUniversität in Hagen, Hagen, Germany

S. Hurst
Institute for Ethics, History, and the Humanities, University of Geneva, Geneva, Switzerland

G. Merkel
Department of Police Sciences, University of Applied Sciences for Administration and Services Kiel-Altenholz, Altenholz, Germany

F. Molnár-Gábor
Heidelberg Academy of Sciences and Humanities, Heidelberg, Germany

J.-M. Rickli
Geneva Center for Security Policy, Geneva, Switzerland

J. Scheibner
College of Business, Government and Law, Flinders University, Bedford Park, Australia

R. Yuste (✉)
The NeuroTechnology Center, Columbia University, New York, NY, USA
e-mail: rmy5@columbia.edu

P. Kellmeyer (✉)
Human-Technology Interaction Lab, Department of Neurosurgery, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg im Breisgau, Germany
e-mail: philipp.kellmeyer@uniklinik-freiburg.de

hand, expanding the volume and variety of brain data available for research is crucial for advancing our scientific understanding of the human brain and providing preventive, diagnostic and therapeutic solutions for patients with neurological or psychiatric disorders [9, 10]. Several large-scale research programs, such as the US BRAIN Initiative and the EU Human Brain Project, are working on advancing measurement tools and computational methods in neuroscience and neurotechnology. These projects could benefit from increased data availability in the medical or consumer domain.

On the other hand, as brain data become part of a wider digital ecosystem, they are subject to the same risks and vulnerabilities as other digital data. These include re-identification, hacking, unauthorized reuse, asymmetric commodification, privacy-sensitive data mining, digital surveillance and co-opting data for other non-benign purposes [2, 11]. Most importantly, brain-related measurements in the non-medical domain are rarely available in isolation. They can be combined with other digitally available information and contextualized against online queries, social media, self-tracked data, DNA and geolocation. Advances in big data analytics and machine learning (ML) portend an unprecedented capacity to infer and identify patterns and predict outcomes by aggregating data from multiple sources [12–16].

Given the increased availability of brain data and recent emphasis in national and international policy-making on data governance, the following question arises: how should brain data be regulated? In particular: what kind of governance framework is needed to maximize the benefits of brain data processing for scientific research and medicine whilst ensuring ethical use in other areas?

What Makes Brain Data Important?

The notion of "brain data" is often used without a clear conceptual characterization. To promote clarity for regulation and governance purposes, we propose the following working definition: *Human brain data* are *quantitative data about human brain structure, activity and function.* These include direct measurements of brain structure, activity and/or function (e.g., neuronal firing or summed bioelectric signals from EEG) and indirect functional indicators (i.e., blood flow in fMRI and fNIRS). These types of brain data can be combined with non-neural contextual data, such as voice recordings, smartphone usage data or neuropsychological assessments, that can be used to support inferences about mental processes in a broader sense (Fig. 1). Compared to other measurements of the human body, the risks associated with the collection and processing of brain data are distinctive in terms of quality and magnitude. This is due to inherent properties of brain data and their resulting ethical and legal implications.

At the neurobiological level, brain data are the most direct correlates of mental states, as all cognitive and emotional activity is generated by the brain. Although current neurotechnologies, especially non-invasive techniques, are not yet able to *decode thoughts* —in the sense of providing a full, granular and real-time account of the neural patterns of specific cognitive processes—they increasingly allow to infer the engagement of perceptual and cognitive processes from patterns of brain activation, a process known as *reverse inference* [17]. This occurs through invasive and non-invasive methods to record (and manipulate) neuronal circuits as well as AI and ML-driven data analytics. In laboratory animals, it is now possible to decode visual perception and manipulate it with high precision [18, 19]. In studies with human subjects, researchers have used fMRI scans and high-density electrocorticography signals to accurately decode mental imagery and silent speech [20, 21]. Recent work on intracranial EEG recordings of speech-related brain activity has achieved remarkable accuracy in identifying brain activity patterns related to inner speech [22] while ML techniques have helped enhance the analysis of cognitive processes also from EEG measurements [23, 24].

Finally, research has shown that predictive inferences about mental states can be drawn also from non-neural data sources such as behavioural and digital phenotyping data [25]. Since network neuroscience models and ML techniques are increasingly acquiring inferential power, brain data analytics will likely result, in the long term, in a greater disclosure of mental information. Big data approaches combining brain data and contextual data may offer additional inferential resources for such predictive analytics and allow for more far-reaching and personalized inferences, especially regarding mental content. Mental decoding can improve our scientific understanding of mental illness and holds promise for the targeted

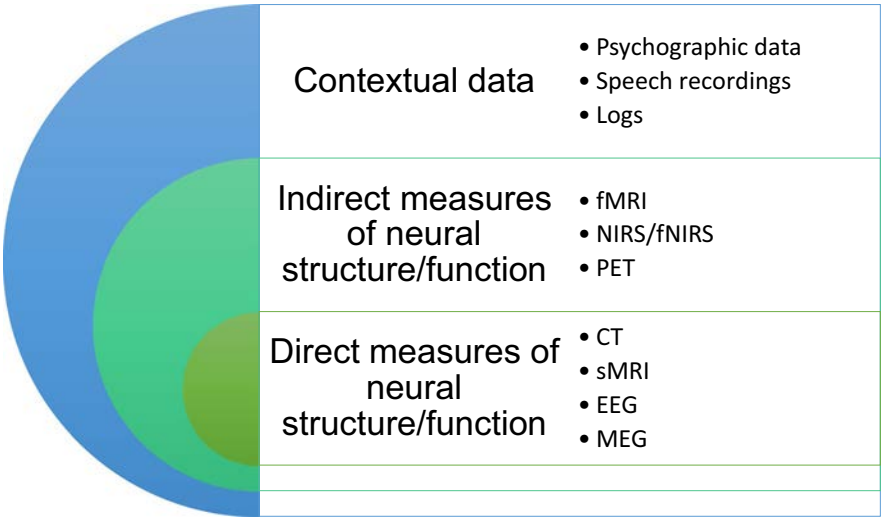


Fig. 1 Brain data taxonomy. CT, computed tomography; MEG, magnetoelectroencephalography; EEG, electroencephalography; PET, positron emission tomography; (f)NIRS, (functional)near-infrared spectroscopy; (f)MRI, functional magnetic resonance imaging, (s)MRI structural MRI. The first category consists of methods for directly measuring electrical activity associated with neuronal activity. The second consists of methods for indirectly measuring neuronal activity, which operate under the principle that neural activity is supported by increased local blood flow and metabolic activity. The third class consists of active or passive digital phenotyping data related to perception, cognition, emotion and behavior. The data types presented in this taxonomy should be considered as explicative of each data category, not as an exhaustive typology

modulation of mental states. At the same time, it raises privacy and security challenges.

Even without decoding mental information, current inferential models based on brain data can make privacy-sensitive inferences about present and future brain function or health status. These inferences and predictions, including early signatures of cognitive decline, can be made about both individuals and groups [26]. Since brain data can be stored digitally, more information will become inferable in the future, as scientific understanding of brain processes and decoding algorithms improve. Furthermore, brain data have higher temporal resolution and potential for real-time interaction compared to other biomedical data such as genetic data. This enables more time-sensitive access to brain activity, e.g., for real-time brain-computer interfacing. Finally, brain data are not "read-only" but are often available in a "read-and-write" format due to neuromodulation such as via electromagnetic brain stimulation techniques, optochemistry and optogenetics. This opens the prospect of targeted and direct influence on a person's mental life and personal identity.

It should be highlighted that many neurotechnologies currently available in the consumer space have limited precision [27]. However, with the current pace of technological progress, increasing market growth and the frequent spillover of biomedical technology into the non-medical sector, brain data processing for non-medical purposes raises the need for anticipatory ethics and foresight governance.

Ethical and Legal Challenges of Brain Data

These unique properties of brain data raise substantive ethical and legal challenges. Since the human brain governs not only life-maintaining physiological processes but also cognitive, affective, volitional, and social faculties [28–30], brain data raise challenges for fundamental normative and legal constructs such as personal identity, autonomy, freedom of thought, moral agency, mental privacy and mental integrity. The notion of "freedom of thought", for example, has been historically characterized as the right and freedom to protect the externalizations of thought such as choice (*freedom of choice*), language (*freedom of*

speech) and behavior (e.g. *freedom of expression*). Brain data processing may solicit a literal reinterpretation of the right to freedom of thought. Similarly, the notions of personhood and personal identity are highly dependent on individual brain function and directly affected by changes to brain activity via neuromodulation.

Further, brain data processing raises novel challenges for the notion of mental privacy for two reasons. First, privacy is predicated upon the conscious ability of the individual to filter the flow of data and intentionally seclude private information. Brain data, in contrast, are mostly elusive to conscious control, hence cannot always be intentionally secluded. While this problem is shared with other data types (e.g., genetic data), it acquires greater ethical complexity in the neural domain. Specifically, brain data admit no separation between the data processed and the system that makes decisions about their processing (the human brain). Second, brain information is the ultimate resort of informational privacy since it includes unexecuted behavior, inner speech or other non-externalized action. In principle, mental privacy can be preserved even if individual behavior is constantly surveilled through activity tracking, personal digital technology, self-quantification or simple observation. It could be argued that when one agrees to allow brain data to be acquired, one seems to surrender the right to mental privacy, at least to some degree. However, in scenarios where brain data collection is either mandated (e.g. in the military sector or workplace) or competitively advantageous (e.g. Facebook's plan to make brain-typing faster than the touch-screen), the risk of sharing data under explicit or implicit coercion is concrete.

AI-driven brain data processing may allow access to mental information and bring privacy debates into partially uncharted territory. Legal systems are well-equipped to protect the '*locus externus*' (behavior, verbal utterances, written text etc.) but less-equipped to protect the '*locus internus*' (e.g. unspoken information, preconscious preferences, attitudes, and beliefs). Data subjects may lose control over their brain data in several ways: (i) by consenting to the collection of their data without being adequately informed (e.g. on a device's Terms of Use due to the complexity of the subject matter); (ii) by providing informed consent to the processing of their data for a certain purpose but remaining unaware of further reuses of their data for different purposes (including scraping by third parties); (iii) by being coerced to have their data collected (e.g. via employer's mandate or in an interrogation context).

The nature of brain data might also compromise the ability of data subjects to exercise their rights to access, edit and delete their own data. For example, a data subject might not possess a computer powerful enough to process data from a BCI [31]. Likewise, deleting brain data may substantially decrease the accuracy of ML models generated with these data. Finally, brain data processing generates a risk of "neurodiscrimination", i.e., discrimination based on a person's neural signatures (indicating, for example, a dementia predisposition), or mental health, personality traits, cognitive performance, intentions and emotional states.

Gaps in the Current Ethical, Legal Framework

We identify four intimately interconnected areas that require attention and proactive governance to ensure the safe and responsible use of brain data outside of the biomedical domain:

Gaps in Supranational and International Law No mandatory governance framework focused on brain data currently exists in supranational or international law. *Prima facie*, brain data are personal data, as codified inter alia in the legally binding European Union's *General Data Protection Regulation* (GDPR), the non-binding 2013 OECD's *Privacy Guidelines* and the upcoming Council of Europe's (CoE) *Modernized Convention for the Protection of Individuals with Regard to the Processing of Personal Data*, and the *European Convention on Human Rights* (ECHR), particularly Article 8. Under these instruments, personal data are defined as any information related to an identified or identifiable natural person (Art. 4 GDPR; Art. 1 OECD Privacy Guidelines, Art. 2a CoE).

The right to privacy, enshrined in Article 8 ECHR, includes the right to data protection. Art. 8 ECHR protects sensitive information, which includes personal data revealing, for example, political opinions, information about a person's health, racial origin, or sexual orientation. With respect to genetic and

biometric data (e.g., cell samples, voice samples), the ECtHR found that, due to rapid technological developments, it is not possible to anticipate and understand all the adverse effects that the collection of such data may entail with respect to private life, and that therefore the collection of any genetic or biometric data constitutes per se an interference with Art. 8 ECHR. The ECtHR might follow a similar approach with regard to brain data.

However, there are a number of limits with this definition of brain data as personal data as defined by GDPR. Firstly, the GDPR is not applicable if brain data are anonymized even though the technical difficulty of anonymizing brain data leaves open the potential for re-identification. Research shows the feasibility of re-identifying data subjects based on electrophysiological measurements or neuroimaging data, predicting present emotional states and future behavior from brain data, as well as decoding information either from the neural activity of data subjects or their digital phenotypes [24, 32]. Because of the technology involved in the processing of brain data and its high contextualization, the likeliness that anonymized brain data (or data *thought to be anonymized*) will become re-identifiable is non-negligible.

Secondly, unique characteristics of brain data pose challenges to safeguarding the rights of data subjects. A prominent example is the *right to be forgotten*, i.e., one's right to request a data controller to delete his/her personal data. A key characteristic of brain data is that they are potentially re-identifiable and elude conscious control. Therefore, even if a person is initially able to have their data deleted, the data controller or others might use those data to derivatively reconnect them to the person concerned. Most importantly, in the case of brain data involving 'unconscious' information, the data controller might be able to retain data the individual is not aware of. Finally, data deletion by consumer BCI companies may be difficult to obtain due the impact that such erasure would have on the accuracy of predictive models [31].

Thirdly, the GDPR allows derogations to the rights of data subjects if data (including the special categories of data listed in Article 9 (1) GDPR) is processed for research or statistical purposes. Those research exemptions also apply to research conducted by private companies, as pointed out by Recital 159 to the GDPR, which names "privately funded research" as part of the science privileged by the GDPR. This implies that processing of brain data by both public and private actors (e.g., government agencies or consumer neurotechnology companies), may rely on derogations from the main GDPR rules. Nevertheless, it is unclear under which conditions the research exemption for the purpose limitation principle defined in Article 5 (1) (b) GDPR applies to brain data collected in the consumer context.

Further, brain data may undermine another principle of data protection law, namely *purpose limitation*. By default, any personal data (including health data) can only be collected for specific purposes that need to be specified at the time when consent is given by the data subject or other legal basis is drawn on, that means ahead of starting data collection and processing. However, the exact specification of purposes is very difficult for brain data because current technology cannot pre-emptively discern purpose-specific data from the myriads of brain signals. Tools for selective filtering such as the *Brain-Computer Interface Anonymizer* are in early stages of development [33]. The GDPR allows framing purposes in a broader manner in specific cases. Nevertheless, data security measures that intend to balance risks for the rights and interests of the data subject and the interests in the data processing are difficult to define in case processing purposes are framed in a broader manner, such as based on broad consent for scientific research (recital 33 GDPR) or based on the processing for scientific research purposes, Art. 9(2)(j) GDPR in conjunction with Art. 89(1) GDPR. Last but not least, the GDPR introduces the fiction that secondary processing for scientific research purposes is compatible with the initial purpose (Art. 5(1)(b) GDPR). Arguably, commercial scientific research, as any other research, underlies transparency obligations that are higher if for-profit benefits are gained based on research conducted with the data.

Finally, safeguards provided by data protection law may not adequately scale to group-level data. This lack of adequate scaling raises a twofold group-privacy risk: first, third parties can make inferences about a group of data subjects based on one or multiple features inherent in the brain data and shared by all individuals in the group (e.g., slower reaction time to cognitive tests). Second, individuals could be unwittingly identified through their brain data, however anonymized, as part of a hitherto unsuspected group (e.g., people showing prodromal signatures of

cognitive decline) and subsequently discriminated against.

To complicate things, brain data generated from consumer neurotechnologies may not constitute 'health data' hence are subject to lower protections compared to data from clinical applications because the application of these devices does not fall under medical device regulation regimes [34].

Gaps in Ethics and Soft Law The collection and processing of brain data within biomedical or clinical research is further governed by research ethics guidelines for the protection of human subjects. These include the *Belmont Report* and the *Declaration of Helsinki* by the World Medical Association, as well as through oversight mechanisms such as Institutional Review Boards. These instruments are critical to uphold the rights and responsibilities of the research community in the conduct of biomedical and clinical research. However, they do not apply in the consumer, neuromarketing, workplace or military domains. In the consumer space, simply prompting users to accept a service's Terms of Use places the responsibility on users to understand these terms and does not guarantee informed decision-making [35]. Even if consent can be obtained in a broad manner, current ethical safeguards are ill-suited to guide the entire data lifecycle. This is particularly true given the trend towards perpetual recycling and re-contextualization of previously collected data [36]. Further, ML allows to draw post-hoc private and confidential inferences from non-sensitive data, prompting further need for the protection of data subjects [12]. Based on these considerations, experts have called for ethical guidelines for novel consumer neurotechnologies to fill persisting gaps in data governance [37].

Gaps in Responsible Innovation Currently, most applications that collect and process brain data outside the clinical and medical research context do not seek compliance with the EU Medical Device Regulation (2017/745) or approval from the US Food & Drug Administration (FDA). Approval from these agencies is only necessary for software and devices with a medical purpose. This bypassing of the relevant medical device regulation is generally predicated on the non-medical scope of these devices and programs. However, a further challenge arises: even though brain stimulation products are covered under Annex XVI, No. 6, the Regulation does not cover brain data processing for purposes other than neuromodulation. We call for expanding the purview of this regulation as to include devices with which users (including vulnerable individuals and groups) may share their brain data for non-medical yet health-related purposes, such as cognitive monitoring and mental wellbeing. Such devices are currently not classified as medical devices and are regularly marketed for wellness, relaxation and other non-medical purposes. They also do not fall under the scope of application based on Annex XVI of the MDR as they often do not include brain stimulation. Furthermore, providing increased guidance for users through clear labelling of such products as not suitable for health-related and medical purposes could enhance transparency and contribute to the fulfilment of information obligations. Finally, consumer and military neurotechnologies can collect medically relevant parameters (e.g. via EEG measurements) and often claim to draw inferences about cognition or psychological wellbeing. Many wearable devices and applications are available for commercial, personal and even health-related use without relevant labelling required by data quality standards [27]. Typically, users of consumer neurotechnology devices or services have no information about how in-house brain function databases are compiled. Further, users have no guarantee that such databases are sufficiently representative to provide valid assessments of individual or group-level cognitive function and affective state [38]. Insufficiently validated applications may incorporate bias, provide false information or even cause harm to the users such as when users make health-related decisions based on these apps. Additional hazard may be posed by malicious hacking, eavesdropping, unauthorized access by third parties, unsecured data transmissions, re-identification of anonymized data and identity theft. Some of these risks also extend to the clinical and biomedical research field.

Neurotechnological devices that are deliberately developed to fall outside of medical device regulations, are often marketed as direct-to-consumer products. Therefore, they fall under the purview of consumer protection laws and regulation. However, current consumer protection (e.g. in the EU and the US) is a legal patchwork that may often allow companies to find regulatory loopholes [39]. Therefore, lawmakers and regulatory agencies should jointly work on defining a clear set of regulatory approaches to

consumer neurotechnology devices that apply within markets and may be harmonized across international markets.

An important step towards innovation governance was recently marked by the Recommendation on Responsible Innovation in Neurotechnology, which was adopted by the OECD in December 2019, setting the first international standards for responsible innovation in this domain [40].

Gaps in International Human Rights Frameworks and Further Lacunae Human Rights instruments, such as the UN's *Universal Declaration of Human Rights* (1948), which is legally binding as part of customary international law,[1] were drafted long before brain data became measurable outside the clinic and amenable to big data analytics. Given this, they did not explicitly spell out requirements for gaining access to and using brain data in a manner that protects individual rights. Whereas the conditions for legitimate use of human genetic data have been delineated in UNESCO's soft law *International Declaration on Human Genetic Data* (2003), human brain data remain without explicit safeguards and lack comparable protection by human rights instruments. In response to this, scholars have called for expanding the existing human rights framework as to explicitly include rights that are purposively designed to protect the brain and mind domain of a person, hence called neurorights. These rights can be seen either as evolutionary interpretations of existing rights or as new rights. Further, they constitute both rights in the legal sense (in accordance with international human rights law) and in the philosophical sense (in accordance with right-based moral philosophy) [41].

Further, there is no specific international treaty that addresses the dual-use or potential weaponization of brain data for military purposes. Dual-use research and technology collecting human brain data is therefore a pressing anticipatory governance concern as neurotechnology evolves and is increasingly researched in the military setting.

[1] It is, however, disputed whether the UDHR forms part of customary international law and it is difficult to conclude at this stage that the UDHR forms entirely part of it. However, some parts of the UDHR may be considered customary international law, e.g. the prohibition of torture.

Towards a Multi-Level Governance Framework

Advancing the use of brain data in neuroscience and medicine while simultaneously preventing ethical-legal risks requires a delicate balancing act. As brain data intersect several domains of human activity and regulation, it is unlikely that a one-size-fits-all approach to governance can be effective. Therefore, a comprehensive framework for global governance should operate adaptively at multiple levels. Based on the previously identified gaps, we propose four primary areas of regulatory intervention: binding regulation, ethics and soft law, responsible innovation, and human rights (Fig. 2).

A. Binding Regulation

Mandatory governance efforts seek to define and locate brain data within the supra-and-international data protection landscape. We suggest that brain data should be considered a special category of personal data that warrants heightened protection during collection and processing. If brain data are not considered a special category of personal data, they could be lawfully processed in ways that go beyond the limited circumstances set out in Article 9 of the GDPR. For example, they could be lawfully processed for purposes that are not health-related (e.g., for predicting consumer behaviour or for psychographic profiling). Further, they could be used for research activities that are not in the public interest and in the absence of an impact assessment. We posit that singling out brain data as a special category would help govern the non-medical use of these data while safeguarding their processing for scientific and biomedical purposes. This approach is consistent with the risk-based approach of the GDPR and could mimic the framing of other special categories of personal data such as genetic data (which includes chromosomal, DNA or RNA data; Article 4(13)). This would allow to protect brain data also prior to analysis, when they cannot be linked back to an identifiable individual or when they are generated by non-medical devices.

Additional provisions may clarify conditions for collecting and processing brain data in the non-medical space. At the data privacy level (e.g., as according to the GDPR), device and software manufacturers should ensure data protection "by design and by

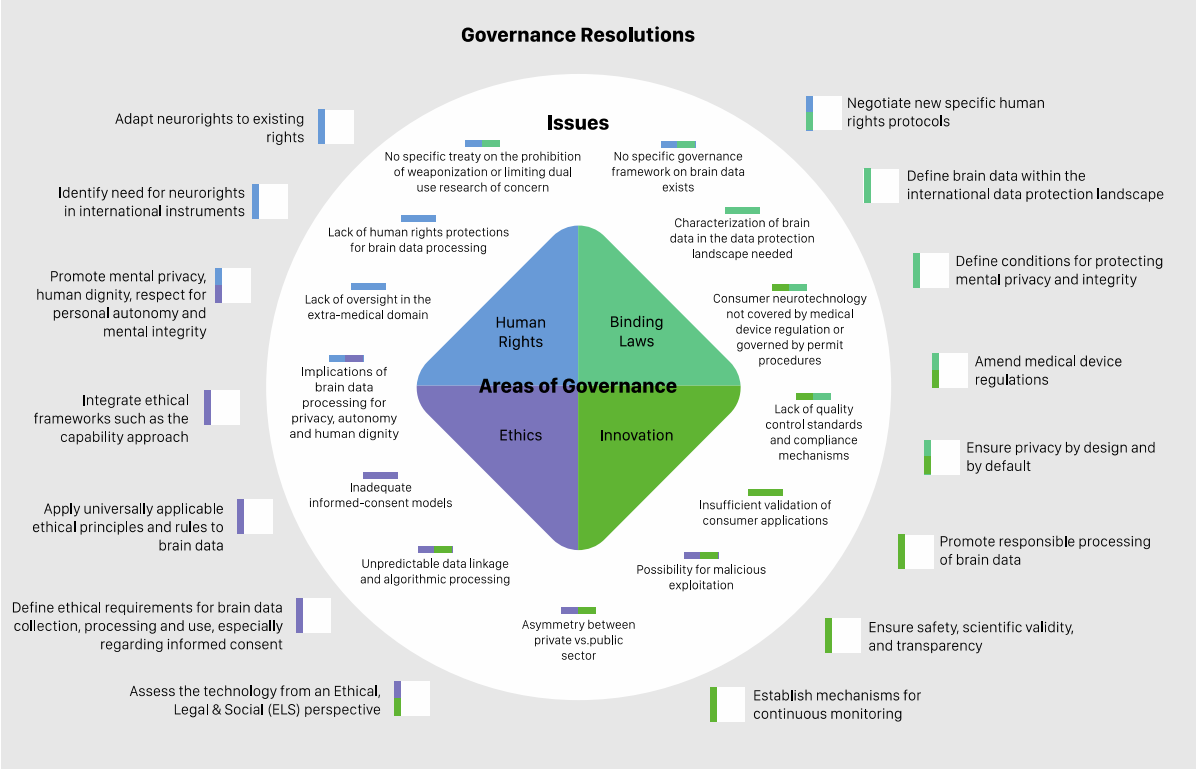


Fig. 2 Overview of normative requirements and levels of governance

default" (GDPR, Article 25). Further, data processors and controllers should use pseudonymization and encryption to guarantee data security (GDPR Articles 32–34) and implement the principle of data minimization. Additional measures may include protecting against third-party apps linked to consumer neurotechnology applications. Finally, the exact conditions and safeguards under which the research exemptions, introduced by Union or Member State law on the basis of Art 89 (2) GDPR, can permit brain data processing by private companies should be clarified. To fill a gap in international regulation, we contend that brain data indicating neurological or mental illness originating from non-medical neurotechnology should not be accessible by third-party actors such as health insurance providers. Access to such information would require the user's explicit and written (or digitally provided) consent.

More broadly, risks for privacy and human dignity specific to brain data analytics must be disclosed. In particular, regulators must consider whether a right to *mental privacy* and *mental integrity* should be granted to data subjects. These rights would grant subjects increased control and protection of data containing information about their sensory, cognitive, affective and volitional processes. In addition to data protection law, criminal and civil laws could reinforce these privacy rights by protecting a person's brain activity against unconsented exploration and modulation. Labor law offers grounds to protect employees from the misuse of their brain data in an employment context, e.g., by prohibiting employers from collecting brain data for productivity monitoring and terminating employment contracts based on brain data.

Another critical issue is the coercive collection of brain data. Governance frameworks should protect the ability of people to make free and competent decisions about the collection and processing of their personal brain data, a principle known as *cognitive liberty*. The *European Convention on Human Rights* (ECHR), which protects the rights to privacy and freedom of thought (Arts. 8 and 9) offers the suitable conceptual and normative framework to prevent coercive uses. If the CoE Modernised

Convention comes into force, it could serve as a solid basis for further specification and a model for other world regions.

In order to increase compliance and promote sustained scientific validation of new devices and algorithms in the gray zone between the medical and the non-medical domain, calibrated amendments to current medical device regulations should be considered. Currently, most consumer neurotechnology companies avoid classification of their products as medical devices by marketing them for wellness, relaxation and other non-medical purposes [27]. Nonetheless, users (including vulnerable people) may use those devices and share their brain data for health-related purposes, such as cognitive monitoring and mental wellbeing. A step towards reform was taken by the EU's amendments to the *Medical Devices Regulation*. These amendments will apply from May 2021 and cover also brain stimulation products without an intended medical purpose as medical devices (Annex XVI, No 6). However, it does not cover brain data processing for purposes other than neuromodulation. Furthermore, it remains highly uncertain whether and how regulatory agencies will take enforcement action.

Apart from peaceful purposes, the limits of exploring and modulating brain function for military usages must be defined. This is especially relevant as large military research agencies, such as the DARPA in the US, actively pursue brain stimulation technologies for modulating cognitive functions, such as memory and learning [5]. In an international context, brain data (as the decisive parameter for calibrating such neuromodulation devices) could thus become a commodity in a neurotechnology "arms race" as other nations also pursue military neurotechnology research and development. This arms race could involve both the development of novel military neurotechnology and the dual-use (repurposing) of consumer or medical technology [6]. The laws of war that are applicable during armed conflict [42] (so-called *international humanitarian law*) do not explicitly protect combatants against the violation of their mental integrity. Pending more in-depth analysis regarding the use of neurotechnology and the processing of brain data in the military context, there may be a need to draft legal guidelines—similar to those guiding autonomous weapons—that protect soldiers against brain data misuse during both wartime and peacetime.

B. Ethical Guidelines and Soft Law

Despite the difficulties of cross-border data transfers, brain data sharing practices are generally not restricted by national borders and regulatory frameworks. Therefore, internationally applicable ethical principles and rules are needed to govern the collection and processing of brain data. Research ethics procedures such as review through ethics committees and Institutional Review Boards (IRBs, which, in some countries and some areas of research, are part of binding law) are well-established governance mechanisms for the clinic and human neuroscience research. However, these procedures are insufficiently agile to respond to the novel challenges posed by the current big-data digital ecosystem, especially the innovation dynamics and business models of AI-based technology in the neuroelectronics marketplace. Similarly, the evaluative criteria of ethics review are not geared towards the current information-intensive ecosystem.

We posit that legitimate interest alone is insufficient to provide the ethical basis for brain data processing. In addition, consent should also be considered as critical ethical requirement for a brain data governance framework. This is consistent with the opinion of the European Group on Ethics in Science and New Technologies, which proposed to include individual consent as a requirement for further processing of health data in the EU regulation [43].

When collecting and/or processing identifiable brain data, private data collectors must conduct a legitimate interest assessment, check that the processing is necessary and there is no less intrusive (non-neural) way to achieve the same result. Further, they must document that *explicit* informed consent for a specific usage was obtained prior to data collection except in cases of medical emergency. Data collectors should be required to apply explicit informed consent procedures that go beyond the mere acceptance of ToU for consumer products. These procedures should transparently disclose and address, not less than: (i) how brain data are used, i.e. which information is decoded and with which accuracy; (ii) in which storage facility and on what medium data will be stored and the duration of storage; iii) the criteria and mechanisms by which access to the brain data is granted, monitored and revoked; (iv) how brain data are reused and shared; (v) what anonymization/pseudonymization and information security measures are

implemented; (vi) how individuals will be informed if their data are hacked, leaked or accidentally disclosed, and; (vii) what legal entity is liable for data breaches and other regulatory lapses. Novel digital technologies for informed consent (eConsent) have shown potential to enhance the practicability and efficacy of consent procedures [44]. In practice, in a clinical setting or in the consumer space, adherence to these procedures could be governed and monitored by Data Use and Access Committees, e.g., adjunct to IRBs in the clinic or consumer protection agencies.

We argue that the default consent for governing brain data use should be an opt-in approach. Accordingly, individuals have to explicitly *opt in* to sharing their brain data or link this data with other contextual information (e.g., social media profiles). Ethical guidelines should extend beyond mere rule-compliance and promote the respectful use of brain data.

C. Responsible Innovation

Responsible Research and Innovation (RRI) is now a widely accepted approach for guiding emerging sciences and technologies and promotes first and foremost the responsible collection and processing of brain data by both public and private actors. RRI principles can help develop safer and more reliable systems as well as increase preparedness to deal with unintended consequences. These include the adoption of community-agreed technical standards (e.g., within the neuroengineering community [45]), adequate validation and best practices by neurotechnology researchers, companies and other stakeholders in a consensus process.

Service providers who collect and process brain data should ensure safety, scientific validity, accountability and transparency. At the safety level, usage of brain data should consider and prevent inherent risks of algorithmic processing including bias, privacy violation, and cybersecurity vulnerabilities. Data collectors and processors should ensure data minimization, for instance by only providing data from some EEG channels or by selectively filtering certain frequencies in the data.

Novel privacy-preserving technologies can help both medical and non-medical processors. Technical approaches to improve protection from leakage and unwarranted access include homomorphic encryption, multi-party computation, federated learning, and differential privacy [46]. Differential privacy is particularly well-suited for brain data because it allows sharing aggregate data whilst preventing inferences from being drawn about individuals. Nonetheless, some risks can only be discovered once the systems have been deployed. Accordingly, developers shall establish mechanisms for continuous analysis, monitoring and mitigation of risk once software and devices are on the market.

Finally, data collectors and processors should ensure high standards of scientific validity for both devices and datasets. Consumer service providers should be prevented from advertising unsubstantiated paramedical claims (e.g. “improving mental wellbeing”) that are loosely founded, if at all, on scientific evidence [27]. Adequate testing and careful risk–benefit analysis should guide development and deployment of brain data processing systems. This will likely improve not only the safety, but also the efficacy, user-friendliness and precision of future devices. Similarly, regulators should take a proactive stance on the ethical, legal and social implications of these technologies. This proactive stance requires constant interaction between all stakeholders to identify suitable means for standardization, such as value-sensitive design. Oversight mechanisms involving binding regulation, soft law and ethical guidelines shall make sure that these standards are met by laying down the necessary permit procedures.

D. Human Rights

Brain data are inherent to and in principle accessible from all human beings, regardless of ethnicity, gender, nationality or religion. Further, they reflect the inner workings of our minds as they correlate with thoughts, emotions and other mental faculties. Therefore, the prospect of unsupervised deciphering of, interfering with and commodifying brain data raises serious human rights challenges. We posit that non-medical brain data processing for legitimate interest should not be pursued when the controller’s interest conflicts with the fundamental rights and freedoms of the data subject. Human rights inform legislation, ethical guidelines and societal norms across the globe, and thus offer an international normative framework where brain data protection needs to be

embedded. Interdisciplinary research investigating the intersection between brain data and human rights is ongoing and proposals for protecting neuro-specific rights, called *neurorights*, have been advanced [41, 46–50]. Governance frameworks should determine whether clauses can be added to human rights treaties or whether a new universal soft law instrument is necessary. This instrument could be modelled after the 2005 *UNESCO Declaration on Bioethics and Human Rights*. Furthermore, it should be determined whether neurorights should be interpreted as new rights or as adaptive interpretations of existing legally binding human rights and moral principles. These rights include the right to privacy, the right to freedom of thought, mental integrity and human dignity. The Universal Declaration of Human Rights, in particular, grounds human rights in the inherent and equal dignity of all human beings.

The normative force and universal claim of human rights often makes it difficult to translate these rights into guidance for context-sensitive action. For this reason, frameworks such as the capabilities approach [51] can be very helpful to translate the general requirements of human rights into actionable and shared international policy goals that promote human flourishing, human dignity and well-being in the context of brain data processing. Several national (e.g., Chile's recently approved Constitutional Reform and ongoing Neuroprotection Bill as well as Spain's Charter of Digital Rights) and international organizations actors (the United Nations, the Council of Europe, the EU Parliament, and the OECD) are putting "neurorights" on their agenda.[2]

Conclusions

International governance should ensure the positive impact of brain data processing on science, health, well-being, human dignity and human rights, while preventing potential risks for individuals and communities. We delineate a roadmap towards a global governance framework on brain data that can fill current ethical and legal gaps. We call upon professional societies, national and international organizations, as well as unrepresented or underrepresented communities and stakeholders (e.g., patient organizations) to take up the challenge and coordinate a joint effort at their adoption. Any move towards an international framework should be aware of cultural diversity and responsive to a pluralistic global society. Finally, following recent challenges in AI governance, we should avoid the uncoordinated proliferation of normative guidance in the absence of adequate strategies for harmonization, standardization and implementation.

Acknowledgements We would like to thank the two anonymous reviewers for valuable feedback on our manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This consensus paper was developed within the scope of a workshop funded by the Brocher Foundation which took place in Hermance, Switzerland, on November 25–27, 2019. The consensus development process involved a series of group activities including interactive panels, expert groups, and plenary sessions. Author Ricardo Chavarriaga acknowledges the support of the Digitalization Initiative of the Zurich Higher Education Institutions (DIZH) and the Polymath Fellowship program of the Geneva Center for Security Policy. Author Marcello Ienca has been supported in part by the ERA-NET NEURON project HYBRIDMIND (Swiss National Science Foundation 32NE30_199436). The work of author Philipp Kellmeyer was partially supported by a grant (00.001.2019) from the Klaus Tschira Foundation, Germany.

Declarations

Conflicts of Interests The authors declare that they have no conflicts of interest.

[2] See *inter alia*: Chile: https://spectrum.ieee.org/neurotech-neurorights; Spain: https://www.jurist.org/news/2021/07/spain-president-proposes-digital-rights-charter-outlining-fundamental-rights-of-individuals-online/; CoE: https://www.coe.int/en/web/bioethics/round-table-on-the-human-rights-issues-raised-by-the-applications-of-neurotechnologies; OECD: https://www.oecd.org/science/recommendation-on-responsible-innovation-in-neurotechnology.htm (all links retrieved March 30.th 2022).

References

1. Insel, T.R. 2017. Digital phenotyping: Technology for a new science of behavior. *JAMA* 318 (13): 1215–1216.
2. Ienca, M., P. Haselager, and E.J. Emanuel. 2018. Brain leaks and consumer neurotechnology. *Nature Biotechnology* 36 (9): 805–810.
3. Wang Y, S. Hong, C. Tai. 2019. China's Efforts to Lead the Way in AI Start in Its Classrooms. *The Wall Street Journal* [published Online First: Oct. 24, 2019].
4. Chen, S. 2018. Forget the Facebook leak': China is mining data directly from workers' brains on an industrial scale. *South China Morning Post* 29.
5. Emondi, A. 2019. Next-Generation Nonsurgical Neurotechnology: DARPA. Link: https://www.darpa.mil/program/nextgeneration-nonsurgical-neurotechnology.
6. Ienca, M., F. Jotterand, and B.S. Elger. 2018. From healthcare to warfare and reverse: How should we regulate dual-use neurotechnology? *Neuron* 97 (2): 269–274.
7. Frégnac, Y. 2017. Big data and the industrialization of neuroscience: A safe roadmap for understanding the brain? *Science* 358 (6362): 470–477. https://doi.org/10.1126/science.aan8866.
8. Landhuis, E. 2017. Neuroscience: Big brain, big data. *Nature* 541 (7638): 559–561. https://doi.org/10.1038/541559a.
9. Yuste, R., and C. Bargmann. 2017. Toward a global BRAIN initiative. *Cell* 168 (6): 956–959.
10. Grillner, S., N. Ip, C. Koch, et al. 2016. Worldwide initiatives to advance brain research. *Nature neuroscience* 19 (9): 1118–1122.
11. Castelluccia, C. 2020. From Dataveillance to Datapulation : The Dark Side of Targeted Persuasive Technologies. *Preprint*. Available at: https://hal.inria.fr/hal-02904926/
12. Price, W.N., and I.G. Cohen. 2019. Privacy in the age of medical big data. *Nature Medicine* 25 (1): 37. https://doi.org/10.1038/s41591-018-0272-7.
13. Zhou, L., S. Pan, J. Wang, et al. 2017. Machine learning on big data: Opportunities and challenges. *Neurocomputing* 237: 350–361.
14. Kellmeyer, P. 2018. Big Brain Data: On the Responsible Use of Brain Data from Clinical and Consumer-Directed Neurotechnological Devices. *Neuroethics* 1–16.
15. Wolkenstein, A., R.J. Jox, and O. Friedrich. 2018. Brain–computer interfaces: Lessons to be learned from the ethics of algorithms. *Cambridge Quarterly of Healthcare Ethics* 27 (4): 635–646.
16. Kellmeyer, P. 2020. Ethical issues in the application of machine learning to brain disorders. In *Machine learning*, eds. Mechelli A, Vieira S., 329–42. Academic Press.
17. Poldrack, R.A. 2011. Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron* 72 (5): 692–697. https://doi.org/10.1016/j.neuron.2011.11.001 [published Online First: 2011/12/14].
18. Marshel, J.H., Y.S. Kim, T.A. Machado, et al. 2019. Cortical layer-specific critical dynamics triggering perception. *Science* 365 (6453): eaaw5202. https://doi.org/10.1126/science.aaw5202.
19. Carrillo-Reid, L., S. Han, W. Yang, et al. 2019. Controlling Visually Guided Behavior by Holographic Recalling of Cortical Ensembles. *Cell* 178 (2): 447–57 e5. https://doi.org/10.1016/j.cell.2019.05.045.
20. Kay, K.N., T. Naselaris, R.J. Prenger, et al. 2008. Identifying natural images from human brain activity. *Nature* 452 (7185): 352–355.
21. Horikawa, T., M. Tamaki, Y. Miyawaki, et al. 2013. Neural decoding of visual imagery during sleep. *Science* 340 (6132): 639. https://doi.org/10.1126/science.1234330.
22. Moses, D.A., M.K. Leonard, J.G. Makin, et al. 2019. Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nature Communications* 10 (1): 3096. https://doi.org/10.1038/s41467-019-10994-4.
23. Hubbard, J., A. Kikumoto, and U. Mayr. 2019. EEG decoding reveals the strength and temporal dynamics of goal-relevant representations. *Scientific Reports* 9 (1): 9051. https://doi.org/10.1038/s41598-019-45333-6.
24. Omurtag, A., H. Aghajani, and H.O. Keles. 2017. Decoding human mental states by whole-head EEG+fNIRS during category fluency task performance. *Journal of Neural Engineering* 14 (6): 066003. https://doi.org/10.1088/1741-2552/aa814b [published Online First: 2017/07/22].
25. Faurholt-Jepsen, M., J. Busk, M. Frost, et al. 2016. Voice analysis as an objective state marker in bipolar disorder. *Translational Psychiatry* 6 (7): e856–e956. https://doi.org/10.1038/tp.2016.123.
26. Gordon, B.A., T.M. Blazey, Y. Su, et al. 2018. Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal dominant Alzheimer's disease: A longitudinal study. *The Lancet Neurology* 17 (3): 241–250.
27. Wexler, A., and P.B. Reiner. 2019. Oversight of direct-to-consumer neurotechnologies. *Science* 363 (6424): 234. https://doi.org/10.1126/science.aav0223.
28. Glanz, O., J. Derix, R. Kaur, et al. 2018. Real-life speech production and perception have a shared premotor-cortical substrate. *Scientific Reports* 8 (1): 8898. https://doi.org/10.1038/s41598-018-26801-x.
29. Koch, C., M. Massimini, M. Boly, et al. 2016. Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience* 17 (5): 307.
30. Bauer, P.J., T. Pathman, C. Inman, et al. 2017. Neural correlates of autobiographical memory retrieval in children and adults. *Memory* 25 (4): 450–466.
31. Greenberg, A. 2019. Inside the mind's eye: An international perspective on data privacy law in the age of brain machine interfaces. 29: 79.
32. Schwarz, C.G., W.K. Kremers, T.M. Therneau, et al. 2019. Identification of anonymous MRI research participants with face-recognition software. *New England Journal of Medicine* 381 (17): 1684–1686. https://doi.org/10.1056/NEJMc1908881.
33. Chizeck. H. J., T. Bonaci. 2014. Brain-Computer Interface Anonymizer. Patent Number: US20140228701A1.
34. Rainey, S., K. McGillivray, S. Akintoye, et al. 2020. Is the European Data Protection Regulation sufficient to deal with emerging data concerns relating to neurotechnology? *Journal of Law and the Biosciences*; 7(1):lsaa051 https://doi.org/10.1093/jlb/lsaa051.
35. Obar, J.A., and A. Oeldorf-Hirsch. 2020. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23 (1): 128–147.

36. Vayena, E., A. Mastroianni, and J. Kahn. 2013. Caught in the web: informed consent for online health research. *Science Translational Medicine* 5 (173): 173fs6.
37. Goering, S., and R. Yuste. 2016. On the necessity of ethical guidelines for novel neurotechnologies. *Cell* 167 (4): 882–885. https://doi.org/10.1016/j.cell.2016.10.029.
38. Eaton, M.L., and J. Illes. 2007. Commercializing cognitive neurotechnology—the ethical terrain. *Nature biotechnology* 25 (4): 393–397.
39. Dasgupta, I. 2020. Assessing current mechanisms for the regulation of direct-to-consumer neurotechnology. In *Developments in neuroethics and bioethics*, 233–65. Elsevier.
40. OECD-Council. 2019. OECD Recommendation on Responsible Innovation in Neurotechnology: Organisation for Economic Co-operation and Development. Available at: https://www.oecd.org/science/recommendation-on-responsible-innovation-inneurotechnology.htm.
41. Baselga-Garriga, C., P. Rodriguez, and R. Yuste. 2022. *Neuro rights: A human rights solution to ethical issues of neurotechnologies*, 157–161. Protecting the Mind: Springer.
42. Voeneky, S. 2020**.** Implementation and enforcement of international humanitarian law. In *The handbook of international humanitarian law*, ed. Fleck D, 4th ed, 647–700.
43. EGE. 2016. Opinion on the ethical implications of new health technologies and citizen participation: European Group on Ethics in Science and New Technologies. Available at: https://op.europa.eu/en/publication-detail/-/publication/e86c21fa-ef2f-11e5-8529-01aa75ed71a1.
44. Kuehn, B.M. 2013. Groups experiment with digital tools for patient consent. *JAMA* 310 (7): 678–680. https://doi.org/10.1001/jama.2013.194643.
45. IEEE. 2020. Standards Roadmap: Neurotechnologies for Brain-Machine Interfacing, 1–100. IEEE.
46. Yuste, R., S. Goering, G. Bi, et al. 2017. Four ethical priorities for neurotechnologies and AI. *Nature News* 551 (7679): 159.
47. Ienca, M., and R. Andorno. 2017. Towards new human rights in the age of neuroscience and neurotechnology. *Life Sciences, Society and Policy* 13 (1): 5. https://doi.org/10.1186/s40504-017-0050-1 [published Online First: 2017/04/27].
48. Ienca, M. 2021. On neurorights. *Frontiers in Human Neuroscience*, 15.
49. Herrera-Ferrá, K., J. M. Muñoz, H. Nicolini, et al. 2022. Contextual and cultural perspectives on neurorights: Reflections toward an international consensus. *AJOB Neuroscience* 1–9.
50. Kellmeyer, P. 2022. Neurorights: A Human-Rights Based Approach for Governing Neurotechnologies. In *The Cambridge Handbook of Responsible Artificial Intelligence - Interdisciplinary Perspectives*. 1st ed: Cambridge University Press.
51. Sen, A. 2005. Human rights and capabilities. *Journal of Human Development* 6 (2): 151–166.

Machine-Learning Based Monitoring of Cognitive Workload in Rescue Missions with Drones

Fabio Dell'Agnola, *Member, IEEE,* Ping-Keng Jao, Adriana Arza, *Member, IEEE,* Ricardo Chavarriaga, *Senior Member, IEEE,* José del R. Millán, *Fellow, IEEE,* Dario Floreano, *Senior Member, IEEE* and David Atienza, *Fellow, IEEE*

***Abstract*— In search and rescue missions, drone operations are challenging and cognitively demanding. High levels of cognitive workload can affect rescuers' performance, leading to failure with catastrophic outcomes. To face this problem, we propose a machine learning algorithm for real-time cognitive workload monitoring to understand if a search and rescue operator has to be replaced or if more resources are required. Our multimodal cognitive workload monitoring model combines the information of 25 features extracted from physiological signals, such as respiration, electrocardiogram, photoplethysmogram, and skin temperature, acquired in a noninvasive way. To reduce both subject and day inter-variability of the signals, we explore different feature normalization techniques, and introduce a novel weighted-learning method based on support vector machines suitable for subject-specific optimizations. On an unseen test set acquired from 34 volunteers, our proposed subject-specific model is able to distinguish between low and high cognitive workloads with an average accuracy of 87.3% and 91.2% while controlling a drone simulator using both a traditional controller and a new-generation controller, respectively.**

***Index Terms*— Cognitive Workload Monitoring, Physiological Signals, Machine Learning, Human-Robot Interaction, Wearable Systems, Search and Rescue Missions.**

I. Introduction

THANKS to recent enhancements in both robotics and human-robot interfaces, the interest in deploying robots in search and rescue (SAR) missions is growing [1]. However, limitations exist in their effective and efficient utilization in real-life missions. The main limitation is that robot teleoperation is a non-intuitive and challenging task. Thus, SAR robots are still constrained to simple missions and highly trained professionals. [2], [3]. Moreover, rescuers have to simultaneously focus on multiple tasks and deal with both scarcity of human resources and time pressure. This situation is cognitively highly demanding and can negatively affect performance [4], [5]. Consequently, operating under high cognitive workload (CWL) may severely compromise the execution of a mission and leads to failure with catastrophic outcomes [6]. Therefore, there is a need to monitor CWL to ensure efficient execution of SAR missions.

To assess CWL, researchers typically use surveys [7], performance metrics [8], [9], and information from physiological signals [10]. However, surveys only provide subjective and sporadic measurements, and are not always reliable [11]. Although performance metrics provide objective measurements, reliable metrics are difficult to set as every rescue mission is unique. On the other hand, physiological signals can be noninvasively acquired without disturbing the rescuers' work. Thus, the use of physiological signals seem the most promising solution to assess Cognitive Workload Monitoring (CWM) [10], [12], [13].

Several studies combine physiological signals with different machine-learning algorithms for CWM in different fields [13], [14]. However, to the best of our knowledge, we are the first to address CWM of drone pilots involved in SAR missions [8], [15], [16]. Now, we extend our previous works by presenting a subject-specific CWM approach based on noninvasive physiological signals that is suitable for new drone control solutions, such as FlyJacket [17]. In particular, this work proposes the following contributions:

- We explore different feature normalization techniques to reduce both inter-subject and inter-day variability;
- We provide a new weighted-learning method for Support Vector Machine (SVM), suitable for subject-specific optimizations. This SVM based method uses two regularization terms, one for learning the general behaviour and another for tuning the model to fit the characteristics of a particular data subset;
- We prove the ability of our method to detect low and high CWL levels while controlling a drone simulator with traditional and advanced controllers, achieving an accuracy of 87.3% and 91.2%, respectively. These results are obtained on unseen data acquired from 34 participants while flying a drone simulator and mapping a graphic representation of a disaster situation. Our results are higher than the latest state-of-the-art studies in SAR missions with drones (see Table I).

II. Related Work

CWL characterization and estimation have been addressed by a large number of studies [12], [25], which characterize

This work was partially supported as a part of NCCR Robotics, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (Grant No. 51NF40_185543) and by the ONR-G through the Award Grant No. N62909-17-1-2006.

F. Dell'Agnola, A. Arza, and D. Atienza are with the Embedded Systems Laboratory (ESL), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland (e-mail: fabio.dellagnola@alumni.epfl.ch, adriana.arza@epfl.ch, david.atienza@epfl.ch).

P.-K. Jao, R. Chavarriaga, and J.d.R. Millán are with EPFL, Lausanne 1015, Switzerland. J.d.R. Millán is also with the Dept. of Electrical and Computer Engineering & the Dept. of Neurology, The University of Texas at Austin, Austin, TX 78712, US (e-mail: ping-keng.jao@alumni.epfl.ch, ricardo.chavarriaga@alumni.epfl.ch, jose.millan@epfl.ch).

D. Floreano is with the Laboratory of Intelligent Systems (LIS), EPFL, Lausanne 1015, Switzerland (e-mail: dario.floreano@epfl.ch).

TABLE I
SUMMARY OF THE STATE-OF-THE-ART STUDIES USING MULTIPLE PHYSIOLOGICAL SIGNALS

Study	Performed Tasks	Physiological Signals	Window Length (Overlap)	Classifier (Classes)	Results Acc.	Sens.	Spec.
Momeni et al. [15]	Simulated SAR with drones	ECG, RSP, PPG, SKT	60s (30s)	XGB (2)	86%*	-	-
Dell'Ágnola et al. [18]	Simulated SAR with drones	ECG, RSP, PPG, SKT	60s (0s)	XGB (2)	80.2%*	79.6%*	71.7%*
Montesinos et al. [19]	Arithmetic tasks	ECG, PPG, RSP, SKT, EDA	60s (30s)	RF (2)	84.13%*	-	-
Chen et al. [20]	Real car driving	ECG, RSP, EDA	100s (90s)	SVM (3)	89.7%	88.5%	94.2%
Solovey et al. [21]	Driving in highway	ECG, EDA	30s (0s)	LR (2)	90%	-	-
Giakoumis et al. [22]	Video-game	ECG, EDA	25s (0s)	LDA (2)	94.96%	94.96%	94.96%
Tjolleng et al. [23]	Simulated driving task	ECG	100s (0s)	ANN (3)	82%	78%	91%
Gjoreski et al. [24]	Daily life activities	PPG, SKT, EDA	300s (150s)	SVM (2)	98.96%	70.44%	99.88%

ECG-electrocardiogram, RSP-respiratory activity, PPG-photoplethysmogram, SKT-skin temperature, EDA-electrodermal activity. XGB-Extreme Gradient Boosting, RF-Random Forest, SVM-Support Vector Machine, LR-Logistic Regression, LDA-Linear Discriminant Analysis, ANN-Artificial Neural Network.
* Results based on an unseen test set, all the other are limited to cross-validation.

either the performance or the distress of a person involved in a particular task or situation. In this section, we review the state-of-the-art machine learning (ML) techniques detecting CWL induced by high cognitive tasks. In particular, we analyze those works using unobtrusively measured physiological signals. Although interesting for their results, studies relying on obtrusive measurements (e.g., electroencephalography [26]) are not included in this analysis since their integration into a jacket is difficult or unattainable. The same applies to works placing sensors in locations other than the torso, such as the head [27].

Table I summarizes the most recent and significant studies including the performed task to induce CWL, measured physiological signals, signal segmentation (i.e., window length and overlap), applied machine-learning methods, targeted classes, and classification results (i.e., Accuracy, Sensitivity, and Specificity). Our analysis identifies the following common methodological steps: signal acquisition and preprocessing (filtering and segmentation), feature extraction, feature normalization, dimension reduction or feature selection, and classification or regression. However, although the methodology is well established, discrepancies are found in different steps. Hence, in the following, we review these discrepancies.

First, significant differences have been observed on the physiological measures, which are electrodermal activity (EDA) [19]–[22], [24], [28], electrocardiogram (ECG) [18], [20]–[23], [29], photoplethysmogram (PPG) [15], respiratory activity (RSP) [15], [20], and peripheral skin temperature (SKT) [15], [18]. Although using multiple physiological signals can increase the detection accuracy of CWL levels [15], the type and number of signals, and in particular the features set, often differ and strictly depend on the case study (e.g., the type of task used to induce different levels of CWL) [10], [29]. Thus, there is no clear definition of the best selection of signals and features to assess CWL in general.

Then, the segmentation window used to extract the features from the signals also depends on the case study. In particular, the window lengths reported in Table I vary from 25 to 300 seconds. Moreover, different window overlaps are applied either to increase the size of the dataset [15] or to provide more frequent estimations in time [20], [24]. These differences can be explained by the fact that physiological methods do not provide a direct measurement of the workload, but rather they give information about how the individuals themselves respond to a particular load [10]. So, a different signal segmentation may be applied depending on the dynamic of the physiological response induced by a particular CWL.

An additional aspect observed in our literature review is that features are often normalized to standardize their ranges. The normalization help to reduce intra- and inter-subject variability caused by age, time of day and other factors [30]. However, not all studies report whether a normalization was applied [30], or clearly explaining how it was done and distinguishing between training and test sets. To properly emulate and test the system's behaviour, test data should be normalized based on the parameters obtained from the training set [30].

Moreover, the choice of machine-learning methods clearly differ. The train data size and the system requirements specification (e.g., computational complexity, power and latency) may explain the different selections of machine-learning algorithms. In fact, as most of the studies typically start with a limited amount of data, simple models like Support Vector Machine (SVM) [20], [24], [31], Linear Discriminant Analysis (LDA) [22], [28], Logistic Regression (LR) [21], and Decision Tree (DT) [19], are the most used machine-learning techniques. In contrast, complex models such as Artificial Neural Networks (ANN) [23], Random Forest (RF) [19], [31], and very recent models like Extreme Gradient Boosting (XGB) [15], have been less used so far. In any case, even if SVM has been the most used classifier in this field, there is no consistent indication of whether it is the best model or not for different case studies.

Finally, our review shows that the highest accuracy levels are in the range of 80 to 99%. This wide range is mainly due to the diverse experimental protocols, methodologies, and number of considered classes in each study. Also, the highest accuracies reported by different studies may be affected by overfitting since their model evaluation is limited to cross-validation [20]–[24]. However, a proper estimation of a model's generalization power requires a final test on new unseen data, a set never used in training [15], [18], [19].

In conclusion, there is a need to investigate further the contribution of each physiological signal, the impact of data normalization, and the performance of the selected classifier on unseen data in the context of rescue missions with drones, which are not appropriately covered in the literature.

Besides, workload is multidimensional [7] and results from the aggregation of three broad aspects [10], [32]. First, the workload depends on the task's type (mental or physical demand), and the load level (e.g., tasks amount and difficulty). Second, it is affected by time, namely, by the duration of the temporal demand. Third, the subjective psychological experiences modulate the level of workload perceived by a subject (i.e., subject's capabilities, learning skills, and effort). So, it is necessary to investigate CWL in the particular field of interest and, also, consider each person's subjective workload level, as suggested in [33].

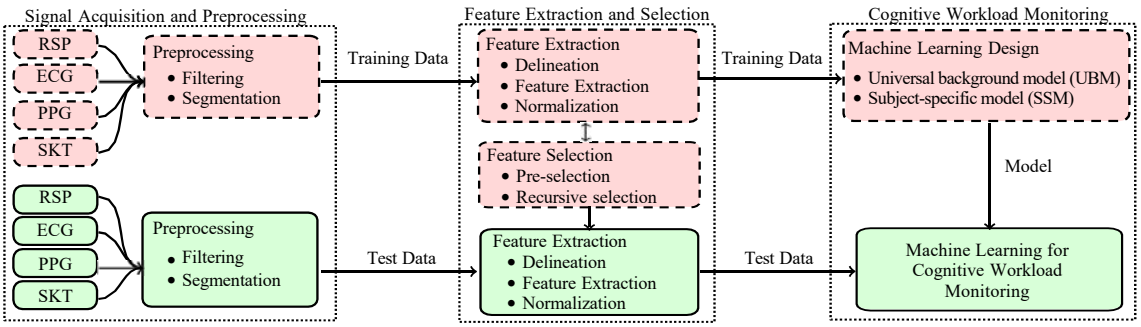


Fig. 1. Process overview for the design of a CWM method. Blocks with dashed lines represent the applied design/optimization methods and blocs with solid lines represent the final system.

III. CWM System

The general design of an ML algorithm suitable to develop a wearable embedded system for online CWM is shown in Fig. 1, namely, blocks with solid lines. Instead, the blocks with dashed lines represent the different statistical pattern recognition methods applied to experimental data for designing such a system. All our analyses are done offline, but the final system is tested, emulating online processing (i.e., using causal spectral filters and computing the features using past information).

The system is divided into three main steps shown in Fig. 1 with the dotted lines, i.e., Signal Acquisition and Preprocessing, Feature Extraction and Selection, and CWM. In the first step of the CWM system, sliding window is applied for signal segmentation, which defines the time resolution of the workload monitoring system. The preprocessing consists of removing artifacts from the signals. In this work, we collected experimental data for both design and evaluation of the proposed CWM method.

Next, the features extraction and selection step includes generating a feature vector that best represents the physiological response induced by different workloads. For an exhaustive investigation, we chose an exploratory approach in which we extract a large number of different features in both time and frequency domains. Then, since physiological signals exhibit high intra- and inter-subject variability due to age, gender, time of day and other factors [30], we investigate different normalization methods. Subsequently, we apply different features selection methods to define the best subset of features to be used in the final system.

Finally, the CWM step includes the prediction of a discrete CWL level. For the design of the CWM method, we consider the most common machine-learning techniques based on pattern recognition algorithms suitable for implementation in embedded systems. Moreover, we consider a personalized weighted-learning approach to assess the person-dependent variance in the physiological response of an induced workload. Performance of our method is then evaluated based on NASA Task Load Index (NASA-TLX), a subjective and multidimensional assessment tool that rates perceived workload [7].

IV. Signal Acquisition and Preprocessing

For a thorough exploration of the physiological changes induced by cognitive workload, we measure RSP, ECG, PPG, SKT, EDA, and EEG, which are signals that are typically used in the literature [34], [35]. The effect of cognitive workload on EEG was analyzed and presented in a different work [36]. Here, we focus on the remaining signals, which sensors can be integrates into a wearable system, such as FlyJacket [17].

Their main physiological manifestations related to CWL are reported in Table II and described in Sec. IV-A.

A. The physiological process behind CWL

While performing a very demanding task, the need for more oxygen is driven by the autonomic nervous system (ANS) activation. The latter involves both a sympathetic nervous system (SNS) activation and parasympathetic nervous system (PSNS) counterbalance. This increased oxygen demand triggers faster and deeper respiration [37]. Therefore, RSP should be measured to track CWL changes [20].

The ANS activation also triggers a cardiac response, which is also affected by the Hypotalai-Adrena (HPA) axis. This response is associated with variabilities in heart rate, defined as heart rate variability (HRV) obtained by monitoring the ECG signal. Consequently, the above relationship can explain the heart's ability to respond to multiple physiological and environmental stimuli [8]. The neurohypophysis activation, the HPA axis, and the ANS lead to blood volume changes, peripheral blood vessels resistance, and cardiac response derived from the pulse wave. Features from the PPG are used to detect those physiological changes induced by cognitive tasks [24], [37].

Moreover, it has been proved that cognitive tasks cause peripheral vasoconstriction [24], [37], regulated by the vasoregulatory system and driven by both neurohypophysis and SNS. Thus, SKT is required to detect the variations in peripheral temperature that are associated with peripheral vasoconstriction.

Finally, EDA is one of the most commonly used measures in studies involving emotional arousal. According to [38], EDA is traditionally measured at the fingers or palms, while foot and shoulders seems to be valid alternatives for ambulatory measurement. However, we cannot confirm their findings, as our EDA measurements from the shoulder did not show any significant response. Therefore, EDA measurements were not considered in this work.

B. Signal preprocessing

The first preprocessing step consists of removing the artifacts from the signals with causal filters [16]. We apply a baseline wander with cutoff frequency at 0.3 Hz to both ECG and PPG signals. Next, we also apply a 32nd-order bandpass

TABLE II
PHYSIOLOGICAL MANIFESTATIONS RELATED TO INDUCED CWL.

Physiological measures	Measurable physiological manifestation to workload response	Sensor body position
Peripheral skin temperature	Neurohypophysis and Sympathetic Nervous System (SNS) activation	Finger
Respiration	SNS activation and Parasympathetic Nervous System (PSNS) counterbalance	Thorax
Electrocardiogram	Both Hypotalai-Adrena (HPA) axis and SNS activation, and PSNS counterbalance	Thorax
Photoplethysmography	Neurohypophysis, HPA axis, and SNS activation, and PSNS counterbalance	Ear

FIR filter with linear phase and Hamming window with cut-off frequencies at 0.3 and 30 Hz for ECG and at 0.1 and 5 Hz for PPG [37]. In the case of the RSP signal, we employ a 4th-order Butterworth IIR bandpass filter with cutoff frequencies at 0.03 and 0.9 Hz. Nevertheless, because of the slow response time of the SKT thermistor (1.1 sec.), which avoid the high frequency noise, no filter is applied to the acquired SKT signal.

Finally, we apply a time-series segmentation of all the acquired physiological signals, which are thus divided into a sequence of samples in windows of 60 seconds.

V. FEATURES EXTRACTION AND SELECTION

Following our methodology described in Section III, we perform an offline investigation to select the features to be considered in the final system. That is, we first extract a broad features set from the segmented signals for an exhaustive assessment of the person's physiological response to CWL. Then, we select the best features set rich in discriminatory information concerning the physiological states induced by different CWL levels. normalized and given as input to the developed CWM algorithm.

A. Feature extraction

For the design of the CWM system, our feature extraction process includes three main steps. First, we delineate the segmented signal to detect points of interest (e.g., signal onset, peak, offset, etc.). Second, we extract physiological markers, a combination of different delineated points and provide information about the person's physiological state (e.g., heart rate). Finally, we compute features in both time and frequency domains. For the time domain, we use standard statistical features (i.e., mean, median, mode, standard deviation, variance, root mean square, and power), extracted either from the physiological markers or from the segmented signals directly. However, in the frequency domain, the features are computed specific to the characteristic of the physiology of each signal, which are listed and detailed next.

Following an extensive literature review and by applying our experience from previous projects [8], we increased the number of analytical methods applied to a single physiological signal segment to extract 384 features: 127 from RSP, 38 from ECG, 190 from PPG, 2 from SKT, and 27 from RSA. However, applying our feature selection method, the final system uses only 25 features, 10 from RSP, 2 from ECG, 10 from PPG, 2 from SKT, and 1 from RSA. These 25 features are listed in Table IV. From EDA, we aimed to compute the mean skin conductance level and the number of skin conductance responses per minute as in [38]. Though we used dedicated electrodes (recommended by Biopac), our EDA signal was rudely flat across participants suggesting a poor SNS activation on the shoulders for our study case. Thus, the signal was discarded. More details about the delineation and feature extraction for each considered signal are provided next.

Fig. 2 shows a schematic representation of the signal processing and feature extraction process.

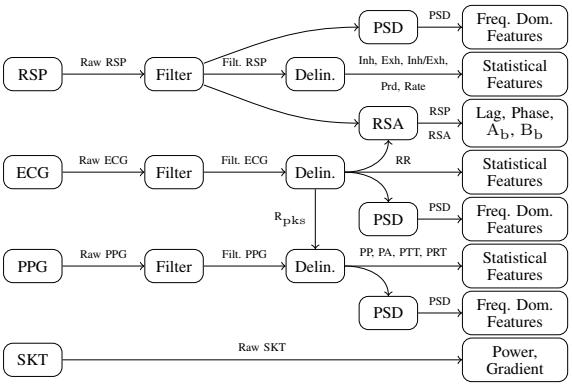


Fig. 2. Schematic representation of the signal processing and feature extraction processes.

1) Respiratory activity (RSP): To extract the features from the RSP signal, we first delineate the signal based on the differences between adjacent samples of the filtered signal defined as:

$$\Delta x[k] = x[k] - x[k-1] \tag{1}$$

Then, by comparing both current and previous values, we detect from the sign of Δx the falling and rising edge, which coincide with inhalation (RSP-peaks) and exhalation (RSP-valleys) end, respectively. Then, all peaks and valleys pairs having a difference smaller than 20% of the mean RSP amplitude are removed [31].

Next, from the delineated RSP, we extract the following physiological markers: inhalation (Inh) and exhalation (Exh) time, the Inh/Enh ratio, Inh and Exh amplitudes, respiratory period ($\mathrm{RSP}_{\mathrm{Prd}}$), and respiratory rate ($\mathrm{RSP}_{\mathrm{Rate}}$). Besides, we compute their numerical differences using Eq. 1. Finally, we calculate the segmented RSP signal's statistical features, its difference (Eq. 1), and all the aforementioned RSP physiological markers. In the frequency domain, we compute the power of the segmented signal in four different bands of equal bandwidth (i.e., 0-0.25, 0.25-0.5, 0.5-0.75, and 0.75-1 Hz), as reported in [28]. We also consider the normalized band powers, obtained by dividing each of the above band powers by the total power in the 0-1 Hz band.

2) Electrocardiogram (ECG): We compute the so-called normal-to-normal (NN) intervals from the filtered ECG signal, the intervals between normal QRS complexes detected with the delineation method described in [39]. Then, we compute features in the time domain describing the Heart Rate Variability (HRV) [40], which are statistical features of the successive NN-intervals and of the interval differences of successive NN-intervals. We also computed the number of interval differences of successive NN-intervals greater than 50 ms (NN50) and the proportion derived by dividing NN50 by the total number of NN-intervals (pNN50) within the processing window.

Additionally, we obtain several geometrical features from the Poincaré (or Lorenz) plot indicating vagal and sympathetic

functions. In particular, we extract the length of the transverse axis (T), vertical to the line $\mathrm{NN}_k = \mathrm{NN}_{k+1}$; the length of the longitudinal axis (L), parallel with the line $\mathrm{NN}_k = \mathrm{NN}_{k+1}$; the Cardiac Sympathetic Index (CSI), defined as L/T; the modified CSI (L^2/T); and the Cardiac Vagal Index (CVI) as $log_{10}(LT)$ [40].

Moreover, we extract HRV features from the frequency-domain, as proposed in [40]. That is, the power in two frequency bands, namely, low-frequency (LF: between 0.04 and 0.15 Hz) and high-frequency (HF: between 0.15 and 0.4 Hz). LF and HF powers are obtained from estimating of the Lomb-Scargle Power Spectral Density (PSD) of the NN intervals [41]. The power values are divided by the total power minus the very-low-frequency (VLF) component (frequency $\leq$ 0.04 Hz). Also, we compute the power sum LF + 1/HF and the ratio LF/HF.

Furthermore, we extract novel features from the HF band. The first one, called $\mathrm{RR}_{\mathrm{HF\ gauss}}$, is the mean frequency of a Gaussian distribution used to fit the Lomb-Scargle PSD estimated in the HF band. This feature describes the shifting in frequency of the PSD in the HF band, where the shift is mainly caused by the RSP activity [42]. The second one is called $\mathrm{RR}_{\mathrm{HF\ pond}}$ and is defined as:

$$\mathrm{RR}_{\mathrm{HF\ pond}} = \frac{\sum_{f\in \mathrm{HF}} f\ \mathrm{PSD}\{RR[k]\}(f)}{\sum_{f\in \mathrm{HF}} PSD\{RR[k]\}(f)} \tag{2}$$

Finally, we also compute the power of the HF divided in 5 sub-bands of equal length ($\mathrm{RR}_{\mathrm{HF\ sband\ Xn}}$), where the subscript index $X = \{1, \cdots, 5\}$.

3) Photoplethysmogram (PPG): According to [37], we delineate the PPG signal and extract the following physiological markers: the Pulse Period (PP), the time interval between two consecutive pulse peaks; the Pulse Amplitude (PA), the difference between the pulse peak and the pulse onset; the Pulse Transit Time ($\mathrm{PTT}_{\mathrm{M}}$), the time interval between the R-Peak in the ECG signal and the instant when the PPG pulse reaches half of its onset-to-peak amplitude; the Pulse Rise Time (PRT), the time interval between the pulse onset and the pulse peak; and the Pulse Rise Speed (PRS), the ratio between amplitude difference and time interval computed from the pulse wave points located at 75% and 25% of the onset-to-peak amplitude, respectively.

To have accurate estimations of PTT and PRT, in the literature [16], the use of both ECG and PPG signals has been proposed. Using both enables trade-offs between accuracy and complexity of the sensing wearable system.

From each of the aforementioned PPG physiological markers, we extract features in the time and frequency domains, following the HRV methodology applied to NN-intervals.

4) Peripheral Skin Temperature (SKT): From the SKT signal, we directly extract the $\mathrm{SKT}_{\mathrm{Gradient}}$ and $\mathrm{SKT}_{\mathrm{Power}}$ of the signal. The SKT Gradient is computed as the mean of the difference between the portion of samples recorded during the first second of the window, acquired at a sampling frequency f_s, and the samples from the final one second of the window. Then, the SKT Power is the signal average power of computed over the entire window of samples.

5) Respiratory Sinus Arrhythmia (RSA): Respiratory sinus arrhythmia (RSA) is the natural variation in the heart rate associated with the respiratory cycle. and measured from the ECG signal. RSA has been used as a noninvasive measure of cardiac vagal tone, as a marker of PSNS tone [43] and thus, it can be used as a marker of the disruption of homeostasis induced by a highly demanding task. Since RSA and cardiac vagal tone can dissociate under certain circumstances [44], we consider the hypotheses that these differences could come from external factors, such as, a need to compensate for CWL changes.

RSA is estimated from the non-uniform time series of successive NN-intervals, which we interpolate using a linear function and resample at 2 kHz. Then, we filter the resulting uniform time series of successive NN-intervals with a 4th order band-pass Butterworth filter with cutting frequency at 0.15 and 0.4 Hz yielding a RSA.

From the computed RSA we extract features that aim to evaluate the agreement with the measured RSP signal, but first, both signals (RSP and RSA) are normalized to zero mean and unit variance. The first feature is the time delay of the RSA with respect to the RSP ($\mathrm{RSA}_{\mathrm{Lag}}$), estimated by computing the cross-correlation of RSA and RSP. We also compute the phase shift between the two signals, given by Eq. 3.

$$\mathrm{RSA}_{\mathrm{Phase}} = \cos^{-1}\left(\frac{\mathrm{RSP}\cdot\mathrm{RSA}}{\|\mathrm{RSP}\|\cdot\|\mathrm{RSA}\|}\right) \tag{3}$$

Subsequently, we extract features based on the Tukey mean-difference plot, also called the Bland-Altman plot [45], to compare both RSA and RSP measurements. To this end, we compute the statistical features of the difference between the two signals and the mean of the two:

$$R_0 = \mathrm{RSP} - \mathrm{RSA} \tag{4}$$

$$A_0 = (\mathrm{RSP} + \mathrm{RSA})/2 \tag{5}$$

We also consider the statistical features of different log transformations of the measurements, as follows:

$$R_b = log_b(\mathrm{RSP}) - log_b(\mathrm{RSA}) \tag{6}$$

$$A_b = (log_b(\mathrm{RSP}) + log_b(\mathrm{RSA}))/2, \forall b = \{n, 2, 10\} \tag{7}$$

where b denote the logarithm base (i.e., n, 2, and 10).

B. Features Normalization

Since the relative range of each feature varies widely, a normalization is applied so that each one contributes approximately equally to the classification problem. Hence, we apply a min-max normalization scaling the features within a 0-1 range. The general formula is given as:

$$\mathbf{x}' = \frac{\mathbf{x} - \min(\mathbf{x}^\dagger)}{\max(\mathbf{x}^\dagger) - \min(\mathbf{x}^\dagger)} \tag{8}$$

where $\mathbf{x}$ is an original value, $\mathbf{x}'$ is the normalized one, and $\mathbf{x}^\dagger$ represents the original value of the training set.

Moreover, to address the problem related to both inter-subject and inter-day variability [8], [30], we found from the computational vision community, a task-specific normalization method [46], which inspired us to consider the following three types of normalization. First, the total normalization (TN) is based on the full training set. Second, the subject dependent normalization (SN) consists on normalizing based on each training subset relative to a specific subject. Finally, the day and subject-dependent normalization (DSN) affects each portion of the training set relative to a specific day and subject. Thus, the training and the test sets are scaled accordingly, using the parameters obtained only from the training set.

Finally, we select the best normalization strategy that better emphasizes the discriminant power of the features and their ability to classify the problem. In other words, we select the method that gives the highest Fisher Discriminant Ratio (FDR) [47] of the normalized feature sets, obtained by applying one

of the three different normalization methods (i.e., TN, SN, or DSN). Then, we evaluate the classification performance of an SVM that uses for each normalized set an equal number of normalized features. The results are reported in Sec. VIII-B.

C. Features Selection

Given the large features number considered for the exhaustive characterization of CWL, we divide the feature selection process into two main steps. First, as a pre-reduction to suppress the features that do not give any discriminatory information, we apply filter methods, particularly effective in computation time and robust to overfitting. Then, to select the most important features considering their possible interactions, we apply embedded methods that simultaneously perform feature selection and classification. Both feature selection steps are performed once with data from the training set.

The pre-reduction of the feature space involves three methods. First, a two-sample Student's t-test selects statistically discriminant features. Second, the discriminant features are ranked based on their FDR, which gives a score based on their ability to discriminate the problem. Lastly, we remove the features that give any redundant information, the less discriminant features that are strongly correlated with others (i.e., a Pearson's correlation coefficient above 0.95) [48].

For the final feature selection, we apply Recursive Features Elimination (RFE) [49], an embedded method that uses an external estimator to assign weights to features. These weights are then used to prune the least important features from the current set. This procedure recursively prunes the selected features until all feature weights are different from 0. In this work, we apply RFE based on different classifiers (i.e., LR, LDA, SVM, RF, and XGB), which we name RFE-LR, RFE-LDA, RFE-SVM, RFE-RF, and RFE-XGB, respectively.

VI. Cognitive Workload Monitoring

For the cognitive workload monitoring, we explore the use of different machine-learning algorithms. In particular, we investigate the use of linear models, namely LR, LDA, SVM, and Gaussian Naive Bayes (GNB) for a feasibility check. Then, we investigate the use of non-linear models, such as k-Nearest Neighbour (k-NN), Quadratic Discriminant Analysis (QDA), SVM with a Radial Basis Function (RBF) kernel, DT, RF, and XGB, to reduce the bias. The accuracy of each model in detecting high levels of CWL is evaluated based on a 5-fold cross-validation (CV) over the training set.

Moreover, we consider a personalized weighted-learning approach to deal with the person-dependent variance. To this aim, we compare the performance of the Universal Background Model (UBM) and the Subject-Specific Model (SSM) [50].

A. Model for Cognitive Workload Monitoring

To estimate CWL, we chose a linear SVM that has the following prediction model [51]:

$$y(x) = \mathbf{w}^T\mathbf{x} + b \tag{9}$$

where $\mathbf{x}$ is the input vector, $\mathbf{w}$ is the weight vector, and b is the offset. The corresponding optimal hyperplane separating the two classes is defined by the relation:

$$y(x) = \mathbf{w}^T\mathbf{x} + b = 0 \tag{10}$$

Thus, an input vector $\mathbf{x}$ is then assigned to class 1 if $y(x) \geq 0$ and to class -1 otherwise. Although we use the same prediction model for UBM and SSM, the difference lies in the objective function. All the details are given in Sec. VI-B and VI-C.

The parameters of both UBM and SSM are chosen based on a 5-fold CV on the training set. We use a stratified split for this validation that preserves the same percentage for each target class as in the complete training set and preserves the same percentage of data relative to the subject of interest. Then, the generalization of both models is tested on an unseen test set.

The performance of the models is evaluated based on: accuracy, the proportion of both true positives and true negatives results among the total number of cases; precision, or confidence, the proportion of predicted positive cases that are correctly real positives; recall, or sensitivity, the proportion of real positive cases that are correctly predicted positive; Receiver Operating Characteristic (ROC); and in particular, based on the F1-score, the weighted average of the precision and recall.

B. Training of the Universal Background Model

The considered UBM is based on SVM with soft margins [51], which relax the condition for the optimal hyperplane (Eq. 10) and allow possible overlaps of the class-conditional distributions. As for a normal soft-margin SVM, the objective function of the UBM is defined as follows:

$$\begin{aligned} \underset{w,b,\xi_i}{\arg\min} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i\in D}\xi_i, \\ \text{subject to} \quad & t_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0; \quad (i \in D) \end{aligned} \tag{11}$$

where the regularization term C and the non-negative variables ξ_i relax the constraints of an otherwise hard-margin SVM. The data x in the training dataset D comprises N input vectors $x_1, \cdots, x_N$, with corresponding target values $t_1, \cdots, t_N$, and where $t_i \in \{-1, 1\}$. The parameter C is analogous to the inverse of a regularization coefficient because it controls the trade-off between minimizing training errors and controlling model complexity. A regularization term $C = 0.1$ is chosen from a $\log_{10}$ scale ranges from 0.001 to 1000 based on a stratified 5-fold CV on the training set.

C. Training of the Subject-Specific Model

As well as for the UBM, the considered SSM is based on a soft-margin SVM. However, to adapt the model to a specific subject, we modify the objective function of the original soft margin SVM (Eq. 11) including two different soft-margins. The first soft-margin (C_s) changes the importance degree given to false estimations of samples coming from a particular subset of data, which can be a particular subject (S). Thus, the term weighed by C_s allows a minimization of the errors (ξ) for all the x in the training set related to a specific subject ($x \in S$). Instead, the second soft-margin (C) affects the rest of the dataset minimizing the errors ξ for all the x in the training set that are related to other subjects ($x \notin S$).

Therefore, the SSM final objective function is defined as:

$$\begin{aligned} \underset{w,b,\xi_i}{\arg\min} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i\notin S}\xi_i + C_s\sum_{i\in S}\xi_i \\ \text{subject to} \quad & t_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0; \quad (i \in D) \\ & C_s > C \end{aligned} \tag{12}$$

With this model, we state a preference for margins that classify the training data correctly, but we soften the constraints to allow for non-separable data with different penalties. To promote the minimization of the total sum of the penalties

$\xi_i \ \forall i \in S$, despite the minimization of the total sun penalties $\xi_i \ \forall i \notin S$, we chose C_s to be greater thar usual, the regularization terms have to be large en avoid under-fitting, but not too much to avoid over-fi well. Based on a stratified 5-fold CV on the training s regularization terms $C = 0.001$ and $C_s = 0.1$ are chos a $\log_{10}$ scale in ranges of 0.001–0.1 and 0.1–100, resp Although both regularization terms seem to be bounde considered range, we keep the lower bounds to avoid under-fitting problems.

VII. Experimental setup

Collecting data in a real SAR mission is complex of the random frequency of events and the many v still undefined. Therefore, for collecting clean data, bu CWM model, and validating our approach, we used th lator for search and rescue mission with drones reporte With the help of a certified instructor of the Swiss fire we designed the following two study protocols, where based on a repeated-measures design using counterba The first study was conducted to characterize CWI through physiological signals using a gamepad as controller, to build a model for real-time monitoring, and to evaluate the contribution of the subject-specific approach.

The second protocol was designed to evaluate the system's quality using a new advanced controller, the FlyJacket [17]. In contrast with the gamepad controller, where the movements were limited to the thumbs, the FlyJacket implies both arms and torso movements. Therefore, when comparing tasks involving different types of movements, there is a risk of yielding a performance overestimation. Thus, to avoid as much as possible any possible miss-classification caused by movement artifacts, we trained the machine-learning algorithm with the data from Study 1 (Trial 1) with the gamepad and did the SSM final tuning with data from Study 2 with FlyJacket. Finally, our models were tested also on unseen data of Study 2 with FlyJacket. The details of both studies are in the following sections.

The signal processing, features extraction, machine-learning design, and classification were done using Matlab R2016a [52]. The RSP, ECG, PPG, SKT, and EDA were recorded with the Biopac MP160 system at 2 kHz of sampling frequency. We also recorded EEG, but because of the difficult integration of such a sensor into a jacket, it is not used in this work. Instead, it is analysed in [36], as previously mentioned. Finally, through an analog input of the Biopac system, a trigger signal provided by the simulator advises the task execution.

A. Search and rescue drone simulator

As presented in [8] and [36], the simulator presents a simplified SAR scenario, where the drone pilot has to deal with two different activities, flying and mapping. The flying activity consists in flying a drone following a randomly generated trajectory depicted by spherical waypoints. Instead, the mapping activity consists of mapping a disaster area situation, represented by cubes of 4 different colors randomly distributed over the flying trajectory. The colors were chosen according to the regulation of the Swiss Firefighters [53].

We modulate both flying and mapping activities to induce different levels of CWL as in [8], [36] i.e., medium/high workload level with Flying (F) and Mapping 3 objects (3M), and high level of CWL with Flying and Mapping 3 objects (F3M). Also, a flying sequence controlled by an auto-pilot is used as Baseline (B) to have participants in a same framework for the entire experiment. B task has the lowest expected workload level of this study.

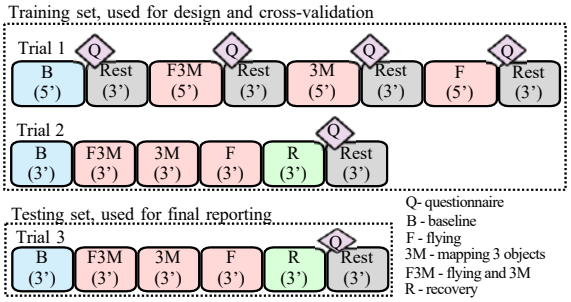


Fig. 3. Protocol of the experiment with the gamepad.

B. Study protocol 1: Use of a gamepad

During this study, participants sat in front of a screen and controlled the simulator with a gamepad from Logitech. To collect clean data, participants were asked not to talk and to avoid any kind of unnecessary movements during the tasks. For proving the feasibility of detecting cognitive workload with constrained sensor placement, clean data were needed. Hence, we asked the participants not to talk and avoid unnecessary movements while performing the tasks. However, we cannot completely avoid the presence of some artifacts. Therefore, in this context, different methods can be applied to make sure the input data can be used for our proposed algorithm. In particular, different approaches in wearables have been shown to be effective for noise removal (e.g., for speech [54], [55] and movement [56], [57] artifacts), which are needed in real-life scenario. The study started with a setup phase (explanation about the experiment, request of the participant consent, and sensor placement), followed by a warm-up phase up to 10 minutes to get familiar with the simulator [58].

The study protocol is shown in Fig 3. Participants performed the first trial, starting with a five-minute baseline, and followed by a sequence including F3M, 3M, and F, executed in a randomized order. A resting period of 3 minutes was enforced after each task. This period also allowed participants to fill a questionnaire (Q), based on the NASA-TLX procedure.

Finally, the participants performed two additional trials, namely Trial 2 and Trial 3. Each trial started with a baseline and continued with a randomized sequence of F3M, 3M, and F, and ended with a recovery (R) phase followed by a resting period, in which the NASA-TLX was filled again. Each task presented in Trial 2 and 3 lasted three minutes.

As shown in Fig. 3, we used all data acquired during both Trial 1 and Trial 2 for both training and CV, and all data collected during Trial 3 as the final unseen test set. We are conscious that this split does not truly respect independent temporality of data because all data sets (i.e., training, CV, and unseen test sets) are taken from the same day and not from a day that is not used for testing (as it should be in a real application). Therefore, this choice implies a daily training phase, which can be seen as a daily calibration of the system. However, as we expect an inter-day variability of the physiological responses [8], [30], we assume that a daily calibration of the system will be required. This calibration process consists of tuning the model for the correct baseline level by using a couple of minutes of data collected under both low and high workloads. A further investigation over different

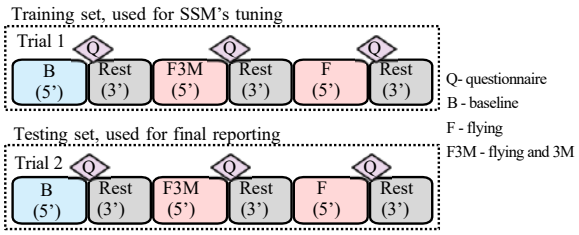


Fig. 4. Protocol of the experiment with FlyJacket.

days could potentially avoid the need for such a calibration, but this analysis is left for a future study.

C. Study protocol 2: Use of FlyJacket

In this study, the drone simulator is controlled with the FlyJacket and two Oculus Touch controllers to map the disaster situation. The study also started with a setup and warm-up phase. Then, participants performed two trials, as shown in Fig. 4, which started with a five-minute baseline followed by a F3M and F sequence executed in a randomized order. Again, three-minute resting period was enforced after each task, where the participants filled the questionnaire.

This second study is a reduced version of the first one since it aims to prove the feasibility of detecting low and high CWL levels with the proposed method. Hence, we designed this study protocol with only two trials, with three tasks of five minutes each, and recording F for a different study [15].

D. Research participants

Study 1 with the gamepad was done by 24 participants (6 females and 18 males) aged between 21 and 39 years old (27.7 ± 4.8), who performed the study protocol twice in two sessions on different days. Study 2 with the FlyJacket was done by 10 additional participants (3 females and 7 males) aged between 22 and 30 years old (26.8 ± 2.3), on a single day session. All participants provided informed consent to participate in both studies. The inclusion criteria were being healthy, free of any cardiac abnormalities, and were receiving no medical treatment. The Cantonal Ethics Commissions approved this study for Human Research Vaud and Geneva (PB2017-00295).

VIII. Experimental Results

Given the recorded data set from Study 1, we select the best combination of normalization, feature selection, and classification methods suitable for CWM. The methods are obtained based on the cross-validations workflow including 747 observations. Finally, we show the performance of the proposed methods on two unseen test sets, including 260 and 57 observations from Study 1 and 2, respectively.

A. Self-perception of induced cognitive workload

The reported overall workload on each task perceived by the 34 participants based on the NASA-TLX is shown in Fig. 5. A one-way ANOVA conducted on the influence of the tasks confirms that participants have perceived different levels of workload. Furthermore, a multiple pairwise comparison analysis using the Student's t-test with up to 164 samples revealed statistically significant mean differences, except for 3M vs F (p-value < 0.001). The comparisons with the 3M task

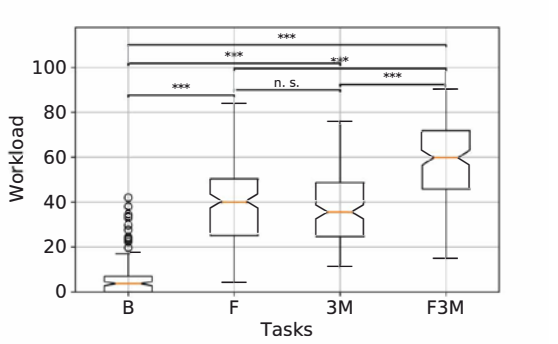


Fig. 5. Cognitive workload perceived by participants.

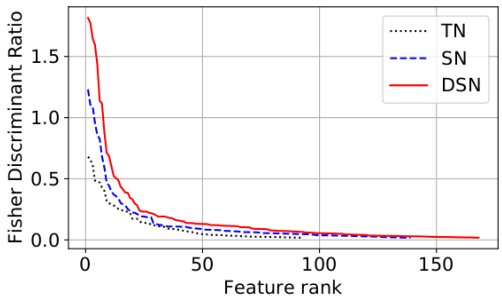


Fig. 6. Normalization methods impact on FDR.

were limited to 144 samples, as Study 2 with the FlyJacket setup does not include the 3M task.

However, as shown in Fig. 5, the perceived CWL level has a large variance. A two-way ANOVA reveals that such a large variance comes from a significant ($p < 0.001$) effect of task, day, and subject on the level of CWL, $F(3,414) = 1637.19$, $F(1,414) = 28.70$, $F(33,414) = 48.93$, respectively. Therefore, the NASA-TLX results confirm the need for both a day- and a subject-specific approach.

Although there is a significant difference in the perceived workload between most tasks, Fig. 5 shows that the distribution of both F and 3M presented a considerable overlap with F3M. Instead, the difference between tasks B and F3M is clear. Thus, as our main goal is to detect low and high levels of CWL, we focus on the extreme cases induced by tasks B and F3M, respectively. F and 3M conditions were analysed in a different work [15], which targets a three-class CWM.

B. Features discriminant power emphasized by normalization

To reduce the variance introduced by the different participants and performing the experiment on different days, we investigated different normalization approaches (i.e., TN, SN, and DSN) as described in Section V-B. We firstly evaluated the effect of each normalization approach on the features discriminant power based on their FDR. Results are shown in Fig. 6, where DSN better emphasises the discriminant power of the features. Compared with TN, the FDR of the most important feature is emphasized by a factor of 80.9% or 166.9%, over SN or DSN, respectively.

Secondly, following our methodology (see Sec. III), we compare how each normalization approach contributes to the classification problem using a linear SVM model. We noticed that the normalization affects the feature selection process,

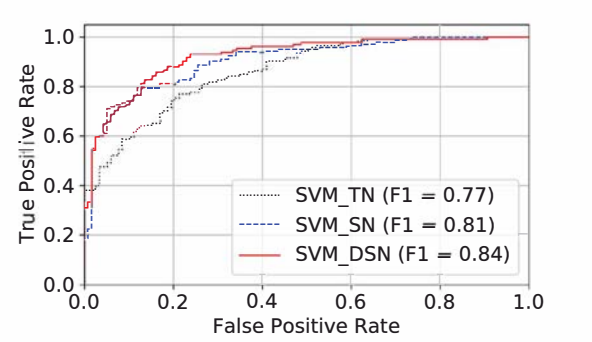


Fig. 7. Normalization methods impact on CWM.

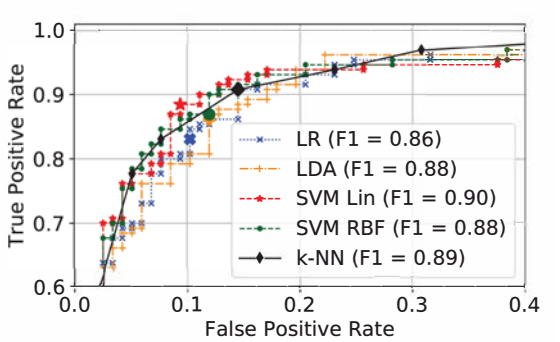


Fig. 8. Best classifiers comparison on CV. Bigger markers denote the performance of the different models based on their corresponding cross-validated threshold or offset b.

which selects 14 features after TN or SN, or 25 features after DSN. Therefore, to avoid biased results caused by the use of a different number of features, we used for this comparison the first 14 most discriminant features selected by RFE-SVM after TN, SN, or DSN normalization. Fig. 7 shows the ROC and the F1-score of the SVM combined with the different normalization methods, where it can be seen that once again DSN outperforms both TN and SN.

Our results show that feature normalization plays an important role during both features selection and classification. DSN normalization gives better results (a bigger F1-score) compared to SN and TN. Similar trends are obtained by applying RFE with other classifiers, such as LR or LDA. Therefore, we select DNS as normalization method.

C. Physiological featuring of cognitive workload

By applying the filter methods presented in Section V-C, we eliminated 282 non-informative features from the normalized (based on DSN) 384 features initially considered for an exhaustive CWL characterization. In particular, we reduced the feature space dimension from 384 down to 168 features with the two-sample Student's t-test and down to 102 features by checking their linear correlation.

Although the above pre-selection step drastically reduced the feature space, using that amount of features requires models with high capacity. It may lead to overfitting if trained with a limited dataset like ours. Therefore, to obtain a reasonable feature set that can be used for CWM, a further dimension reduction based on embedded methods was applied, as presented in Section V-C.

The features space was reduced from 102 to 5, 10, 12 and 25 by applying RFE-XGB, RFE-LR, RFE-LDA, RFE-SVM, respectively. RFE found a consistent set of features based on LR, LDA, and SVM, see Table III. For the case of RFE-XGB, we used a low-complex model to avoid overfitting and inconsistent results. In particular, we limited the model to 10 estimators and three maximum depth of each decision tree. Such a low-complex RFE-XGB showed a drastic lower selection compared to other methods.

Without banning the ensemble methods from building complex models, RFE does not converge to the same result if executed several times. In contrast, by limiting the model complexity, RFE provides a reproducible result. However, this trick does not help the RFE-RF method that does not converge to a consistent solution. This model always selects a different set of features, even if the model complexity is reduced (i.e., number of estimators and maximum tree depth). Hence, such complex models are not suitable for small datasets.

The feature set obtained after applying both filter and embedded methods are shown in Table IV. Although selected features vary between 5 and 25 depending on the applied embedded methods, a common subset of features is identified. We observed that the features obtained by RFE-LR, RFE-LDA and RFE-XGB are almost all included in the feature set obtained by RFE-SVM. In particular, $\mathrm{RSP}_{\mathrm{Rate\ Median}}$ and $\mathrm{SKT}_{\mathrm{Power}}$ are selected by all the four methods, followed by $\mathrm{RSP}_{\mathrm{Prd\ Median}}$, $\mathrm{SKT}_{\mathrm{Gradient}}$, $\mathrm{RSA}_{\mathrm{R2\ Std}}$, $\mathrm{PRT}_{\mathrm{Median}}$, $\mathrm{RSP}_{\mathrm{Rate\ Diff\ RMS}}$ and $\mathrm{PP}_{\mathrm{Median}}$, selected by three methods out of four. Based on this result, the above eight features seem to be the most important ones in terms of CWL characterization in the context of this experiment.

Additionally, we investigated the effect of using the different feature sets obtained with the considered RFE methods on different classification methods. Results are presented in Table III, where we report both the training and the CV accuracy. A significant difference between training and CV accuracy indicates a sign of overfitting (e.g., QDA with 102 features). Moreover, we report the best CV F1-score for each applied RFE method. While there seem to be no significant differences across methods, the highest best F1-score and the best CV accuracy are reached when linear SVM is applied on both RFE and classification. Therefore, RFE-SVM is the employed feature selection method hereafter.

D. Classifiers for cognitive workload monitoring

A ROC curve is used to further evaluate the performance of the considered classifiers in CV, reported in Fig. 8 . In particular, for greater clarity of the illustration, we only report the best classifiers results ($\mathrm{AUC} \geqslant 0.94$), namely LR, LDA, k-NN, linear SVM, and SVM with RBF kernel. Our results show that, with the amount of data we have, the use of non-linear models does not increase the detection accuracy. Instead, non-linear models tend to introduce a larger variance between training and CV- accuracy. Linear SVM shows a higher F1-score and better ROC curve, in particular by comparing the bigger markers representing the performance of the models based on their corresponding cross-validated threshold or offset b. Therefore, a linear SVM was selected for our further investigation.

Although selecting the SVM reaches the highest classification accuracy, it may not be the optimal solution for embedded implementations. Other solutions considering fewer features may be preferred for implementations in low-power embedded systems, where power consumption may play an important role. However, our results indicate certain flexibility in selecting the number of features to be used, since the best F1-score is quite similar for all the applied feature selection embedded methods.

TABLE III
FEATURE SELECTION PERFORMANCE COMBINING RECURSIVE FEATURE ELIMINATION (RFE) AND CLASSIFICATION METHODS.

Embedded Feature Selection		Training and cross-validation accuracy of different classifiers, left and right values, respectively																		F1-score
Method	Features	LR		LDA		QDA		SVM_{Lin}		SVM_{RBF}		GNB		k-NN		DT		RF		XGB		Best
Pre-selection	102	93	85	94	84	100	82	88	87	92	87*	84	85	91	85	90	85	90	87	94	87	89
RFE-XGB	5	85	85	85	86	87	86	85	87*	87	88*	85	88*	90	87*	89	84	90	85	90	85	88
RFE-LR	10	90	86	91	85	90	87*	89	86	91	86	86	86	91	85	90	80	90	85	92	84	88
RFE-LDA	12	90	87	91	87	92	87	87	86	92	88	86	85	92	87*	88	76	90	85	91	85	89
RFE-SVM	25	91	86	92	87	95	88	89	88*	93	87	88	88	92	87	90	84	92	85	93	85	90

* highlights the classifier having the best F1-score on cross-validation for the particular feature selection method. Logistic Regression (LR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Support Vector Machine (SVM), with linear and Radial Basis Function (RBF) kernels, Gaussian Naive Bayes (GNB), Nearest Neighbour (k-NN), Decision Tree (DT), Random forest (RF), and Extreme Gradient Boosting (XGB).

TABLE IV
MOST IMPORTANT FEATURES USED TO DETECT LOW AND HIGH LEVELS OF CWL, B AND F3M TASKS, RESPECTIVELY.

Physiological Features	Task B $\mu \pm \sigma$	Task F3M $\mu \pm \sigma$	p-Val$< 10^{-x}$ x
$RSP_{\text{Rate Mean}}$ [1,4]	0.28 ± 0.22	0.71 ± 0.23	107
$RSP_{\text{Rate Median}}$ [1,2,3,4]	0.28 ± 0.23	0.71 ± 0.23	106
$RSP_{\text{Prd Mean}}$ [1,2,3]	0.61 ± 0.25	0.22 ± 0.21	90
$Inh_{\text{Time Median}}$ [1]	0.53 ± 0.31	0.20 ± 0.23	52
$Exh_{\text{Time Median}}$ [1]	0.63 ± 0.29	0.32 ± 0.28	45
$Inh_{\text{Time Mean}}$ [1]	0.50 ± 0.31	0.22 ± 0.25	38
$Inh_{\text{Time RMS}}$ [3,4]	0.44 ± 0.30	0.19 ± 0.25	31
$RSA_{\text{R2 Std}}$ [1,3,4]	0.40 ± 0.28	0.57 ± 0.28	17
$RSP_{\text{Pks Mode}}$ [1]	0.61 ± 0.29	0.50 ± 0.30	08
$RSP_{\text{Rate Diff RMS}}$ [1,2,3]	0.33 ± 0.29	0.42 ± 0.29	06
RSP_{PSD3n} [1]	0.35 ± 0.29	0.43 ± 0.30	05
RSP_{PSD1n} [1]	0.48 ± 0.35	0.41 ± 0.31	04
$RR_{\text{HF gauss}}$ [1,3]	0.32 ± 0.23	0.68 ± 0.26	74
$RR_{\text{HF sband 3n}}$ [1]	0.47 ± 0.32	0.30 ± 0.28	15
$RR_{\text{Lorenz L2}}$ [2]	0.49 ± 0.30	0.35 ± 0.26	11
RR_{CVI} [2]	0.54 ± 0.29	0.42 ± 0.28	10
$PP_{\text{HF sband 5n}}$ [1]	0.23 ± 0.25	0.46 ± 0.30	28
PA_{RMS} [1,2]	0.53 ± 0.35	0.32 ± 0.27	20
$PA_{\text{Lorenz L}}$ [1]	0.44 ± 0.33	0.26 ± 0.25	17
PRS_{Mean} [2]	0.38 ± 0.35	0.55 ± 0.31	13
PP_{CSI} [1]	0.46 ± 0.29	0.33 ± 0.26	11
$PA_{\text{CSI modified}}$ [2]	0.40 ± 0.30	0.28 ± 0.27	09
PRT_{Median} [1,2,3]	0.44 ± 0.31	0.56 ± 0.31	08
$PTT_{\text{M Mode2}}$ [1]	0.50 ± 0.35	0.58 ± 0.28	05
PP_{Median} [1,2,3]	0.55 ± 0.31	0.47 ± 0.28	05
$PTT_{\text{M HF pond}}$ [1]	0.47 ± 0.28	0.54 ± 0.29	05
PRT_{LFp1oHF} [1]	0.38 ± 0.31	0.30 ± 0.27	05
PP_{Mode2} [1]	0.55 ± 0.33	0.49 ± 0.28	04
SKT_{Power} [1,2,3,4]	0.61 ± 0.35	0.37 ± 0.30	24
SKT_{Gradient} [1,2,3]	0.57 ± 0.29	0.38 ± 0.26	20

Selected feature with: [1]SVM-RFE, [2]LDA-RFE, [3]LR-RFE, and [4]XGB-RFE.

Besides, the CV-accuracy reported in Table III after RFE is delimited between 84 and 88%, except for DT. The CV-accuracy variability seems to be more dependent on the selected classifier (difference $> 4.5\%$) rather than the selected number of features (difference $< 3.5\%$). In fact, a linear SVM with an input of only five features can provide a reduced implementation complexity with a loss of only 1% of classification accuracy.

E. Classification improved with the SSM

Once we have selected the set of features (i.e., 25 features with RFE-SVM) and a linear SVM as classification method, we tested the subject-specific approach contribution compared to a general model (i.e., SSM vs. UBM). First, we trained the models as described in Section VI. The regularization term $C = 0.1$ of the UBM was selected based on a 5-fold CV on the training set. For the SSM, we selected $C = 0.001$ and $C_S = 0.1$ being the most common regularization terms found with a 5-fold CV on the training (data of Study 1).

TABLE V
PERFORMANCE OF THE UNIVERSAL BACKGROUND MODEL (UBM) VS. SUBJECT-SPECIFIC MODEL (SSM) ON AN UNSEEN DATA.

Study	Model	class	precision	recall	F1-score	samples
Study 1 Gamepad	UBM	B	0.81	0.76	0.79	123
		F3M	0.80	0.84	0.82	137
		avg	0.80	0.80	0.80	260
	SSM	B	0.89	0.83	0.86	123
		F3M	0.86	0.91	0.88	137
		avg	0.87	0.87	0.87	260
Study 2 FlyJacket	UBM	B	0.87	0.93	0.90	29
		F3M	0.92	0.86	0.89	28
		avg	0.90	0.89	0.89	57
	SSM	B	0.88	0.97	0.92	29
		F3M	0.96	0.86	0.91	28
		avg	0.92	0.91	0.91	57

Table V reports the comparison between UBM and SSM, tested on an unseen test set emulating an online CWM. The average accuracy of the UBM is 80.4%, and it is improved to 87.3% by the use of the SSM. The SSM shows a statistically significant improvement of the classification performance indicated by both the Wilcoxon rank-sum test [59] and the McNemar's test [60] over the 260 samples (p-value < 0.01). SSM improves the results for all the participants on CV, while one participant over 24 does not show the expected improvement on the final test set. This result may be explained by the need for more training data that could be used to better fit this participant's physiological response.

Furthermore, as shown in Table V, the higher performance of the SSM compared to the UBM is also confirmed on the test set acquired using FlyJacket (Study 2, Sec VII-C). In fact, the UBM reached a global accuracy of 89.5% that is improved to 91.2% using the SSM. However, the improvement (1 sample over 57) is not statistically significant, shown by both Wilcoxon rank-sum and McNemar's tests. For statistical results, additional data are needed. Nevertheless, a single misclassified sample in SAR missions can have a significant impact.

SSM obtains better performance than UBM because uses all the observations with a different weight. Those from other participants contribute to learn the general behaviour, with a regularization term C that allows a higher misclassification of such observations. Then, specific subject observations tune the margins between classes with a regularization term C_S to reinforce each specific subject. In light of the above, we can conclude that the personalized model performs in general better than the universal model.

Our results for the SSM are comparable with the state-of-the-art (See Table I), in particular with the work presented in [15], where the authors achieved an accuracy of 86%. Although with similar accuracy, our model is less complex and uses a reduced feature number. Another important difference is the test set selection, which was random in [15]. Instead, as a test set, we selected data from the last trial performed by

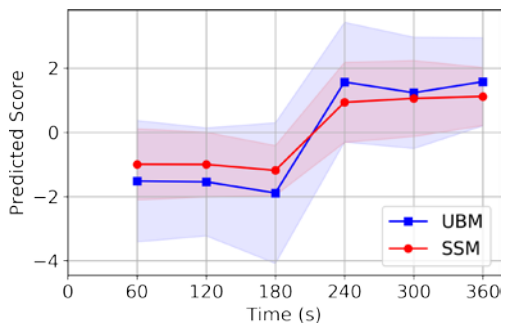


Fig. 9. Models performance comparison on a simulated online CWM (every 60s). The first 180 s correspond to B task followed by F3M task for other 180 s.

each subject, namely, Trial 3. For any classification problem that breaks the interchangeability hypothesis, such as the time-dependent CWM, a random training/test split should be avoided, as it yields a biased model evaluation. With a random split, the model learns from prospective data, commonly not available when designing and training a prediction model. Besides, the model is evaluated based on retrospective data, which are too similar to the training data. Hence, the classifier tends to look better than it is. Therefore, to estimate how well a model will work with new data, a time-dependent training/test split should be considered.

Also, the improved classification performance with FlyJacket vs. gamepad is assigned to the increased amount of training data. In the case of FlyJacket, the classifier weights were tuned based will all collected data from Study 1 and including Trial 1 of Study 2. Nevertheless, considering less training data (ignoring Day 2 of Study 1) reduces the accuracy of UBM from 86% to 82%.

Then, we assessed if our model using FlyJacket could suffer from possible movement artifacts, as they would differ from Task B to Task F3M. Thus, we minimized this risk because 93.5% of the samples used to train the classifier comes from Study 1, with the gamepad, in which the movements were minimal and limited to the thumbs. Moreover, all the features, normalization coefficients, and regularization terms were chosen using data only from Study 1. Thus, movement artifacts cannot significantly influence our classification results.

F. Emulated online cognitive workload monitoring

A visual representation of the emulated online CWM of both UBM and SSM is shown in Fig. 9. Since the order of the tasks was randomized, we only report the 76 samples of the sequences having consecutive transitions between B and F3M tasks. This analysis is based on Study 1 performed with the gamepad (Trial 3). During the first 180 seconds, participants performed the B task, a low workload level. For the last 180 seconds, participants performed the F3M task, a higher workload level. The detection was done on the test set, where features were extracted from a 60-second sliding window with no overlap. Negative and positive scores denote low and high workloads, respectively. A Wilcoxon rank-sum test with 76 samples indicates that the scores before and after 180 seconds are significantly different (p-value $< 10^{-8}$).

Another interesting aspect to note from Fig. 9 is the contradictory difference between the averaged predicted scores of the UBM and SSM. As the SSM is performing better than the UBM, we would expect to see a bigger absolute value of the SSM averaged score than the one of the UBM. However, the upper margin of the standard deviation of the predicted score reported in the interval between 60 and 180 seconds (Task B) and the lower margin in the interval between 240 and 360 seconds (Task F3M) seems similar for both UBM and SSM. This behaviour may be explained by the attempt of the SVM to choose the hyperplane that maximized the distance from it to the nearest data point on each side. Thus, as the SVM tends to maximize the margins, the SVM-based SSM performance may be limited to a consistent but marginal improvement.

Finally, comparing Fig. 5 and Fig. 9, we can see that both perceived and detected CWL are affected by a large variance. However, as shown in Fig. 9, such a variance is partially reduced using the SSM, which contributes better to fit the physiological response of a single subject.

IX. Conclusion

In this work, we have proposed a reliable subject-specific machine-learning algorithm for real-time CWM in SAR missions with drones. Our multimodal CWM model combines the information of features extracted from physiological signals (i.e., RSP, ECG, PPG, and SKT) noninvasively acquired. After an exhaustive investigation involving up to 384 features, we have selected only 25 required to get the highest classification accuracy. In addition, we have explored different feature normalization techniques to reduce both subject and day inter-variability, showing that a combination of day and subject normalization improves the detection accuracy.

Moreover, we have introduced a novel SVM based weighted-learning method suitable for subject-specific optimizations. With such a method, we distinguish between low and high CWL with an accuracy of 87.3%, on an unseen test set. Furthermore, we tested our model on ten new subjects using an advanced controller, reaching an average accuracy of 91.2%. Therefore, our model is valid to monitor CWL from rescuers piloting a drone with either traditional or advanced controllers.

The proposed methodology paves the way for detecting high levels of cognitive workload with sensors that can be included into a jacket. Our model can already operate in real-time to obtain information of the cognitive workload of the user. Such information can be used to improve shared-control systems by modulating the human-robot interaction and dynamically adapt the level of assistance, which will ensure an efficient execution of the missions. However, further investigations in real-life scenarios are needed to model other stressful conditions, which are not reproducible in laboratory tests. Moreover, there is a need to address a fine-grained detection in order to define a threshold for preventing a possible pilot's overload that could compromise the outcome of a search and rescue mission.

References

[1] I. Management Association, *Robotics: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications*, ser. Essential reference. IGI Global, 2013.

[2] J. Casper and R. R. Murphy, "Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2003.

[3] J. Y. C. Chen *et al.*, "Supervisory control of multiple robots: Human-performance issues and user-interface design," *IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews)*, vol. 41, no. 4, pp. 435–454, Jul 2011.

[4] K. H. Teigen, "Yerkes-dodson: A law for all seasons," *Theory & Psychology*, vol. 4, no. 4, pp. 525–547, Nov 1994.

[5] A. Marinescu *et al.*, "Exploring the relationship between mental workload, variation in performance and physiological parameters," *IFAC-PapersOnLine*, vol. 49, no. 19, pp. 591–596, 2016.

[6] G. F. Wilson, "An analysis of mental workload in pilots during flight using multiple psychophysiological measures," *The International Journal of Aviation Psychology*, vol. 12, no. 1, pp. 3–18, Jan 2002.

[7] S. G. Hart and L. E. Staveland, *Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research*, ser. Advances in Psychology. Elsevier, 1988, vol. 52, pp. 139–183.

[8] F. Dell'Agnola *et al.*, "Physiological characterization of need for assistance in rescue missions with drones," in *IEEE International Conference on Consumer Electronics (ICCE)*, 2018.

[9] H. Mansikka *et al.*, "Dissociation between mental workload, performance, and task awareness in pilots of high performance aircraft," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 1, pp. 1–9, Feb 2019.

[10] B. Cain, "A Review of the Mental Workload Literature. Toronto," *Defence Research and Development Canada*, no. 1998, 2007.

[11] B. Ahmed *et al.*, "ReBreathe: A calibration protocol that improves stress/relax classification by relabeling deep breathing relaxation exercises," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 150–161, Apr 2016.

[12] M. Ranchet *et al.*, "Cognitive workload across the spectrum of cognitive impairments: A systematic review of physiological measures," *Neuroscience & Biobehavioral Reviews*, vol. 80, pp. 516–537, Sep 2017.

[13] J. Heard *et al.*, "A survey of workload assessment algorithms," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 434–451, Oct 2018.

[14] G. Borghini *et al.*, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neuroscience & Biobehavioral Reviews*, vol. 44, pp. 58–75, Jul 2014.

[15] N. Momeni *et al.*, "Real-time cognitive workload monitoring based on machine learning using physiological signals in rescue missions," in *41th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019.

[16] F. Dell'Agnola *et al.*, "MBioTracker: Multimodal self-aware biomonitoring wearable system for online workload detection," *IEEE Transactions on Biomedical Circuits and Systems*, 2021.

[17] C. Rognon *et al.*, "Flyjacket: An upper body soft exoskeleton for immersive drone control," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2362–2369, Jul 2018.

[18] F. Dell'Agnola *et al.*, "Cognitive workload monitoring in virtual reality based rescue missions with drones," in *Int Conf Human-Computer Interact*, vol. 12190 LNCS. Copenhagen, Denmark: Springer, Cham, jul 2020, pp. 397–409.

[19] V. Montesinos *et al.*, "Multi-modal acute stress recognition using off-the-shelf wearable devices," in *41th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019.

[20] L.-l. Chen *et al.*, "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers," *Expert Systems with Applications*, vol. 85, pp. 279–291, Nov 2017.

[21] E. T. Solovey *et al.*, "Classifying driver workload using physiological and driving performance data: two field studies," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014.

[22] D. Giakoumis *et al.*, "Subject-dependent biosignal features for increased accuracy in psychological stress detection," *International Journal of Human-Computer Studies*, vol. 71, no. 4, pp. 425–439, Apr 2013.

[23] A. Tjolleng *et al.*, "Classification of a driver's cognitive workload levels using artificial neural network on ecg signals," *Applied Ergonomics*, vol. 59, pp. 326–332, Mar 2017.

[24] M. Gjoreski *et al.*, "Monitoring stress with a wrist device using context," *Journal of Biomedical Informatics*, vol. 73, pp. 159–170, Sep 2017.

[25] F. T. Eggemeier *et al.*, "Workload assessment in multi-task environments," in *Multiple Task Performance*, 1991.

[26] J. C. Christensen *et al.*, "The effects of day-to-day variability of physiological data on operator functional state classification," *NeuroImage*, vol. 59, no. 1, pp. 57–63, 2012, neuroergonomics: The human brain in action and at work.

[27] T. Luong *et al.*, "Towards real-time recognition of users mental workload using integrated physiological sensors into a vr hmd," in *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2020, pp. 425–437.

[28] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, Jun 2005.

[29] T. Heine *et al.*, "Electrocardiographic features for the measurement of drivers' mental workload," *Applied Ergonomics*, vol. 61, pp. 31–43, May 2017.

[30] D. Novak *et al.*, "A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing," *Interacting with Computers*, vol. 24, no. 3, pp. 154–172, May 2012.

[31] L. Han *et al.*, "Detecting work-related stress with a wearable device," *Computers in Industry*, vol. 90, pp. 42–49, Sep 2017.

[32] R. J. Lysaght *et al.*, "Operator workload: Comprehensive review and evaluation of operator workload methodologies," *United States Army Research Institute for the Behavioral Sciences, Technical Report*, 1989.

[33] D. Carneiro *et al.*, "New methods for stress assessment and monitoring at the workplace," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 237–254, Apr 2019.

[34] D. Novak *et al.*, "Workload estimation in physical human-robot interaction using physiological measurements," *Interacting with Computers*, vol. 27, no. 6, 2015.

[35] B. Cinaz *et al.*, "Monitoring of mental workload levels during an everyday life office-work scenario," *Personal and Ubiquitous Computing*, vol. 17, no. 2, pp. 229–239, Feb 2013.

[36] P.-K. Jao *et al.*, "EEG correlates of difficulty levels in dynamical transitions of simulated flying and mapping tasks," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 2, pp. 99–108, 2020.

[37] A. Arza *et al.*, "Measuring acute stress response through physiological signals: towards a quantitative assessment of stress," *Medical and Biological Engineering and Computing*, pp. 1–17, Aug 2018.

[38] M. van Dooren *et al.*, "Emotional sweating across the body: comparing 16 different skin conductance measurement locations." *Physiology & behavior*, vol. 106, no. 2, pp. 298–304, may 2012.

[39] J. Pan and W. J. Tompkins, "A Reat-Time QRS Detection Algorithm," *IEEE transactions on bio-medical engineering*, vol. BME-32, no. 3, pp. 230–236, 1985.

[40] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, Mar 1996.

[41] W. H. Press and G. B. Rybicki, "Fast Algorithm for Spectral Analysis of Unevenly Sampled Data," *Astrophysical Journal, Part 1*, vol. 338, pp. 277–280, 1989.

[42] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, vol. 5, no. 258, Sep 2017.

[43] S. W. Porges, "Cardiac vagal tone: A physiological index of stress," *Neuroscience & Biobehavioral Reviews*, vol. 19, no. 2, pp. 225–233, Jun 1995.

[44] P. Grossman and E. W. Taylor, "Toward understanding respiratory sinus arrhythmia: Relations to cardiac vagal tone, evolution and biobehavioral functions," *Biological Psychology*, vol. 74, no. 2, pp. 263–285, Feb 2007.

[45] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between 2 methods of clinical measurement," *Lancet*, vol. 8476, pp. 307–310, 1986.

[46] W. Zhang *et al.*, "Task-specific normalization for continual learning of blind image quality models," 2021.

[47] S. Theodoridis *et al.*, *Introduction to Pattern Recognition*. Boston: Academic Press, 2010.

[48] H. Akoglu, "User's guide to correlation coefficients," *Turkish journal of emergency medicine*, vol. 18, no. 3, pp. 91–93, 2018.

[49] I. Guyon *et al.*, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.

[50] D. Reynolds, *Universal Background Models*. Boston, MA: Springer US, 2009, pp. 1349–1352.

[51] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, Jan 2006.

[52] MATLAB, *R2016a*. Natick, Massachusetts: The MathWorks Inc., 2016.

[53] D. Goepfert *et al.*, *Reglement Einsatzführung*. Bern: Feuerwehr Koordination Schweiz FKS, 2015.

[54] A. Mondal *et al.*, "A noise reduction technique based on nonlinear kernel function for heart sound analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 3, pp. 775–784, 2018.

[55] J. Martín *et al.*, "On the regressand noise problem: Model robustness and synergy with regression-adapted noise filters," *IEEE Access*, vol. 9, pp. 145 800–145 816, 2021.

[56] S. Ansari *et al.*, "Motion artifact suppression in impedance pneumography signal for portable monitoring of respiration: An adaptive approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 2, pp. 387–398, 2017.

[57] Q. Zhang *et al.*, "Motion artifact removal for ppg signals based on accurate fundamental frequency estimation and notch filtering," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 2965–2968.

[58] J. Miehlbradt *et al.*, "Data-driven body-machine interface for the accurate control of drones," *Proceedings of the National Academy of Sciences*, vol. 115, no. 31, pp. 7913–7918, Jul 2018.

[59] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics*, vol. 1, no. 6, pp. 80–83, 1945.

[60] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

Journal of *Imaging*

MDPI

Article

Unsupervised Domain Adaptation for Vertebrae Detection and Identification in 3D CT Volumes Using a Domain Sanity Loss

Pascal Sager [1], Sebastian Salzmann [1], Felice Burn [2,†] and Thilo Stadelmann [1,3,*,†]

[1] Centre for AI, Technikumstrasse 71, Zurich University of Applied Sciences, 8400 Winterthur, Switzerland
[2] Cantonal Hospital Aarau, AI and Data Science CoE, Tellstrasse 25, 5001 Aarau, Switzerland
[3] ECLT European Centre for Living Technology, 30123 Venice, Italy
* Correspondence: stdm@zhaw.ch; Tel.: +41-58-934-7208
† These authors contributed equally to this work.

Abstract: A variety of medical computer vision applications analyze 2D slices of computed tomography (CT) scans, whereas axial slices from the body trunk region are usually identified based on their relative position to the spine. A limitation of such systems is that either the correct slices must be extracted manually or labels of the vertebrae are required for each CT scan to develop an automated extraction system. In this paper, we propose an unsupervised domain adaptation (UDA) approach for vertebrae detection and identification based on a novel Domain Sanity Loss (DSL) function. With UDA the model's knowledge learned on a publicly available (source) data set can be transferred to the target domain without using target labels, where the target domain is defined by the specific setup (CT modality, study protocols, applied pre- and processing) at the point of use (e.g., a specific clinic with its specific CT study protocols). With our approach, a model is trained on the source and target data set in parallel. The model optimizes a supervised loss for labeled samples from the source domain and the DSL loss function based on domain-specific "sanity checks" for samples from the unlabeled target domain. Without using labels from the target domain, we are able to identify vertebra centroids with an accuracy of 72.8%. By adding only ten target labels during training the accuracy increases to 89.2%, which is on par with the current state-of-the-art for full supervised learning, while using about 20 times less labels. Thus, our model can be used to extract 2D slices from 3D CT scans on arbitrary data sets fully automatically without requiring an extensive labeling effort, contributing to the clinical adoption of medical imaging by hospitals.

Keywords: unsupervised domain adaptation; semi-supervised learning; vertebrae detection; vertebrae identification; transfer learning; semantic segmentation; data centrism; deep learning

Citation: Sager, P.; Salzmann, S.; Burn, F.; Stadelmann, T. Unsupervised Domain Adaptation for Vertebrae Detection and Identification in 3D CT Volumes Using a Domain Sanity Loss. *J. Imaging* **2022**, *8*, 222. https://doi.org/10.3390/jimaging8080222

Academic Editor: Pier Luigi Mazzeo

Received: 30 June 2022
Accepted: 12 August 2022
Published: 19 August 2022

1. Introduction

Fine-tuned AI-driven software tools allow an automated analysis of digital images and play a highly relevant role in different industries, especially in healthcare [1]. Computed tomography (CT) images provide accurate information about structural anatomy, morphology, as well as quantitative and qualitative composition of body parts [2]. They usually consist of multiple 2D slices stacked as a batch and form therefore a 3D data set. CT scan processing often relies on the feature extraction capabilities of modern deep learning architectures [3], and many modern deep learning systems process 3D scans as a whole [4,5]. An alternative to 3D scan processing is to extract representative 2D slices first [6], which, for example, can be used for preoperative surgical assessment as well as to examine metabolic, pulmonary, and neurological diseases [7,8]. Such relevant 2D slices of the upper body are usually identified based on their relation to the spine [8–10] and can either be extracted manually [10] or automatically, where automatic systems therefore need to be able to recognize the vertebrae and extract the slice containing the relevant information [11]. Usually, this requires knowledge of vertebrae locations, i.e., manually created labels for a multitude of 3D CT volumes, to train respective systems.

In this paper, we present an approach to identify vertebrae of the spine automatically without the need of excessive labeling of own data (or even no labels at all), thereby heralding a data-centric approach [12] based on un- or semi-supervised learning [13]. To this end, our contribution is the development and evaluation of a novel method that requires no labels at all to achieve reliable vertebrae detection and identification and, if given less than 5% of the labels we perform on par with comparable supervised approaches. Thus, our approach reduces the labor-intensive labeling effort that can hinder applicability in medical institutions. An overview of our approach is given in Figure 1. The quality of our results allows the extraction of representative 2D slices from 3D volumes within an automated machine learning (ML) pipeline.

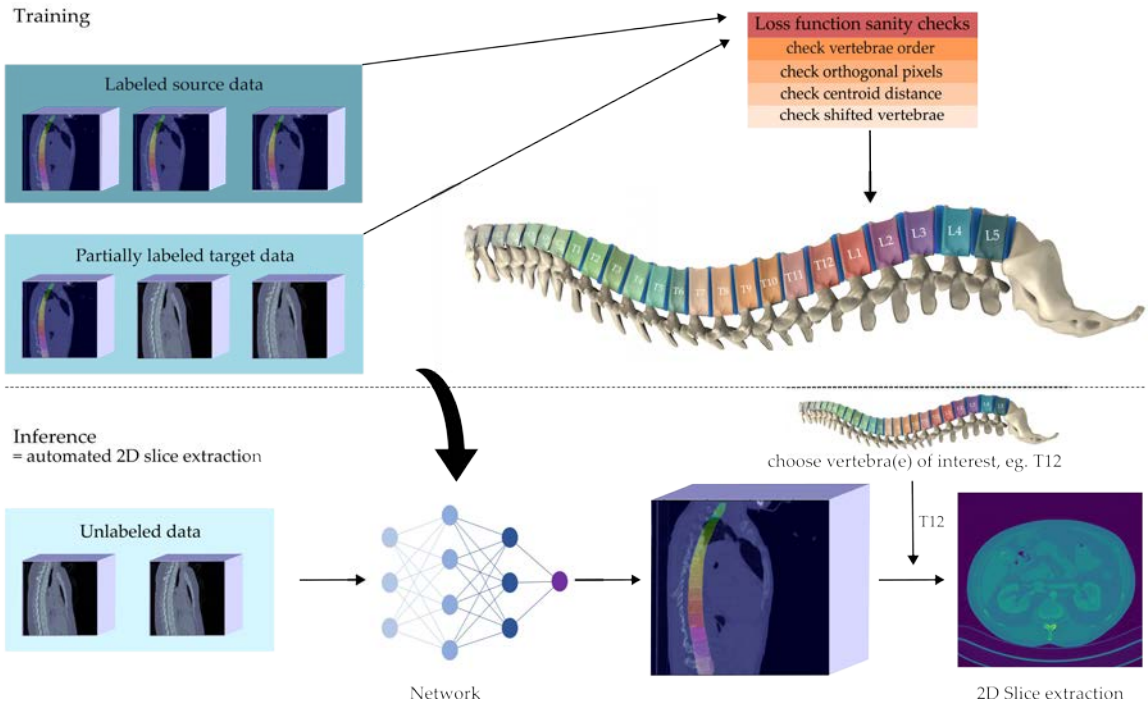


Figure 1. Visual abstract of our work: We train a vertebrae Detection and Identification module simultaneously on a publicly available data set (source domain) and a second custom data set (target domain). We require only a few labels from the custom data set. With the help of a loss function that is inspired by anatomical domain knowledge the proposed model is able to identify vertebrae centroids with state-of-the-art performance, reducing the need for target-domain labels by a factor of 20. We see its main application within ML-pipelines to extract representative 2D slices out of 3D volumes, representing a step towards fully automated systems for downstream 2D slice analysis.

The remainder of the paper is organized as follows: In Section 2, we review the related work and argue why we build upon the work of McCouat and Glocker [14]. In Section 3, we explain how we extended the "Detection" module with post-processing and propose a new unsupervised loss function for the "Identification" module. In Section 4, we present the results of our method in detail and show how well vertebrae can be detected and identified with only a few labels. In Section 5, we conclude that our method facilitates the application in medical institutions, as very good results are obtained with an order of magnitude fewer labels than comparable methods require. Furthermore, we identify limitations and suggest future research directions.

2. Related Work

The detection and identification of vertebrae is well studied. However, many methods for vertebrae identification make prior assumptions. For example, Zhou et al. [15] assume that the first sacrum vertebra (S1) is within the image while Yi et al. [16] assume that always the same vertebrae are visible. The model of Altini et al. [17] on the other hand requires manual input with meta-information about the first visible vertebra. Other approaches make assumptions about the shape of the spine [18] and therefore do not work well in pathological cases where the spine is deformed. In contrast, this work does not impose such assumptions, enabling processing of a broad range of CT scans even if the images only contain cropped parts of the spine.

Predicting the vertebra centroids directly (i.e., as a regression task) often leads to poor results [19]. Therefore most approaches turn the regression problem into a dense classification problem [14,16,19]. Earlier approaches used classical machine learning models such as random forests to identify vertebra centroids [19] while more recent approaches achieve better results using convolutional neural networks (CNNs). For example, Yang et al. [20] use an encoder-decoder architecture together with multi-level feature concatenation to locate vertebrae. The extracted centroid probability maps are iteratively improved based on the mutual relation of vertebra centroids. Liao et al. [21] achieve state-of-the-art results using a CNN to detect the positions of the centroids, combining it with a recurrent neural network (RNN) to capture the ordering of the vertebrae.

McCouat and Glocker [14] obtained similar results using two separate U-Nets [22] for detecting and identifying vertebrae. Their data set consists of 3D CT scans with labels for the vertebrae centroids. Initially, these sparse labels are converted to dense labels. Then the "Detection" module, the first in the two-stage approach, detects the spine within the 3D volume. To enable training with limited computational resources the 3D volumes are divided into smaller patches. Each of these patches is fed into a 3D U-Net that segments the vertebrae from the background. Once the spine is located the relevant region is extracted from the 3D volume and processed by the second module.

This second stage is the "Identification" module that maps pixels to the corresponding vertebrae. For this purpose, a 2D U-Net is used. The model does not classify each pixel but produces a continuous value for each pixel. Rounding this continuous value results in an integer which is associated with a vertebra (e.g., 1 = C1, 2 = C2, ...). Due to the prediction of continuous values per pixel the L1 loss function can be used to capture the order of the vertebrae. The Identification module predicts a value for each pixel, even if that pixel depicts background and not a vertebra. Since the Detection module classifies the background pixels as 0 the output of the Identification module is multiplied by the output of the Detection module yielding the prediction without background. Finally, the predicted dense labels are converted back to sparse labels by calculating their median position.

In this work, we extend this approach from McCouat and Glocker [14] with unsupervised domain adaptation (UDA) methods. We extend the Detection module with post-processing and the Identification module with a new Domain Sanity Loss (DSL) based on "sanity checks". We build upon their work for the following reasons: (i) The average distance between the predicted and the actual vertebrae centroids is small and considered state-of-the-art; (ii) the models are pure CNN architectures which can be easily extended within the framework of deep learning [23]; (iii) no assumptions are made about neither the shape of the spine nor the visible vertebrae. This way, the model is adapted to the target data, which is considerably easier to train in our experience than the alternative of adapting the data to the model [24].

3. A Method for Unsupervised Domain Adaptation of CT Scans of the Spine

The method of McCouat and Glocker [14] performs well on labeled data sets. However, performance is poor when the trained model is applied to other data sets on which it has not been trained (c.f. Section 4). To process data from other domains, we extend the two modules. The Detection module is extended with post-processing, while the Identification

model is trained with a new DSL loss function. The proposed UDA training procedure for the Identification module leverages publicly available labels and helps the model to adapt to a second data set even without labels. Since we adapt the knowledge learned on one domain to another, we refer to the first domain as the source and the second as the target domain. Our extensions only affect the training process, while the network architecture remains unchanged.

3.1. Detection Module

In accordance with [14] we divide the 3D volumes of the source and target data set into smaller patches of size $[80 \times 80 \times 96]$ and process them with a 3D U-Net. An advantage of processing patches instead of the entire 3D volume is that the model can be trained with limited computational resources. The sparse annotations (i.e., centroid positions of vertebrae) are converted into dense annotations (i.e., pixel-level labels) [14]. Pixels depicting a vertebra are labeled as 1, and pixels depicting background as 0. Adam [25] is used with a learning rate of 1×10^{-3} during training to minimize a binary cross entropy (BCE) loss. The model is trained with a batch size of 16 samples for 70 epochs. After training the model labels pixels either as spine or background. Thus, this module can locate the spine in a 3D volume.

In contrast to [14], we post-process the predictions of our model. This post-processing is helpful because it can be hard for the model to detect parts of the spine in small patches. Processing patches is considered more difficult than processing the entire CT scan because of the lack of context provided by the surrounding pixels. After all patches of a scan are predicted we conduct a connected component analysis on the 3D volume. It identifies all connected groups of pixels that are labeled as spine. Since the spine consists of many pixels, it is retained as the biggest component while smaller components are discarded as artefacts. To remove only artefacts and not the spine from the prediction we weigh the BCE loss by a factor of 1.0 for the spine and 0.1 for the background. By doing so, the spine is detected as a single component with very high accuracy and not removed as an artefact.

3.2. Identification Module and Domain Sanity Loss

The Identification module processes patches of the size $[8 \times 80 \times 320]$ in a 2D U-Net as in [14]. These patches have a large field of view of 80×320 pixels along the sagittal plane thus allowing identification of vertebrae. As conducted in the Detection module, the sparse annotations are converted to dense annotations, background is labeled as 0 and the vertebrae with integers in ascending order (i.e., 1 = C1, 2 = C2, . . . , 26 = S2).

In contrast to [14], we extend this module with an UDA method. Our proposition is based on a novel training process that instead of processing only samples from the source domain is alternatingly feeding mini-batches from the source and target domain into the model. The intuition behind this is that samples from the source domain teach the model vertebrae identification while samples from the target domain help to adapt to the target data set. This 2-way training procedure is shown in Figure 2.

In the first phase, since the source data samples have labels, a supervised L1 loss function is used as suggested by [14]. By predicting continuous values and not label probabilities, this function is able to measure the distance to the ground truth vector rather than merely checking for equality (e.g., prediction C2 is better than prediction C3 for label C1) and thus considers the order of the vertebrae. However, since no labels are available for the target data set no supervised loss function can be used in the second phase. Therefore, we propose the Domain Sanity Loss (DSL) based on “sanity checks” as introduced and illustrated in Figure 3.

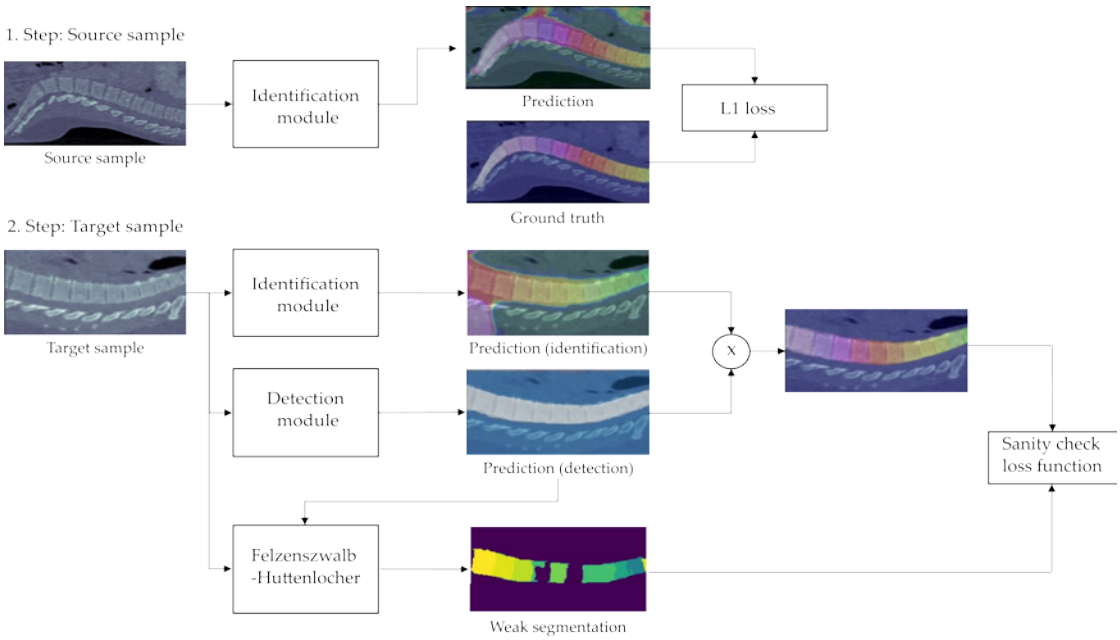


Figure 2. 2-way training process of the Detection module: In step one, L1 distance is used to calculate the loss of a mini-batch of source domain samples. In step two, several "sanity checks" (see Figure 3 for an overview) are calculated to form the loss of a mini-batch of target-domain data. The sanity-check-based DSL loss only considers spine pixels by multiplying the output of the Identification module with the output of the Detection module and employs the Felzenszwalb-Huttenlocher algorithm [26] to create a weak segmentation mask of vertebrae location in an unsupervised way (c.f. Section 3.2).

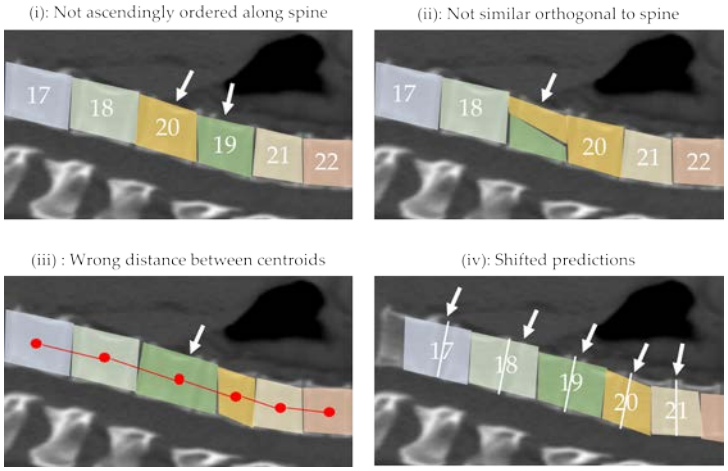


Figure 3. Visual representation of the sanity checks performed by the proposed Domain Sanity Loss (DSL) function; the displayed cases show failures for each check, indicated by the white arrows. Specifically, the DSL loss checks for (**i**) monotonous ascend of predicted vertebrae numbers along the spine; (**ii**) all spine pixels in one column of the image having the same vertebra number; (**iii**) predicted vertebrae centroids having a reasonable distance to each other, based on average distances from the literature; and (**iv**) predictions not being shifted along the spine, based on an unsupervised weak segmentation of the vertebrae (c.f. Figure 2).

The DSL loss is with its four checks purely based on anatomically induced invariances that hold true even for severely deformed spines and hence need no corresponding human-provided labels for any image. As these invariances only apply to pixels belonging to the spine, we multiply the model output with the prediction of the previous Detection module and thereby set all pixel values that do not belong to the spine to zero. In the following, we denote this prediction with removed background as $\hat{y}$: a matrix of the same shape as an input image with the predicted vertebra number for spinal pixels (i.e., 1 = C1, 2 = C2, ...) and 0 otherwise. We denote i as row and j as column indices of $\hat{y}$ and n_{row} and n_{col} as the number of pixels per row and column of the sagittal plane respectively. Furthermore, we define the identification function for boolean values as

$$\mathbb{1}_b(x) = \begin{cases} 1 & \text{if } x \text{ is true} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The first term s_1 of the DSL loss function (c.f. Equation (6) at the end of this subsection) evaluates whether the *vertebrae are sorted in ascending order along the spine*. For a correct prediction, the per-pixel values in $\hat{y}$ along the longitudinal axis must be sorted in ascending order (c.f. Figure 3(i)). We implement this by comparing each predicted pixel $\hat{y}_{i,j}$ with a version of the same prediction $\hat{y}_{i,j+s}$ shifted to the right by s pixels. Thereby we evaluate if a pixel shifted to the right of any given pixel still gets the same or a higher prediction. In doing so, we check whether pixels are sorted ascending from the left to the right. All pixels that do not fulfill this criterion lead to an increase in the loss value. We ignore the pixel values that get shifted outside of the range of the original prediction which is why we only sum up the pixels column wise until n_{col-s}. We define the first loss term as

$$s_1(\hat{y}) = \frac{1}{n_{pix}} \sum_{s=1}^{n_{shift}} \sum_{i=1}^{n_{row}} \sum_{j=1}^{n_{col}-s} \mathbb{1}_b(\hat{y}_{i,j} - \hat{y}_{i,j+s} \geq 0) \tag{2}$$

where n_{shift} is the maximum range of shift, and n_{pix} the number of pixels in $\hat{y}_{i,j}$. Empirically, we found that shifting values $s > 30$ do not enhance the result anymore. We therefore define $n_{shift} = 30$ and thus compare the order of the vertebrae only locally which leads to higher computational efficiency. We divide the number of pixels that violate this constraint by the number of total pixels n_{pix} and therefore $s_1(\hat{y})$ captures the percentage of spinal pixel for which the anatomical order of the vertebrae is not correct.

The second term s_2 of the loss function checks whether the *pixel values orthogonal to the spine are identical*. For this we analyze the pixels that are differently labeled along the sagittal axis (c.f. Figure 3(ii)). We assume the median value of each column j of $\hat{y}_{i,j}$ as label of that column and compare it to all values in that column. We denote v_j as the column vector of $\hat{y}_{i,j}$ at index j. Furthermore, we define a function $median(v_j)$ which calculates the median of a column vector v_j. We assume that the spine is more or less parallel to it (rotation can be checked easily by pre-processing, if necessary). We define the second loss term as

$$s_2(\hat{y}) = \frac{1}{n_{pix}} \sum_{i=1}^{n_{row}} \sum_{j=1}^{n_{col}} \mathbb{1}_b(|\hat{y}_{i,j} - median(v_j)| > 0) \tag{3}$$

For each column, we sum up the number of pixels that are labeled differently than the median and divide this sum by the total number of pixels. Thereby we obtain a factor that indicates how consistent the vertebrae per column and thus orthogonal to the spine are.

The third term s_3 of the DSL loss function evaluates the *distance between the centroids of the predicted vertebrae* (c.f. Figure 3(iii)). We define the distance between vertebra i and j as $\delta_{i,j}$. We denote the average distances of vertebrae as taken from Busscher et al. [27] as $\bar{\delta}_{i,j}$. We denote the upper bound of the summation as $n_{vert} = 25$, which is the number of vertebrae of a spine (26) minus one. The third loss term

$$s_3(\hat{y}) = \frac{1}{n_{vert}} \sum_{i=1}^{n_{vert}} |\delta_{i,i+1} - \bar{\delta}_{i,i+1}| \tag{4}$$

calculates the Euclidean distances between subsequent vertebra centroids and compares it to the gold standard from literature using the L1 loss. If the distance between two vertebrae is equal to the gold standard the loss is 0, otherwise it is bigger than 0. We sum up the distance differences between subsequent vertebrae to the third term $s_3(\hat{y})$. We therefore perform an explicit sanity check on vertebrae distance and an implicit check on vertebrae size.

The fourth term s_4 of the loss function checks *whether the predicted vertebrae are not shifted*. So far it has only been verified whether the spine is anatomically correctly detected. However, the spine itself may be slightly displaced within the image (c.f. Figure 3(iv)). To detect shifts we make use of a weak segmentation mask which is constructed as follows: First, the input scan (and not the mask) is multiplied by the prediction of the Detection module to extract the spine, followed by setting all pixels below an intensity threshold of 180HU to 0 in order to emphasize the edges. We then use the Felzenszwalb-Huttenlocher algorithm [26] to predict a segmentation mask of the vertebrae in a unsupervised manner. As this mask is relatively imprecise it is referred to as a weak mask wm. The predicted mask is further improved by heuristically filtering out components that cannot correspond to a vertebra (e.g., wrong shape) and by merging components that are enclosed in one another.

The weak mask wm has the same shape as the prediction $\hat{y}$. Each pixel in the weak mask is assigned to a connected component $c_k \in wm$. Each c_k has a set of row c_{k_i} and column c_{k_j} coordinates which pairwise represent all pixels of a component. The intuition behind this fourth loss term is that the prediction $\hat{y}$ should have the same label at the coordinates of pixels that belong to the same connected component c_k. For each connected component c_k we extract from $\hat{y}_{i,j}$ the values at the positions $(i,j) \in (c_{k_i}, c_{k_j})$ and define this operation as $v(\hat{y}, c_k)$. Furthermore, we define $u(x)$ which returns the number of unique values in a set x. Based on our definition $u(v(\hat{y}, c_k))$ returns the number of unique values within $\hat{y}_{i,j}$ at the coordinates (c_{k_i}, c_{k_j}) of a connected component c_k.

Per connected component c_k the pixels in the prediction $\hat{y}$ should be labeled identically and thus $u(v(\hat{y}, c_k))$ should return 1. If multiple labels are predicted at the positions of a connected component, $u(v(\hat{y}, c_k))$ returns a value greater than 1. The fourth part of our DSL loss function sums up the number of inconsistent labels per connected component:

$$s_4(m, \hat{y}) = \frac{1}{n_c} \sum_{c_k \in wm} u(v(\hat{y}, c_k)) - 1 \tag{5}$$

The domain-specific DSL loss function therefore consists of four sanity checks that penalize anatomical inconsistencies. To obtain the DSL loss value, we sum the four loss terms:

$$L(m, \hat{y}) = c_1 \cdot s1(\hat{y}) + c_2 \cdot s2(\hat{y}) + c_3 \cdot s3(\hat{y}) + c_4 \cdot s4(m, \hat{y}) \tag{6}$$

where the constants c_s are scaling values that we found experimentally to work well when set to $c_1 = 20$, $c_2 = 1$, $c_3 = 1/40$, and $c_4 = 1/100$ as they bring the four loss parts to an approximately similar scale. To optimize this loss, we use Adam [25] as optimizer with a learning rate of 5×10^{-4}. The model is trained for 100 epochs with a batch size of 32 samples.

3.3. Data Sets

We use the BioMedIA Spine data set [28] as source data set. It consists of 242 spine-focused CT-scans of 125 patients with varying types of pathologies. In most scans, the view is limited to 5–15 vertebrae, while only a few scans depict the entire spine [19]. The scans differ significantly in terms of image noise, physical resolution, and vertical cropping [18]. Each scan is labeled with point-annotations of vertebrae centroids that are extended to

dense labels using the approach outlined in [14]. The data set provides a predefined split which is used for training and testing.

To test the proposed unsupervised domain adaptation schema for vertebrae detection and identification, the COVID19-CT data set [29,30] with 1000+ scans from patients with confirmed COVID-19 diagnosis is used. The scans are composed of 16-bit grayscale images with a size of 512×512 pixels [29]. Most of the scans have an inter-axial distance between 0.5 and 1.5 mm. A radiology experienced physician labeled the vertebra centroids of a random subset with 30 scans, of which 20 are used as a test set and 10 labeled scans optionally together with the remaining scans as training set.

Similar to [14], we divide all samples into smaller patches. To train the Detection module on the source data set we extract 10 patches of the size $[80 \times 80 \times 96]$ from random positions out of each scan. Thereby we ensure that at least 8 out of the 10 patches contain parts of the spine. Since the Detection module is not trained on the target data set, only patches from the labeled source data set are needed. For testing on the source as well as the target data set, we divide the entire scan independent of the position of the spine into patches of the size $[80 \times 80 \times 96]$.

For the training of the Identification module, we extract 300 patches with a shape of $[8 \times 80 \times 320]$ per sample. If labels exist we ensure that each patch contains at least one vertebra. If no labels exist we use the output of the Detection module to locate the spine and extract patches out of this region. For testing, the entire scan is again divided into patches

4. Results

In the following three subsections, we analyze our Detection and Identification module experimentally, comparing them to prior and related work.

4.1. Detection Results with and without Post-Processing

The Detection module detects the spine within the 3D volume well. However, without post-processing many false-positive predictions (i.e., prediction "spine" instead of "background") lead to bad results, especially on the target data set without labels (c.f. Figure 4). A possible reason for this is that the model is trained only on small patches of the original volume. Therefore, the model only learns to identify parts of vertebrae and not how a whole spine looks like. An indication for this is that false negatives are often detected in places with cubic shapes, for example, the bed on which the patient is lying. However since these false predictions consist of far fewer connected pixels than the entire spine our post-processing is able to successfully remove these artefacts.

To highlight how our post-processing improves performance we calculate various metrics. However, these metrics must be interpreted with caution for two reasons: (i) Generated dense annotations, which are calculated based on average sizes of vertebrae bodies are used as ground truth; such annotations are by design less accurate than, for example carefully hand-crafted segmentation masks. (ii) The performance is calculated on the whole volume and not on cropped samples as is conducted in [14]. Since the cropped samples have a much higher proportion of pixels representing the spine these results are not directly comparable. However, the published results of [14] correspond roughly with the performance of our model without post-processing as both are based on the same method

Table 1 shows the results of the Detection module. The accuracy, recall, intersection over union (IoU), and dice-score are calculated for the source data set (BioMedIA) and for the target data set (COVID19-CT). The proposed post-processing clearly improves the performance. For example, the IoU of pixels representing vertebrae in the source data set improves from 67.4% to 78.7%, which is a 16.8% relative improvement. The improvement on the target data set is even more noticeable. Using post-processing on the target data set, IoU improves from 46.4% to 79.1%. While the model without post-processing is considered not accurate enough to detect the vertebrae on the target data set, the model with post-processing is suitable for the subsequent vertebrae identification.

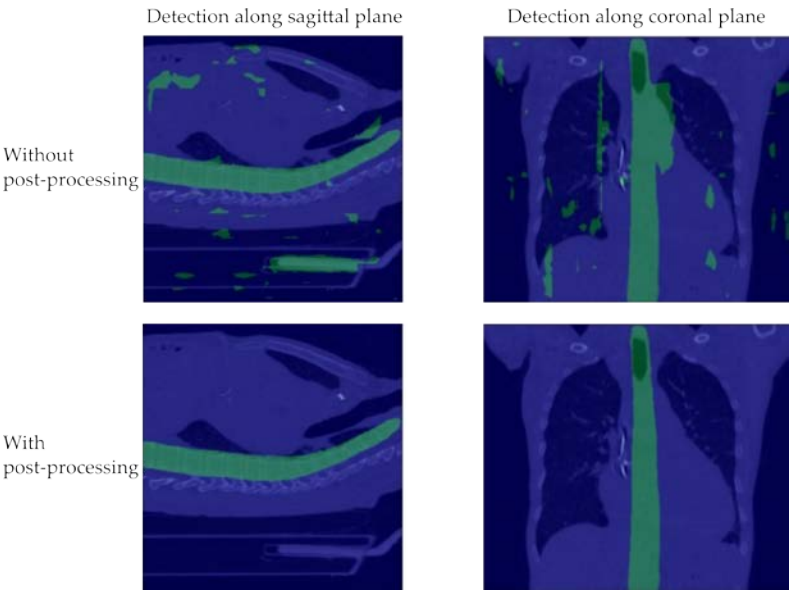


Figure 4. Four randomly selected samples from the target data set (COVID-19 CT) with overlayed predictions for the spine detection with (**bottom row**) and without (**top row**) post-processing. To provide a better grasp of the post-processing's effect, we visualize all predictions within the 3D mask along the sagittal plane (**left**) and along the coronal plane (**right**).

Table 1. Performance evaluation of the Detection module with the highest score for each metric and data set in bold. For each metric, the overall performance for the whole 3D scan and for the vertebrae with ignored background is reported. The positive effect of our proposed the post-processing is visible on the source and the target data sets.

Metric	**BioMedIA (Source Data Set) without Post-Processing**	**with Post-Processing**
Accuracy (overall)	99.2%	**99.5**%
Recall (overall/vertebrae)	99.2%/**94.3%**	**99.5%**/94.1%
IoU (overall/vertebrae)	98.3%/67.4%	**99.0%/78.7%**
Dice (overall/vertebrae)	99.2%/80.2%	**99.5%/88.0%**
Metric	**COVID-19 CT (Target Data Set) without Post-Processing**	**with Post-Processing**
Accuracy (overall)	99.6%	**99.9**%
Recall (overall/vertebrae)	99.6%/**95.1%**	**99.9%/95.1%**
IoU (overall/vertebrae)	99.2%/46.4%	**99.8%/79.1%**
Dice (overall/vertebrae)	99.6%/63.0%	**99.9%/88.0%**

4.2. Identification Results per Spinal Pixel

We trained the Identification module in three different setups: (i) A first model is trained without UDA and only using source labels, corresponding to the same method as proposed in [14]; (ii) a second model is optimized with the proposed DSL loss of Section 3.2; (iii) a third model is given ten random training samples plus their labels from the target data set, used in the same fashion as source samples.

To compare the models with and without UDA, the classification accuracy per pixel is measured. The accuracy is determined by dividing the number of correctly classified pixels by the total number of pixels. Thereby, only the pixels belonging to the spine are taken into account and the background is ignored. As shown in Table 2, the model without UDA (i) achieves a classification rate of 13.3% on the target data set. The model with UDA (ii)

achieves an accuracy of 61.4%. This corresponds to a relative improvement of 462.7% and demonstrates the effectiveness of the proposed approach. If additionally ten samples from the target data set are labeled (iii), the identification rate further improves to 74.2%. We display some predictions in Figure 5. This visualization demonstrates that the vertebrae are well recognized.

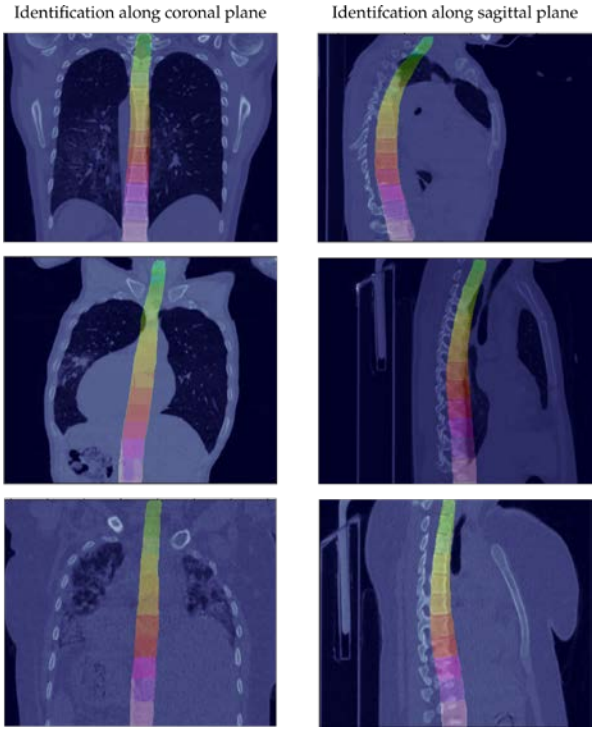


Figure 5. Random samples of prediction from the Identification module on the target data set (COVID-19 CT), showing satisfactory results even when the spine is not well aligned on the coronal and sagittal axis.

Table 2. Classification rate on the COVID19-CT data set for the three trained models with the best classification rate in bold. The effectiveness of un- and semi-supervised domain adaptation is striking.

Classification Rate on COVID-19 CT (Target Data Set)		
Our Method without UDA	**Our Method**	**Our Method (with 10 Labels)**
13.3%	61.4%	**74.2%**

4.3. Identification Results per Vertebra

The results described so far refer to the classification accuracy per pixel. However, the goal is to identify the vertebra centroids and therefore the obtained dense predictions must be converted back into sparse centroid predictions. This is conducted by calculating the median of the dense predictions as described by [14], thereby ignoring outliers in the pixel-level prediction by virtue of the median. The results of the centroid predictions are shown in Table 3. We define the identification rate "ID" as the number of correctly identified vertebrae divided by the total number of vertebrae. We count an identification as correct if the predicted centroid is no more than 20 mm away from the ground truth. We use 20 mm as this is an often used reference distance [14,18,21] and therefore makes our method comparable to other approaches. Only the results on thoracic vertebrae are compared since vertebrae from other regions are underrepresented in the COVID-19 CT

data set (CT scans can be classified into regions depending on the body part they are taken from. Well-known areas are the cervical region (neck level), the thoracic region (chest level) and the lumbar region (pelvis level). For state-of-the-art AI models the thoracic region is the most challenging one because only a middle section of the spine is visible in these scans and therefore vertebrae cannot be counted from the first cervical vertebra (C1), respectively the last sacrum vertebra (S2)).

Table 3. Detection result per vertebra with the best score for each metric and data set in bold. The upper part of the table displays the results on thoracic scans of the source data set, the lower part the results on the target data set. The column "ID" gives the identification rate, column "Mean" reports the average distance to the ground truth centroid in mm and column "Std" gives the standard deviation in mm.

Thoracic Vertebrae BioMedIA (Source Data Set)			
Method	**ID**	**Mean**	**Std**
Chen et al. [31]	76.4%	11.4 mm	16.5 mm
Liao et al. [21]	**84.0%**	7.8 mm	10.2 mm
McCouat and Glocker [14]	79.8%	6.6 mm	7.4 mm
Our method	67.0%	8.4 mm	8.7 mm
Our method (with 10 labels)	80.1%	**6.2 mm**	**7.2 mm**
Thoracic Vertebrae COVID-19 CT (Target Data Set)			
Method	**ID**	**Mean**	**Std**
Our method without UDA	45.6%	17.4 mm	24.2 mm
Our method	72.8%	11.1 mm	20.8 mm
Our method (with 10 labels)	**89.2%**	**8.1 mm**	**20.3 mm**

As before, "our method" corresponds to the model proposed in [14] with additional UDA extensions. The results obtained with this model on the BioMedIA source data set are less accurate than those of the original model without UDA. A reason is that our model was optimized for the target data set only. Furthermore, by using domain adaptation a performance loss on the source data set was consciously accepted in exchange for better results on the target data set. If ten labels from the target data set are added during training the model is superior to the original one on the source data set. Reasons for this are that (i) the post-processing of the Detection module leads to better identification of the spine and (ii) that the COVID-19 CT data set contains a lot of samples from thoracic vertebrae and thus the model is more optimized for this region.

When analyzing the results on the COVID19-CT data set the effectiveness of the proposed domain adaptation is evident. When the model is trained without UDA, only 45.6% of the vertebrae are correctly classified on the target data set. With the proposed domain adaptation methods, the classification rate increases to 72.8%. A comparison with state-of-the-art results on the BioMedIA data set shows (though being unfair because of the different data sets used to achieve the respective numbers) that this is only 11.2 pp. less accurate than the results of Liao et al. [21] and only 7 pp. less accurate than the results of McCouat and Glocker [14], which both trained their model with labels. If ten labeled target samples are added to the training set, an identification rate of 89.2% is achieved. This is 5.2 pp. better than the best results reported so far for the BioMedIA data set. Of course, the comparability of these remarks is limited because the data sets are different, but it underlines that the performance of our method with semi-supervised domain adaptation is remarkable.

5. Conclusions

In this paper, we presented a method to find vertebrae centroids on unlabeled CT data sets, proposing a novel un- and semi-supervised domain adaptation method based on the Domain Sanity Loss function that achieves state-of-the-art results with orders of magnitudes less labels than previous methods. The detection and identification of

vertebrae is important, for example, to extract 2D slices at predefined levels from 3D CT scans. Compared to existing state-of-the-art systems our method has the advantage of requiring much fewer labels while obtaining comparable results. For example, in clinical practice, the BioMedIA [28] data set could be used as source data set and be combined with a custom target data set. Our proposed UDA approach would only require the creation of approximately ten labels of the custom data set, whereas a supervised approach might require several hundred labels. Since less labor-intensive labeling is necessary the transfer of the method to other medical applications and facilities is easier and more cost-efficient.

The main drawback of our method is that it requires more computational resources. While supervised methods use one data set, our UDA method requires a source and a target data set. Using an NVIDIA V100 GPU, training takes about 2 days. However, comparable results with an ID rate of over 86% can be achieved after 16.5 h (with 35 instead of 100 epochs). Thus, the training takes slightly more than twice as long as the original method from McCouat and Glocker [14]. Inference, on the other hand, is identical except for the additional post-processing and therefore takes about the same amount of time.

5.1. Discussion

Specifically, pixel-level classification is often employed in the medical field [4–6]. Training such models in a supervised manner requires labels. Depending on the specific task, labeling a single 3D scan on the pixel level can take an expert up to two weeks [32]. Considering that many applications require several hundred samples, one can conclude that labeling a complete data set is almost prohibitively labor-intensive [33,34], setting harsh limits to AI democratization. Alternatively, representative 2D slices can be used for various applications (c.f. Section 1). These 2D slices are less time-consuming to label, since they are only a cut-out of the 3D data. Thus, not only does our method for extracting 2D slices require very few labels, but it can reduce the labeling effort of downstream ML pipelines because representative 2D slices instead of 3D data can be processed in subsequent systems.

On the COVID-19 CT data set, 89.2% of all vertebra centroid predictions are identified correctly which is in line with (in fact, beyond) the state-of-the-art on other data sets. The mean deviation of the predicted centroid to the ground truth centroid is 8.1 mm. However, this distance is measured in the 3D space. Considering the task of extracting 2D slices the deviation is even smaller because only the error in one direction of the 3D space is relevant. In rare cases vertebrae can be mistaken and the deviation is much bigger leading to a standard deviation of 20.3 mm. Depending on the application, such wrong predictions can simply be filtered out by analyzing the content of the 2D slice as conducted by [11]. However, since this is application-dependent such post-processing is out of scope of our work.

5.2. Limitations and Future Work

The proposed UDA method with DSL loss works very well on our target data set. A limitation, however, is that the fourth loss component s_4 relies on reference distances between subsequent vertebrae from the literature. Therefore, it is assumed that our approach works worse for patients which do not comply with these reference values (e.g., children). A second limitation is uncommon spinal constellations. In very rare cases, for example, patients may have an additional lumbar vertebra L6, a lumbalizated S1, or a sacralizated L5 as normal deviations to the standard spine. Since these constellations are not included in our label set, they therefore cannot be detected.

In principle, our proposed UDA method and a DSL loss based on domain-specific sanity checks is applicable to other domains and problems as well, even outside of medical image processing. For example, we started experimenting with DSL losses for symbol recognition in document analysis tasks [35]: We calculated statistics of symbols such as their size and orientation, and built DSL losses to ensure that the predictions per page comply with these statistics. From the preliminary experiments, we learned that DSL losses will not work well if the data contains a lot of variation which cannot be specified in the loss

function. Furthermore, we found that in this use-case a pre-training is necessary, otherwise the predictions deviate too much from the statistics which hinders the learning process.

With respect to this work, we see further research potential (i) on optimizing performance for patients with a smaller spine and (ii) on reliably detecting and correcting incorrect predictions. The issues for patients with a small spine could be remedied either by using other reference values or by adapting the loss component s_4 to work with ratios instead of absolute distances. Incorrect predictions, on the other hand, could be detected with statistical methods regarding the centroids or by analyzing the corresponding 2D slice on the transversal plane.

On a more general perspective, the DSL loss is considered complementary to process unlabeled data and could serve as a general domain adaptation method. For example, specifying a framework that derives statistics about sizes and relations of objects from the data set and uses them as sanity checks in the loss function could be helpful for various applications.

Author Contributions: Conceptualization, P.S., F.B. and T.S.; methodology, P.S.; software, P.S.; validation, P.S., F.B. and S.S.; formal analysis, P.S.; data curation, F.B.; writing—original draft preparation, P.S.; writing—review and editing, S.S., F.B. and T.S.; visualization, P.S. and S.S.; supervision, F.B. and T.S.; project administration, F.B. and T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and good clinical research practise. An ethical review and approval were waived for this study due the public and free available data set that was used in all methods. The Harvard data set [29,30] is hereby attributed for its contribution in this work. The BioMedIA data set [18,19] is hereby attributed for its contribution in this work. Its data has been provided by the Department of Radiology at University of Washington (http://www.rad.washington.edu/) and is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Informed Consent Statement: Not applicable—Patient consent was waived due to fact that the data set was publicly free available for applied medical research.

Data Availability Statement: We publish our code on https://github.com/sagerpascal/uda-vertebrae-identification (accessed on 17 August 2022) together with a detailed description of how the data sets as well as the annotations can be accessed.

Acknowledgments: The authors are grateful for support with expertise, clinical experience and annotations by the Cantonal Hospital of Aarau.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Esmaeilzadeh, P. Use of AI-based tools for healthcare purposes: A survey study from consumer's perspectives. *BMC Med Informatics Decis. Mak.* **2020**, *20*, 170. [CrossRef] [PubMed]
2. Paris, M. Body Composition Analysis of Computed Tomography Scans in Clinical Populations: The Role of Deep Learning. *Lifestyle Genom.* **2019**, *13*, 1–4. [CrossRef] [PubMed]
3. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef]
4. Koitka, S.; Kroll, L.; Malamutmann, E.; Oezcelik, A.; Nensa, F. Fully automated body composition analysis in routine CT imaging using 3D semantic segmentation convolutional neural networks. *Eur. Radiol.* **2021**, *31*, 1795–1804. doi: 10.1007/s00330-020-07147-3. [CrossRef] [PubMed]
5. Fu, Y.; Ippolito, J.E.; Ludwig, D.R.; Nizamuddin, R.; Li, H.H.; Yang, D. Automatic segmentation of CT images for ventral body composition analysis. *Med. Phys.* **2020**, *47*, 5723–5730. [CrossRef]
6. Weston, A.D.; Korfiatis, P.; Kline, T.L.; Philbrick, K.A.; Kostandy, P.; Sakinis, T.; Sugimoto, M.; Takahashi, N.; Erickson, B.J. Automated Abdominal Segmentation of CT Scans for Body Composition Analysis Using Deep Learning. *Radiology* **2019**, *290*, 669–679. [CrossRef]

7. Schweitzer, L.; Geisler, C.; Pourhassan, M.; Braun, W.; Glüer, C.C.; Bosy-Westphal, A.; Müller, M.J. Estimation of Skeletal Muscle Mass and Visceral Adipose Tissue Volume by a Single Magnetic Resonance Imaging Slice in Healthy Elderly Adults. *J. Nutr.* **2016**, *146*, 2143–2148. [CrossRef]
8. Tolonen, A.; Pakarinen, T.; Sassi, A.; Kyttä, J.; Cancino, W.; Rinta-Kiikka, I.; Pertuz, S.; Arponen, O. Methodology, clinical applications, and future directions of body composition analysis using computed tomography (CT) images: A review. *Eur. J. Radiol.* **2021**, *145*, 109943. [CrossRef]
9. Shen, W.; Punyanitya, M.; Wang, Z.; Gallagher, D.; St-Onge, M.P.; Albu, J.; Heymsfield, S.B.; Heshka, S. Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image. *J. Appl. Physiol.* **2004**, *97*, 2333–2338. [CrossRef]
10. Popuri, K.; Cobzas, D.; Esfandiari, N.; Baracos, V.; Jagersand, M. Body Composition Assessment in Axial CT Images Using FEM-Based Automatic Segmentation of Skeletal Muscle. *IEEE Trans. Med Imaging* **2016**, *35*, 512–520. [CrossRef]
11. Nowak, S.; Theis, M.; Wichtmann, B.D.; Faron, A.; Froelich, M.F.; Tollens, F.; Geißler, H.L.; Block, W.; Luetkens, J.A.; Attenberger, U.I.; et al. End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT. *Eur. Radiol.* **2021**, *32*, 3142–3151. [CrossRef]
12. Stadelmann, T.; Klamt, T.; Merkt, P.H. Data Centrism and the Core of Data Science as a Scientific Discipline. *Arch. Data Sci. Ser. A (Online First)* **2022**, *8*, 1–16. [CrossRef]
13. Simmler, N.; Sager, P.; Andermatt, P.; Chavarriaga, R.; Schilling, F.P.; Rosenthal, M.; Stadelmann, T. A Survey of Un-, Weakly-, and Semi-Supervised Learning Methods for Noisy, Missing and Partial Labels in Industrial Vision Applications. In Proceedings of the 8th Swiss Conference on Data Science (SDS), Lucerne, Switzerland, 9 June 2021; pp. 26–31, ISBN 978-1-6654-3874-2. [CrossRef]
14. McCouat, J.; Glocker, B. Vertebrae Detection and Localization in CT with Two-Stage CNNs and Dense Annotations. *arXiv* **2019**. arXiv:1910.05911. [CrossRef]
15. Zhou, Y.; Liu, Y.; Chen, Q.; Gu, G.; Sui, X. Automatic Lumbar MRI Detection and Identification Based on Deep Learning. *J. Digit. Imaging* **2019**, *32*, 513–520. [CrossRef] [PubMed]
16. Yi, J.; Wu, P.; Huang, Q.; Qu, H.; Metaxas, D.N. Vertebra-Focused Landmark Detection for Scoliosis Assessment. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; pp. 736–740, ISBN 978-1-5386-9330-8. [CrossRef]
17. Altini, N.; De Giosa, G.; Fragasso, N.; Coscia, C.; Sibilano, E.; Prencipe, B.; Hussain, S.M.; Brunetti, A.; Buongiorno, D.; Guerriero, A.; et al. Segmentation and Identification of Vertebrae in CT Scans Using CNN, k-Means Clustering and k-NN. *Informatics* **2021**, *8*, 40. [CrossRef]
18. Glocker, B.; Feulner, J.; Criminisi, A.; Haynor, D.R.; Konukoglu, E. Automatic Localization and Identification of Vertebrae in Arbitrary Field-of-View CT Scans. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI, Nice, France, 1–5 October 2012; Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., et al., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7512, pp. 590–598, ISBN 978-3-642-33454-2. [CrossRef]
19. Glocker, B.; Zikic, D.; Konukoglu, E.; Haynor, D.R.; Criminisi, A. Vertebrae Localization in Pathological Spine CT via Dense Classification from Sparse Annotations. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI, Nagoya, Japan, 22–26 September 2013; Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., et al., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7908, pp. 262–270, ISBN 978-3-642-40763-5.[CrossRef]
20. Yang, D.; Xiong, T.; Xu, D.; Huang, Q.; Liu, D.; Zhou, S.K.; Xu, Z.; Park, J.; Chen, M.; Tran, T.D.; et al. Automatic Vertebra Labeling in Large-Scale 3D CT using Deep Image-to-Image Network with Message Passing and Sparsity Regularization. In Proceedings of the International Conference on Information Processing in Medical Imaging, Boone, NC, USA, 25–30 June 2017; Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.T., Shen, D., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 633–644, ISBN 978-3-319-59050-9. [CrossRef]
21. Liao, H.; Mesfin, A.; Luo, J. Joint Vertebrae Identification and Localization in Spinal CT Images by Combining Short- and Long-Range Contextual Information. *IEEE Trans. Med. Imaging* **2018**, *37*, 1266–1275. [CrossRef]
22. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI, Munich, Germany, 5–9 October 2015; Navab, N.; Hornegger, J.; Wells, W.M.; Frangi, A.F., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241, ISBN 978-3-319-24574-4. [CrossRef]
23. Stadelmann, T.; Tolkachev, V.; Sick, B.; Stampfli, J.; Dürr, O. Beyond ImageNet: deep learning in industrial practice. In *Applied Data Science: Lessons Learned for the Data-Driven Business*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 205–232, ISBN 978-3-030-118200-4. [CrossRef]
24. Amirian, M.; Montoya-Zegarra, J.A.; Gruss, J.; Stebler, Y.D.; Bozkir, A.S.; Calandri, M.; Schwenker, F.; Stadelmann, T. PrepNet: A Convolutional Auto-Encoder to Homogenize CT Scans for Cross-Dataset Medical Image Analysis. In Proceedings of the 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, China, 23–25 October 2021; pp. 1–7, ISBN 978-1-6654-0004-6. [CrossRef]
25. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980. [CrossRef]

26. Felzenszwalb, P.F.; Huttenlocher, D.P. Efficient Graph-Based Image Segmentation. *Int. J. Comput. Vis.* **2004**, *59*, 167–181. [CrossRef]
27. Busscher, I.; Ploegmakers, J.J.W.; Verkerke, G.J.; Veldhuizen, A.G. Comparative anatomical dimensions of the complete human and porcine spine. *Eur. Spine J.* **2010**, *19*, 1104–1114. [CrossRef]
28. Biomedical Image Analysis Group, Imperial College London. BioMedIA Spine Dataset. Available online: https://biomedia.doc.ic.ac.uk/data/spine/ (accessed on 15 November 2021).
29. Shakouri, S.; Bakhshali, M.A.; Layegh, P.; Kiani, B.; Masoumi, F.; Ataei Nakhaei, S.; Mostafavi, S.M. COVID19-CT-dataset: An open-access chest CT image repository of 1000+ patients with confirmed COVID-19 diagnosis. *BMC Res. Notes* **2021**, *14*, 1–3. [CrossRef]
30. Mostafavi, S.M. COVID19-CT-Dataset: An Open-Access Chest CT Image Repository of 1000+ Patients with Confirmed COVID-19 Diagnosis. *BMC Res. Notes* **2021** *14*, 178. [CrossRef]
31. Chen, H.; Shen, C.; Qin, J.; Ni, D.; Shi, L.; Cheng, J.C.Y.; Heng, P.A. Automatic Localization and Identification of Vertebrae in Spine CT via a Joint Learning Model with Deep Neural Networks. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI, Munich, Germany, 5–9 October 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 9349, pp. 515–522. ISBN 978-3-319-24553-9, [CrossRef]
32. Ma, D.; Chow, V.; Popuri, K.; Beg, M.F. Comprehensive Validation of Automated Whole Body Skeletal Muscle, Adipose Tissue, and Bone Segmentation from 3D CT images for Body Composition Analysis: Towards Extended Body Composition. *arXiv* **2021**, arXiv:2106.00652. https://doi.org/10.48550/ARXIV.2106.00652.
33. Cheplygina, V.; de Bruijne, M.; Pluim, J.P.W. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med Image Anal.* **2019**, *54*, 280–296. [CrossRef] [PubMed]
34. Guan, H.; Liu, M. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Trans. Biomed. Eng.* **2022**, *69*, 1173–1185. [CrossRef] [PubMed]
35. Tuggener, L.; Satyawan, Y.P.; Pacha, A.; Schmidhuber, J.; Stadelmann, T. The DeepScoresV2 dataset and benchmark for music object detection. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 9188–9195, ISBN 978-1-7281-8808-9. [CrossRef]

FormulaNet: A Benchmark Dataset for Mathematical Formula Detection

FELIX M. SCHMITT-KOOPMANN[1, 2], ELAINE M. HUANG[2], HANS-PETER HUTTER[1] (Member, IEEE), THILO STADELMANN[3, 4] (Senior Member, IEEE), ALIREZA DARVISHY[1]
[1]ZHAW Institute of Applied Information Technology, Winterthur, Switzerland
[2]UZH People and Computing Lab, Zurich, Switzerland
[3]ZHAW Centre for Artificial Intelligence, Winterthur, Switzerland
[4]ECLT European Centre for Living Technology, Venice, Italy

Corresponding author: Felix M. Schmitt-Koopmann (e-mail: scmx@zhaw.ch).

This work was supported by the Bridge Discovery program of the Swiss National Science Foundation under grant number 194677.

ABSTRACT One unsolved sub-task of document analysis is mathematical formula detection (MFD). Research by ourselves and others has shown that existing MFD datasets with inline and display formula labels are small and have insufficient labeling quality. There is therefore an urgent need for datasets with better quality labeling for future research in the MFD field, as they have a high impact on the performance of the models trained on them. We present an advanced labeling pipeline and a new dataset called FormulaNet in this paper. At over 45k pages, we believe that FormulaNet is the largest MFD dataset with inline formula labels. Our experiments demonstrate substantially improved labeling quality for inline and display formulae detection over existing datasets. Additionally, we provide a math formula detection baseline for FormulaNet with an mAP of 0.754. Our dataset is intended to help address the MFD task and may enable the development of new applications, such as making mathematical formulae accessible in PDFs for visually impaired screen reader users.

INDEX TERMS automatic annotation, dataset, document analysis, deep learning, mathematical formula detection, page object detection

I. INTRODUCTION

THE 2008 United Nations Convention on the Rights of Persons with Disabilities [1] and the 2019 European Accessibility Act [2] require that everyday products and services be usable for people with disabilities. Nevertheless, many technologies remain inaccessible; PDFs are one such technology that frequently present a barrier for readers with visual impairments. This is especially true for scientific PDFs. For example, mathematical formulae in PDFs are usually not tagged with alternative text, making it impossible for screen reader software to read them out in a comprehensible way. Research has shown that most authors of scientific documents are unfamiliar with the concept of PDF accessibility, or lack the tools to support it [3].

Document analysis offers high potential for new applications, including applications for people with disabilities. One such application is the automated addition of accessibility tags to a PDF. Such accessibility tags allow a visually impaired person to read a PDF with a screen reader. Currently, tags must be added manually, which requires a great deal of time, expert knowledge, and awareness [3].

With effective document analysis, the tagging process could be automated or semi-automated, thus reducing the required time and expert knowledge necessary. This could help to increase the overall availability of tagged PDFs and as a result, give visually impaired people more complete access to information. However, the challenges of automated document analysis have not yet been solved. Searching for simple text in documents is currently possible [4]; however, the detection of more complex structures within a text, such as tables, graphs, or formulae remains problematic.

New data-driven approaches have enabled significant advancements in the document analysis field [5]. Most data-driven document analysis solutions work with images of document pages. This has the advantage that the approach can be applied regardless of the document format and version.

The first step planned for our document analysis pipeline is page object detection (POD). It aims to locate logical objects in document pages with a high semantic level, e.g., paragraphs, footnotes, tables, figures, or mathematical

formulae. In the next step, these objects will be processed by formula recognition, figure classification, text analysis, and other means.

The POD task is often divided into subtasks of locating a single logical object at a time. Despite the progress of POD in recent years [4], [6], [7], some objects are still challenging to identify and need to be addressed further. One of these open problems is mathematical formula detection (MFD) [8]. MFD is especially important for scientific documents from STEM fileds (science, technology, engineering, and mathematics), because mathematical formulae are often important objects for the understanding of STEM articles. Automated processing of formulae could help to simplify and improve many tasks, such as searching for mathematical formulae in documents, extracting mathematical formulae, and making mathematical formulae accessible.

In recent years, many MFD models have been proposed [4], [6], [7], but one problem that the authors of this paper have identified is that the MFD datasets they have been evaluated on have been of limited size and quality.

A selection of the most popular POD datasets is presented in Table 1. Existing POD datasets [9]–[15] are of limited value for the MFD issues we are attempted to address because of three reasons. First, most POD dataset were not intended for the MFD task and hence, consider no mathematical formulae or only display formulae but not inline formulae. Second, existing datasets with inline formulae tend to be small for deep learning approaches with less than 10k pages. Third, the mathematical formulae labels have insufficient quality or are incorrect. In this, paper, we propose a new large-scale and high-quality dataset for the MFD task of scientific PDF documents. It is created from the LaTeXsource [16] of papers from arXiv.org [17].

The main contributions of this paper are as follows: (a) a novel large-scale, high-quality dataset for MFD with practical relevance for document accessibility and, in conjunction with the provided baselines, scientific use as a benchmark suite; (b) an advanced fully automated labeling pipeline for constructing similar high-quality datasets of POD of nearly any size.

Due to copyright issues, we can only provide the links to the papers used and the postprocessing scripts to reconstruct FormulaNet, but not the images of FormulaNet. The scripts are publicly available at https://github.com/felix-schmitt/FormulaNet. Due to the compiling of the LaTeXfiles, the resulting pixel values may differ. We observed that on average 0.1% of the binary pixel values and 10.4% of the color pixel values variate.

The remainder of this paper is organized as follows: Chapter II presents related work and existing datasets. Chapter III presents our definition of inline and display formulae and introduces our dataset and labeling pipeline. Chapter IV presents the baseline model and experiments to demonstrate the improvement in labeling quality. Chapter V provides concluding remarks.

TABLE 1. Overview of a selection of the most popular POD datasets.

Dataset	Pages	Inline Labels	Inline-Accuracy	Display-Accuracy
Marmot [9]	400	Yes	76.90%	88.72%
ICDAR 2017 POD (corrected) [10]	2,417	No	-	-
IBEM [11]	8,272	Yes	96.72%	83.38%
FormulaNet	**46,672**	**Yes**	**98.08%**	**97.86%**
GROTOAP2 [12]	119,334	No	-	-
PubLayNet [13]	364,232	No	-	-
TableBank [14]	417,234	No	-	-
DocBank [15]	500,000	No	-	-

II. RELATED WORK AND EXISTING DATASETS

POD has been an active research area for several years [4], [6], [7]. The MFD subtask has been researched since at least 1968 [18] and efforts in this area have increased in recent years. Traditional MFD solutions are rule-based. However, object recognition using deep learning models has achieved good results and is replacing traditional rule-based approaches. Modern MFD models use convolutional neural networks (CNN) and build upon state-of-the-art object detections models, e.g., Faster-RCNN [19], Mask-RCNN [20], and FCOS [21]. The major challenge with MFD is the variation in complexity between small single mathematical elements and large mathematical formulae. Research [23] has shown that deformable CNNs [22], with their adaptive geometric transformation, have the ability to handle large variations in size. Furthermore, Generalized Focal Loss [24] reduces the imbalance issue of positive/negative sampling of large and small objects. As baseline model, we use the 1st place solution of the in ICDAR 2021 Competition on Mathematical Formula Detection [23] with small modifications. It is built upon FCOS and uses both modifications.

The competition [4] showed that MFD models can achieve excellent results in terms of F1 scores, but inline formulae are still challenging for these models and additional work is needed to address. One reason is that large existing POD datasets do not include labels for inline formulae (ref. Table 1) and the ones containing inline formulae are limited in size and labeling quality. We explain this lack of dataset with inline formulae by the fact that inline formulae are uncommon and often not crucial for the understanding of non-STEM documents. Furthermore, the separation between inline formulae and text is not clearly defined, as presented in Chapter III-A. However, STEM documents contain many inline formulae, and their correct processing is important for many applications, such as accessible PDFs.

We are aware of only two publicly available MFD datasets with inline formulae based on not rearranged articles such as omitting content and changing layout. One is the Marmot dataset [9] with 400 pages. Due to its small size, it is not ideal for deep learning approaches. The largest dataset with inline formulae is the IBEM dataset [11] with 8,272 pages, which is 20 times larger than Marmot, but it is still small for deep learning approaches. In comparison, DeepScores [25], an object detection dataset for music scores, which is

a comparable object detection task, contains 300,000 pages. The IBEM dataset was created for the ICDAR 2021 Competition on Mathematical Formula Detection [4] to run the latest performance competition of MFD models. It was created in a fashion similar to FormulaNet, by detecting specific formula patterns in the LaTeXcode. The patterns detected were then used to create the ground truth labels.
The large-scale POD datasets are not designed for the MFD task and hence, contain no inline formulae labels. With FormulaNet, we narrow the gap between MFD datasets and large-scale POD datasets.

III. FORMULANET

This section describes the construction details and characteristics of the FormulaNet dataset. FormulaNet uses papers about High Energy Physics on arXiv.org from the years 2000, 2002, and 2003. We used the High Energy Physics papers for the FormulaNet dataset not only because such PDFs comprise many formulae, but also to make it more comparable to the IBEM dataset, which also uses High Energy Physics papers from arXiv.org.

A. LABEL DEFINITIONS

There are no widely accepted standard definition for inline forumlae or display formulae. For the purposes of this reasearch, we provide working definitions of these terms based on the rules detected from the Marmot, ICDAR, and IBEM datasets:

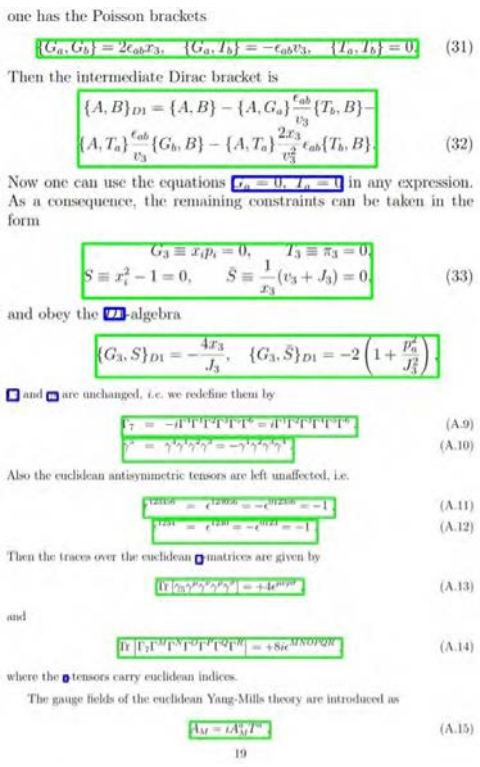
1) Inline Formulae:

We define inline formulae as all math-typed elements embedded in a text, except plain numbers.

An inline formula can consist of a single math element such as γ or a more complex formula consisting of multiple such elements. A single number is not considered as an inline formula for two reasons: First, in the existing datasets most numbers are not labeled as formulae. Second, numbers can already be processed through standard text optical character recognition (OCR). However, if a number comprises math structure elements like super-scripts or fractions, we consider it an inline formula because it is a mathematical construct, and text OCR will likely have problems interpreting it correctly. Mathematical elements within tables are not considered inline formulae because detecting a table structure is a challenging task, and detecting formulae within the table is a subtask of this task. For the same reason, mathematical elements in figures are not labeled as inline formulae, because formulae within figures need to be considered separately, similar to formulae within tables.

2) Display Formulae:

We define display formulae to be all-mathematical elements isolated from the running text. Multiline display formulae are separated depending on the formula references.


one has the Poisson brackets

$$\{G_a, G_b\} = 2\epsilon_{ab}x_3, \quad \{G_a, T_b\} = -\epsilon_{ab}v_3, \quad \{T_a, T_b\} = 0, \tag{31}$$

Then the intermediate Dirac bracket is

$$\{A,B\}_{D1} = \{A,B\} - \{A,G_a\}\frac{\epsilon_{ab}}{v_3}\{T_b,B\} - \{A,T_a\}\frac{\epsilon_{ab}}{v_3}\{G_b,B\} - \{A,T_a\}\frac{2x_3}{v_3^2}\epsilon_{ab}\{T_b,B\}. \tag{32}$$

Now one can use the equations $G_a = 0$, $T_a = 0$ in any expression. As a consequence, the remaining constraints can be taken in the form

$$G_3 \equiv x_i p_i = 0, \quad T_3 \equiv \pi_3 = 0, \quad S \equiv x_i^2 - 1 = 0, \quad \bar{S} \equiv \frac{1}{x_3}(v_3 + J_3) = 0, \tag{33}$$

and obey the $\mathcal{M}$-algebra

$$\{G_3, S\}_{D1} = -\frac{4x_3}{J_3}, \quad \{G_3, \bar{S}\}_{D1} = -2\left(1 + \frac{p_a^2}{J_3^2}\right),$$

$\mathcal{H}$ and $\mathcal{H}_a$ are unchanged, i.e. we redefine them by

$$\bar{\gamma}_7 = -i\gamma^1\gamma^2\gamma^3\gamma^4\gamma^5\gamma^6 = i\gamma^1\gamma^2\gamma^3\gamma^4\gamma^5\gamma^6, \tag{A.9}$$

$$\bar{\gamma}^\mu = \gamma^1\gamma^2\gamma^3\gamma^4\gamma^\mu = -\gamma^\mu\gamma^1\gamma^2\gamma^3\gamma^4 \tag{A.10}$$

Also the euclidean antisymmetric tensors are left unaffected, i.e.

$$\epsilon^{123456} = \epsilon^{126345} = -\epsilon^{012346} = -1, \tag{A.11}$$

$$\epsilon^{1234} = \epsilon^{1230} = -\epsilon^{0123} = -1 \tag{A.12}$$

Then the traces over the euclidean γ-matrices are given by

$$\mathrm{tr}\left[\bar{\gamma}_5\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma\right] = +4\epsilon^{\mu\nu\rho\sigma}, \tag{A.13}$$

and

$$\mathrm{tr}\left[\Gamma_7\Gamma^M\Gamma^N\Gamma^O\Gamma^P\Gamma^Q\Gamma^R\right] = +8i\epsilon^{MNOPQR} \tag{A.14}$$

where the γ tensors carry euclidean indices.

The gauge fields of the euclidean Yang-Mills theory are introduced as

$$A_M = iA_M^a T^a \tag{A.15}$$

19

FIGURE 1. Examples of how multicolumn display labels are separated. Green shows the display formulae and blue the inline formulae.

Formula references are not counted as part of a formula, because they are document structure elements and not part of the formula itself. This has the advantage that the bounding box size does not depend on the existence of a formula reference. Furthermore, we decided to only split up a multiline display formula into separate formulae if there is a formula reference on each line, as shown in Fig. 1. Splitting up a display formula line-by-line would have the effect of dividing a single formula into multiple parts, thus making it more complicated to process.

B. LABELING PROCESS

The labeling pipeline starts from the LaTeXsource files. It involves two labeling steps and one correction step as shown in Fig. 2. The first step is to modify the LaTeXcode to color each LaTeXobject. Depending on the object type, we use one or multiple colors to simplify the later separation. Two methods were combined to colorize the LaTeXcode. The first method uses regular expression search [26] to find predefined sequences in the LaTeXcode which are typical for a logical object class. Then, the sequences identified are colored with the xcolor package [27] and the following command:

```
\textcolor{l_color}{label}
```

The second method colors complete LaTeXenvironments with

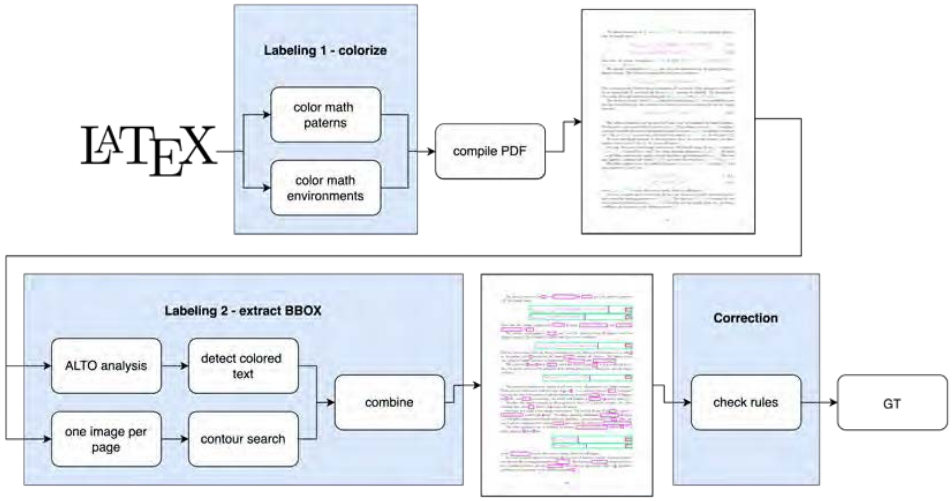


FIGURE 2. Overview of the labeling pipeline.

the following LaTeXcommand:

```
\AtBeginEnvironment{l_env}{l_color}
```

The modified LaTeXfile is used to render a PDF of the paper with the colored logical objects. In the second part, the colored objects of the modified PDF are detected and combined into one bounding box by heuristic rules. A combination of two methods is used to enhance the labeling quality. One method converts the PDF into the ALTO format [28] with pdfalto [29]. The resulting XML files contain information about the elements detected and it allows the identification of all colored elements. Since pdfalto is an OCR engine mainly for text it does not detect all symbols correctly. We therefore apply the second method to find the missing symbols.
For the second step, a PNG image of each page is rendered using a modified version of pdf2image [30] without anti-aliasing. This modification allows us to create images with clear contours which simplifies the contour search (OpenCV implementation [31]). This enables the detection of all missing colored pixels such as bars, heads, and other special math symbols. All BBOXs of the pdfalto and contour search are then combined with heuristic rules. Using only contour search would make it complicated or even impossible to get the correct combination of contours to a BBOX.
The last step is the correction step. It detects labeling errors, and depending on the errors detected, deletes entire pages or even the whole document. The rules applied are based on our observations during developing the pipeline, e.g.:

- These rules indicate an error in the coloring step:
  - If the paper has 3 or fewer pages, the document is discarded.
  - If the paper has no inline or display formulae, the document is discarded.
  - If there exist black pixels in a 30-pixel border of the document, the document is discarded.
- These rules indicate and error in the extracting BBOX step:
  - If there are more than 3 small display formulae, the page is discarded.
  - If there are not enough black pixels in an image, the page is discarded.
  - If the sum of all label areas is less than 10% of the page, the page is discarded.

After the correction step, a txt-file of each page is created with the detected BBOXs and a corresponding JPG image of the page with a resolution of 1447x2048 is saved. If the ratio of the document does not match the image ratio, a white border is added.

C. FORMULANET CHARACTERISTIC

FormulaNet consists of 46,672 pages with 175,685 display labels and 825,838 inline labels. Besides formula labels, FormulaNet contains 11 other labels (display reference, display both, header, table, figure, paragraph, caption, footnote, footnote reference, list, bibliography). We have randomly split the dataset into training (95% of the pages) and test (5% of the pages) sets. The distribution of the labels can be found in Table 2.

IV. COMPARISON WITH OTHER MFD DATASETS

To present the advantages of the proposed dataset, we used the currently best available FCOS model, I.e. [21] with selected modifications from Zhong [23]. We identified two main benefits of this model: First, the FCOS model is an object detection model without anchor boxes. The main advantage of an anchor-free object detection model is that it

TABLE 2. Distribution of the labels of the FormulaNet Dataset

Label	Train (44,338 pages)		Test (2,334 pages)	
	Total	Per Page	Total	Per Page
Inline Formulae	784,978	17.71±13.2	40,860	17.51±13
Display Formulae	166,759	3.76±2.9	8,936	3.83±2.8
Bibliography	1,086	0.02±0.2	56	0.02±0.2
Caption	3,671	0.08±0.3	203	0.09±0.3
Display Reference	144,800	3.27±2.9	7,799	3.34±2.8
Display Formulae + Reference	144,800	3.27±2.9	7,799	3.34±2.8
Footnote	8,109	0.18±0.4	438	0.19±0.4
Footnote Reference	11,576	0.26±0.7	632	0.06±0.3
Header	20,818	0.47±0.6	1,082	0.46±0.6
List	2,539	0.06±0.3	136	0.06±0.3
Paragraph	283,933	6.4±2.7	15,008	6.43±2.6
Table	1,145	0.03±0.2	49	0.02±0.2

avoids the complicated calculations related to anchor boxes and has no anchor box hyper-parameters. Second, it uses the Generalized Focal Loss [24]. This allows the model to handle the large size differences between inline formulae and display formulae. Furthermore, these modifications have shown to be successful in competition [4]. The model is built upon Zhong's implementation [32], which uses the MMDetection toolbox [33]. Since we trained the models with one NVIDIA Tesla-V100, we used the ResNetSt-50 model and not the suggested ResNetSt-101. We trained the model with the training datapoints of the FormulaNet dataset and, for comparison, with the Tr00, Tr01, Tr10, Va00, Va01, Ts00, and Ts01 datapoints of the IBEM dataset. As we used one GPU for training, we increased the batch size from 3 to 5, decreased the learning rate from 10^{-3} to 10^{-4}, and trained it for 24 epochs. The model config files are publicly available on https://github.com/felix-schmitt/FormulaNet and the results can be reproduced by using the framework from Zhong [32].

A. EXPERIMENTS

We demonstrate the high quality of our labels and the resulting advantage for the model training with three experiments. The first experiment, which we call "Labeling Quality", investigates the quality of the labels. The second experiment is named "Dataset Comparison"; it analyses the prediction errors on existing datasets of the model trained with FormulaNet. The third experiment, "Out-of-Sample", investigates the generalization capability of models trained with FormulaNet. All results of the experiments should be interpreted with some caution, as only a randomized sample of the test PDFs was examined, and the evaluation was carried out manually.

Contrary to our definition of display formulae, the Marmot dataset includes the reference number to the display formula bounding box as shown in Fig. 3. Through the different display formula definition, we did not count this as an error in the experiment "Labeling Quality" and we did not count it as an error if the model predicted the display formula without

FIGURE 3. Example image from Marmot dataset. Red shows the GT of Marmot and blue the predicted bounding box. Due to our definition of display formulae, this was counted as correct.

the reference number in the experiment "Dataset Comparison". Detailed experiment results are publicly available on https://github.com/felix-schmitt/FormulaNet/.

1) Labeling Quality

To investigate the labeling quality of the different datasets, we checked 100 randomly sampled pages of each dataset by hand. We counted the correct labels (CL), wrong labels (WL), wrong dimensions (WD), and missed labels (ML). CL BBOXs cover all pixels from the desired formula and no pixels from non-formula elements, while WD BBOXs contain pixels from non-formula elements or cover only parts of the desired formula. WL BBOXs cover no pixels from the corresponding formula or overlap with another BBOX. MLs are formulae that failed to be labeled as such. To make the results comparable, we put them in relation to the correct number of ground truth (CGT) labels, which is the sum of CL, WD, and ML. The pages without any labeling error (PWE) are the percentage of pages without any WL, WD, and ML of inline or display labels. This corresponds to the approximate amount of work required to clean up all errors manually. The results are shown in Table 3. The results for inline labels show that IBEM and FormulaNet have 8 times fewer labeling errors than Marmot, and furthermore, FormulaNet has 30% fewer labeling errors than IBEM. Marmot has the lowest ratio of WL, but the highest ratio of ML. The analysis of the errors revealed that the inline labels of Marmot are very accurate, but are missing many inline formulae compared to the other two datasets. Compared to IBEM, FormulaNet decreases the ratios of all three error types (WL, WD, ML) by 30-80%. One reason is FormularNet's consistent definition of inline formulae, in comparison with IBEM's inconsistent labeling of formulae in figures as inline formulae, as shown in Fig. 4. The results for display formulae shows that the labeling errors of FormulaNet are 10 times less frequent than those of IBEM and Marmot. The lower labeling quality of IBEM and Marmot is primarily caused by not properly splitting and merging the display formulae as shown in Fig. 5.

Additionally, the PWE of FormulaNet shows that fewer than 15% of the pages have any labeling error, which is 4 and 6 times less than IEBM and Marmot, respectively. This also clearly indicates the better labeling quality of FormulaNet compared to IBEM and Marmot.

2) Dataset Comparison

The "Dataset Comparison" experiment investigates whether a model benefits from the high labeling quality of the FormulaNet dataset, and whether a model trained with FormulaNet

TABLE 3. Results of "Labeling Quality" with the three datasets IBEM, Marmot, and FormulaNet. The table shows the ratios of correct labeled labels (CL) over the correct number of GT labels (CGT), wrong labels (WL) over CGT, wrong dimension of the BBOX (WD) over CGT, and missed labels (ML) over CGT for the two label types Inline and Display. Further, it shows the percentage of pages without a labeling error (PWE).

Label	Inline Formulae				Display Formulae				Pages
Dataset	CL/CGT (CL)	WL/CGT (WL)	WD/CGT (WD)	ML/CGT (ML)	CL/CGT (CL)	WL/CGT (WL)	WD/CGT (WD)	ML/CGT (ML)	PWE (PWE)
IBEM	96.72% (1533)	2.08% (33)	1.01% (16)	2.24% (36)	83.38% (316)	2.9% (11)	7.92% (30)	8.64% (33)	59% (41)
Marmot	76.9% (1808)	**0.38%** (9)	2.81% (66)	17.27% (477)	88.72% (354)	12.28% (49)	9.27% (37)	2.16% (8)	11% (89)
FormulaNet	**98.08%** (1529)	0.45% (7)	**0.38%** (6)	**1.54%** (24)	**97.86%** (365)	**0.27%** (1)	**1.61%** (6)	**0.54%** (2)	**84%** (16)

TABLE 4. Results of "Dataset Comparison" experiment with the datasets IBEM Ts10, IBEM Ts11, Marmot, and FormulaNet (test). The table shows the recall, precision for an IoU threshold of 0.5 and an NMS value of 0.4. The non-predicted GT BBOXs (NPs) and the wrongly predicted BBOXs (WPs) are manually checked if an NP should be not a GT (NGT) and if a WP should be a GT (SGT).

Label	Inline Formulae						Display Formulae					
Dataset	Recall	Precision	SGT/WP (WP)	SGT/CGT (SGT)	NGT/NP (NP)	NGT/GT (NGT)	Recall	Precision	SGT/WP (WP)	SGT/CGT (SGT)	NGT/NP (NP)	NGT/GT (NGT)
IBEM Ts10	94.73%	94.52%	36% (50)	1.98% (18)	39.58% (48)	2.09% (19)	94.64%	92.98%	58.33% (12)	4.14% (7)	66.67% (9)	**3.57%** (6)
IBEM Ts11	**95.16%**	**97.8%**	40% (15)	**0.87%** (6)	61.76% (34)	2.99% (21)	89.23%	82.08%	94.74% (38)	16.98% (36)	90.48% (21)	9.74% (19)
Marmot	82.93%	66.13%	84.63% (423)	27.39% (358)	27.65% (170)	4.72% (47)	75.47%	94.49%	28.57% (7)	1.46% (2)	61.54% (39)	15.09% (24)
FormulaNet (test)	94.91%	94.38%	**35.29%** (51)	1.98% (18)	**23.91%** (46)	**1.22%** (11)	**98.96%**	**95.5%**	**0%** (9)	**0%** (0)	**0%** (2)	**0%** (0)

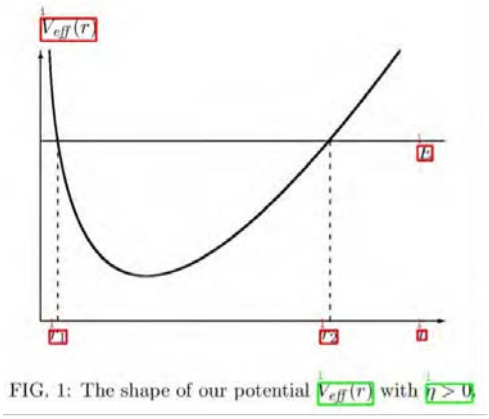


FIGURE 4. Example page from IBEM Ts11. Red shows the GT inline labels, that are not inline labels with our inline definition.

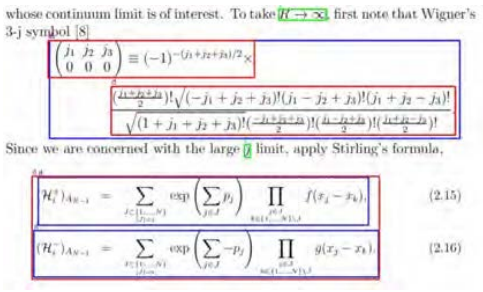


FIGURE 5. Examples from IBEM Ts11 of split and merge errors. Red shows the GT of IBEM and blue shows possible BBOX with our display definition.

can detect errors in existing datasets.

For the experiment, the model was trained with the FormulaNet dataset. We used the trained model to test the predictions on the IBEM Ts10 and IBEM Ts11 and Marmot datasets, and randomly selected 50 pages from each dataset. We used an Intersection of Union (IoU) threshold of 0.5 and an Non-maximum Suppression (NMS) value of 0.4 for the evaluation. Any non-predicted GT BBOXs (NPs) (with IoU smaller than 0.5 or no overlap) were manually checked to determine whether they are a correct GT or should not be a GT (NGT). Moreover, any incorrectly predicted BBOXs (WP) are manually checked for whether they should be a GT (SGT). For comparison, we have added the FormulaNet test set results. The results are shown in Table 4.

The high recall and precision values of the two IBEM test datasets indicate a similar labeling strategy of IBEM and FormulaNet. The model trained on the FormulaNet training set reached a combined F1 score (inline formulae and display formulae) of 94.49% for the 50 pages of IBEM Ts10, 93.97% for IBEM Ts11, and 94.26% for IBEM Ts10 + IBEM Ts11. Since the challenge [4] used an IoU threshold of 0.7, the values are not fully comparable. With an IoU threshold of 0.7 and all pages of Ts10 and Ts11, the model reaches an F1 score of 84.58%, which is only 2% lower than the results in the challenge [4] without using the training data.

The lower precision and recall values on IBEM Ts11 for display formulae are a result of the small number of pages, along with an excessive number of split and merge errors of display formulae (shown in Fig. 5). Additionally, the high

SGT and NGT ratios indicate that many of these errors are errors in the ground truth of IBEM Ts11. These results verify that the model trained with FormulaNet can detect labeling errors in the IBEM dataset.

The recall and precision values for our model tested with the Marmot test dataset are lower compared to the results on the two IBEM datasets. The corresponding accuracy of 88.02% for inline formulae and 76.51% for display formulae (86.81% combined) is slightly lower than the best models trained on Marmot [34]. However, the low NGT ratio and high SGT ratio for inline formulae of the Marmot dataset show that the Marmot inline labels are accurate, but not all inline formulae are in the GT, as the "Labeling Quality" experiment showed as well. The high NGT ratio of display formulas is primarily due to split and merge errors.

The precision and recall values with the FormulaNet test set show that the model accurately predicts inline and display formulae. The four display formulae indicators (SGT/WP, SGT/CGT, NGT/NP, and NGT/GT) are rather low with 0. We explain these zero values due to the small page set of 50 pages and hence few display formulae. However, the zero values indicate that the are only few labeling errors in the dataset and the model has learned very accurately to predict display formulae.

3) Out-of-Sample

For the "Out-of-Sample" experiment, we randomly selected 50 pages from over 1000 arXiv papers from all fields from 2021. We trained our model once with the IBEM dataset and once with the FormulaNet dataset. The trained models predicted the labels of the 50 pages. Since there are no annotations for these pages, we manually checked each BBOX to see if it was correct, incorrect, and if BBOXs were missing from the page. The definitions of CL, WD, and WL are the same as for the experiment "Labeling Quality". The recall is calculated as the ratio of CL over CGT and the precision as the ratio of CL over the sum of CL, WL, and WD. The results are shown in Table 5.

Even on papers from other fields, the model makes better prediction if it is trained with the FormulaNet dataset compared to when it is trained on the IBEM dataset. The model trained with FormulaNet reaches an 11.72% higher recall and a 24.02% better precision for inline labels, and a 12.16% higher recall and a 9.87% better precision for display formulae.

As expected, the performance of both models is substantially lower compared to the performance in the "Dataset Comparison" experiment with the IBEM dataset. There are two reasons for the lower performance. First, we used our CL definition and not an IoU of 0.5 because of the manual evaluation of the results. Second, the papers in this test are not from the same research field as the papers during training (IBEM uses papers from the same research field as FormulaNet).

B. BASELINE RESULTS ON FORMULANET DATASET

For a baseline performance on FormulaNet, we present here the results of two of the models trained with the FormulaNet dataset. The smaller model (FCOS-50) uses the ResNetSt-50 as backbone, as used for the experiments, and the larger model (FCOS-101) is based on the ResNetSt-101 backbone. The evaluation was conducted on the FormulaNet test set with the COCO metric [35]. The models are trained on the training set of the FormulaNet dataset and evaluated on the test set of the FormulaNet dataset after 24 epochs. Table 6 presents the results of 5 runs of the two baseline models. The results show that the larger backbone ResNetSt-101 does not significantly improve the model performance and the dataset is challenging for MFD models. The baseline model configs are publicly available on https://github.com/felix-schmitt/FormulaNet and can be reproduced using the framework of [32].

V. CONCLUSION

In this paper, we presented the FormulaNet dataset, a new dataset to train and benchmark MFD. FormulaNet is the largest dataset comprising labeled display and inline formulae and achieves an unprecedented labeling quality for this problem. FormulaNet was created by an automated labeling pipeline which will make it possible to create large high-quality datasets for future MFD research and benchmarking. Due to our automated labeling process and our proposed definition of inline and display formulae, the labels are very consistent compared with existing datasets. In addition to the FormulaNet dataset, we provide a strong baseline with one of the current best MFD models.

Through the design of the labeling pipeline, the dataset is limited to LaTeXpapers. Furthermore, FormulaNet is based only on High Energy Physics papers from arXiv.org. However, the "Out-of-Sample" experiment showed that the dataset still generalizes well to out-of-sample datapoints.

Given the promising results of our experiments, we are optimistic that FormulaNet can serve as a new Benchmark dataset for MFD to help to advance research in this area, which may finally result in new applications with high impact regarding accessible scientific PDFs.

REFERENCES

[1] "Convention on the Rights of Persons with Disabilities – Articles | United Nations Enable," 2008. Accessed: May 24, 2022. [Online]. Available: https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities/convention-on-the-rights-of-persons-with-disabilities-2.html

[2] Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services (Text with EEA relevance), vol. 151. 2019. Accessed: May 24, 2022. [Online]. Available: http://data.europa.eu/eli/dir/2019/882/oj/eng

[3] A. Jembu Rajkumar, J. Lazar, J. B. Jordan, A. Darvishy, and H.-P. Hutter, "PDF accessibility of research papers: what tools are needed for assessment and remediation?," Jan. 2020, pp. 4185–4194. doi: 10.24251/HICSS.2020.512.

[4] D. Anitei, J. A. Sánchez, J. M. Fuentes, R. Paredes, and J. M. Benedí, "ICDAR 2021 Competition on Mathematical Formula Detection," in Document Analysis and Recognition – ICDAR 2021, Cham, 2021, pp. 783–795. doi: 10.1007/978-3-030-86337-1_52.

TABLE 5. Results of the "Out-of-Sample" experiment with 50 random pages of 1000 arXiv 2021 papers. The table shows the resulting recall, precision, WL over CGT, and WD over CGT for the two label types Inline Formulae and Display Formulae.

Label	Inline Formulae				Display Formulae			
Training Dataset	Recall	Precision	WL/CGT (WL)	WD/CGT (WD)	Recall	Precision	WL/CGT (WL)	WD/CGT (WD)
IBEM	68.5%	68.32%	14.31% (164)	17.63% (202)	73.27%	77.08%	1.98% (2)	19.8% (20)
FormulaNet	**76.53%**	**84.73%**	**5.32%** (61)	**8.29%** (95)	**82.18%**	**84.69%**	**0.99%** (1)	**13.86%** (14)

TABLE 6. Results of the two baseline models (FCOS-50 and FCOS-101). The COCO metric is used for the evaluation.

Model	mAP Inline	mAP Display	mAP	mAP@50	mAP@75
FCOS-50 [$\mu \pm \sigma$]	0.752±0.02	**0.755±0.02**	0.754±0.03	**0.921±0.02**	0.84±0.02
FCOS-101 [$\mu \pm \sigma$]	**0.756±0.02**	0.749±0.03	**0.755±0.03**	0.920±0.02	**0.841±0.02**

[5] T. Stadelmann et al., "Deep Learning in the Wild," in Artificial Neural Networks in Pattern Recognition, Cham, 2018, pp. 17–38. doi: 10.1007/978-3-319-99978-4_2.

[6] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "ICDAR2017 Competition on Recognition of Documents with Complex Layouts - RDCL2017," in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Nov. 2017, vol. 01, pp. 1404–1410. doi: 10.1109/ICDAR.2017.229.

[7] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "ICDAR2019 Competition on Recognition of Documents with Complex Layouts - RDCL2019," in 2019 International Conference on Document Analysis and Recognition (ICDAR), Sep. 2019, pp. 1521–1526. doi: 10.1109/ICDAR.2019.00245.

[8] R. Zanibbi and D. Blostein, "Recognition and retrieval of mathematical expressions," IJDAR, vol. 15, no. 4, pp. 331–357, Dec. 2012, doi: 10.1007/s10032-011-0174-4.

[9] "Marmot Dataset." Accessed: Feb 23, 2022. [Online]. Available: https://www.icst.pku.edu.cn/cpdp/sjzy/

[10] "Icdar-2017 POD(corrected)," DFKI Cloud. Accessed: Feb. 23, 2022. [Online]. Available: https://cloud.dfki.de/owncloud/index.php/s/jrK3f9KEFSkwmgJ

[11] D. Anitei, J. A. Sánchez, and J. M. Benedí, "IBEM Mathematical Formula Detection Dataset." Accessed: May 13, 2021. doi: 10.5281/zenodo.4757865.

[12] D. Tkaczyk, P. Szostek, and Ł. Bolikowski, "GROTOAP2." RepOD, Sep. 29, 2015. doi: 10.18150/8527338.

[13] X. Zhong, J. Tang, and A. J. Yepes, "PubLayNet: largest dataset ever for document layout analysis," 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1015-1022, doi: 10.1109/ICDAR.2019.00166.

[14] M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, and Z. Li, "TableBank: Table Benchmark for Image-based Table Detection and Recognition," in Proceedings of the 12th Language Resources and Evaluation Conference, pp. 1918-1925, May, 2022.

[15] M. Li, Z. Xu, L. Cui, S. Huang, F. Wei, Y.Li, and M. Zhou, "DocBank: A Benchmark Dataset for Document Layout Analysis," in Proceeding of the 28th International Conference on Computational Linguistics, Dec. 2020, pp. 949-960, doi: 10.18653/v1/2020.coling-main.82.

[16] D. Knuth, "TeX Live." LaTeXVersion TeX Live 2021. Accessed: May 31, 2022. [Online]. Available: https://tug.org/texlive/

[17] "arXiv.org e-Print archive." Accessed: Feb. 23, 2022. [Online]. Available: https://arxiv.org/

[18] R. H. Anderson, "Syntax-directed recognition of hand-printed two-dimensional mathematics," in Symposium on Interactive Systems for Experimental Applied Mathematics: Proceedings of the Association for Computing Machinery Inc. Symposium, New York, NY, USA, Aug. 1967, pp. 436–459. doi: 10.1145/2402536.2402585.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in 2017 IEEE International Conference on Computer Vision (ICCV), Oct. 2017, pp. 2980–2988. doi: 10.1109/ICCV.2017.322.

[21] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2019, pp. 9626–9635. doi: 10.1109/ICCV.2019.00972.

[22] J. Dai et al., "Deformable Convolutional Networks," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 764–773.

[23] Y. Zhong et al., "1st Place Solution for ICDAR 2021 Competition on Mathematical Formula Detection," arXiv:2107.05534 [cs], Jul. 2021, Accessed: Feb. 23, 2022. [Online]. Available: http://arxiv.org/abs/2107.05534

[24] X. Li et al., "Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection," in Proceedings of the 34th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, Dezember 2020, pp. 21002–21012.

[25] L. Tuggener, I. Elezi, J. Schmidhuber, M. Pelillo, and T. Stadelmann, "DeepScores-A Dataset for Segmentation, Detection and Classification of Tiny Objects," in 2018 24th International Conference on Pattern Recognition (ICPR), Aug. 2018, pp. 3704–3709. doi: 10.1109/ICPR.2018.8545307.

[26] "regex — Regular expression operations." regex version 2022.1.18. Accessed: May 03,2022. [Online]. Available: https://docs.python.org/3/library/re.html

[27] The LaTeXProject, "xcolor." xcolor version 2.13. Accessed: May 03, 2022. [Online]. Available: https://ctan.org/pkg/xcolor

[28] "Analyzed Layout and Text Object (ALTO) XML Schema," ALTO XML Version 4.2. Accessed: May 03, 2022. [Online]. Available: https://www.loc.gov/standards/alto/

[29] P. Lopez, "pdfalto," pdfalto Version 0.5. Accessed: May 03, 2022. [Online]. Available: https://github.com/kermitt2/pdfalto

[30] Edouard Belvale, "pdf2image" pdf2image Version 1.16.0. Accessed: May 03, 2022. [Online]. Available: https://pypi.org/project/pdf2image/

[31] "OpenCV" py-opencv Version 4.5.5. Accessed: May 03, 2022. [Online] Available: https://opencv.org/

[32] Zhong, "1st Solution For ICDAR 2021 Competition on Mathematical Formula Detection." Accessed: May 23, 2022. [Online]. Available: https://github.com/Yuxiang1995/ICDAR2021_MFD

[33] K. Chen et al., "MMDetection: Open MMLab Detection Toolbox and Benchmark." arXiv, Jun. 17, 2019. Accessed: Mar. 05, 2022. [Online]. Available: http://arxiv.org/abs/1906.07155

[34] M. Z. Afzal, K. A. Hashmi, A. Pagani, M. Liwicki, and D. Stricker, "DeHyFoNet: Deformable Hybrid Network for Formula Detection in Scanned Document Images," Jan. 2022, doi: 10.20944/preprints202201.0090.v1.

[35] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," in Computer Vision – ECCV 2014, Cham, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

FELIX M. SCHMITT-KOOPMANN received the B.Sc. degree in mechanical engineering and the M.Sc. degree in Robotics, Systems, and Control from ETH Zurich, in 2019 and 2021. He is currently a Ph.D. Student with the People and Computing Lab of the University of Zurich, and he is a member of the Institute of Applied Informatics of ZHAW. His research interests include software engineering, accessibility, AI, and document analysis.

ELAINE M. HUANG is a professor of Human-Computer Interaction in the Department of Informatics at the University of Zurich, where she leads the People and Computing Research Group. Prior to joining UZH in 2010, she was a researcher at Motorola Labs and a professor in the department of Computer Science at the University of Calgary. She received her PhD from the College of Computing at the Georgia Institute of Technology in 2006. Her research focuses on the use of technology to address issues of inequality and other societal challenges.

HANS-PETER HUTTER studied Electrical Engineering at ETH Zurich where he received a doctor of technical science degree in 1997 for his work on hybrid HMM/ANN approaches to speech recognition over telepone lines. Hans-Peter joined the UBS Ubilab as post doc where he worked on a european project for HMM-based speaker indentification over the telephone. At the same time he was co-lecturer at ETHZ in two speech processing modules. In 1997 he joined the ZHAW Zurich University of Applied Sciences in Winterthur where he worked as professor in computer science on various projects in the area of speech recognition and user centered design of graphical and voice user interfaces. In 2005 he founded the InIT Institute of applied Information Technology at the ZHAW School of Engineering together with his colleagues and was head of the insitute until 2010. At the same time, he was also head of the Human-Information Interaction group of the InIT which he is still leading today.

THILO STADELMANN studied computer science in Giessen and received his doctor of science degree from Marburg University, Germany, in 2010 for work on multimedia analysis and voice recognition. Thilo worked in engineering and leadership roles in the automotive industry and is professor of AI/ML at the ZHAW School of Engineering in Winterthur, Switzerland, director of the ZHAW Centre for Artificial Intelligence and head of its Computer Vision, Perception and Cognition Group. He is IEEE Senior Member and a fellow of the European Centre for Living Technology in Venice, Italy.

ALIREZA DARVISHY is professor for ICT Accessibility and head of the ICT Accessibility Lab at Zurich University of Applied Sciences in Switzerland. He serves an independent reviewer for European research projects such as the Active Assisted Living (AAL) program, and is principle investigator of the “Accessible Scientific PDFs for All” project, funded by the Swiss National Science Foundation.

Making Sense of the Natural Environment

Christoph von der Malsburg (FIAS, Frankfurt and INI, ETH Zürich), Benjamin Grewe (INI, ETH Zürich), Thilo Stadelmann (CAI, ZHAW Winterthur)

The neural basis of cognition is unclear to this day. We here present a conceptual framework resolving the conflict Fodor & Pylyshyn (1988); Dever (2006) between symbolic and neural approaches. In our scheme, the cortical carriers of meaning are not individual neurons but sets of neurons supporting each other by mutual excitation. These sets and their supporting connectivity are called 'net fragments' or simply 'fragments.' Also fragments activate only as part of larger nets composed of overlapping fragments. Fragments play the role of composite symbols. As each neuron can be part of several fragments, and each fragment can overlap with several alternative other fragments, fragments can be likened to jigsaw puzzle pieces that fit together in innumerable different arrangements. Any such arrangement must, however, conform to a highly non-trivial consistency condition.

Net fragments and the composite nets they form are supported by specific patterns of synaptic connections. These are formed in development and learning by network self-organization, a process studied experimentally Goodhill (2007) and theoretically Willshaw & von der Malsburg (1979); Häussler & von der Malsburg (1983) on the example of the ontogenetic establishment of retinotopic fiber projections. This process selects net structures that are sparse (limited fan-in and fan-out of connections at each neuron) and are self-consistent such that a sufficient number of fibers converge on any one neuron from within the net. The composition rule for fragments to co-activate in a net is that together they form a net that is self-consistent (and would be stable under the process of network self-organization). Any particular large net (that is, set of active neurons) is unlikely to occur more than once in a life-time, so that only relatively small fragments have a chance to be active again and again to thus reach stability under network self-organization. But as these fragments overlap in multiple ways, cortex develops into an overlay of net fragments that supports an infinitude of consistent large-scale nets.

Among all possible thus-defined net structures a particular role is played by those that realize schema application. Each schema is an abstract structural description under which large numbers of instances can be united Bartlett (1932); Minsky (1974); Schank & Abelson (1977). Invariant object recognition has been modeled as schema application Arathorn (2002); Olshausen et al. (1995); Hinton (1981); Kree & Zippelius (1988); von der Malsburg (1988) realizable as a net that is representing schema, instance and the structure-preserving mapping between them Wolfrum et al. (2008). Natural intelligence may be defined as the ability of pursuing vital goals and intentions in varying contexts. Behavioral control has been classically described as schema application Shettleworth (2010). We propose the composition of nets out of fragments as basis for this process von der Malsburg et al. (2022).

In distinction to present-day artificial neural networks the human brain can learn and generalize from very few examples. It is a well-established insight Geman et al. (1992); Wolpert (1996) that such efficiency must be based on a deep structural relationship between learning system and domain. Inherent in our neural representation framework is therefore the claim that also the environment can be seen as a composite of a finite set of structural fragment types.

Keywords: neural representation, network self-organization, compositionality, net fragments, behavioral schema, intentions.

References

Arathorn, D. (2002). *Map-seeking circuits in visual cognition – a computational mechanism for biological and machine vision*. Stanford, California: Standford Univ. Press.

Bartlett, F. (1932). *Remembering, a study in experimental and social psychology*. Cambridge: Cambridge University Press.

Dever, J. (2006). Compositionality. In E. Lepore & B. Smith (Eds.), *The oxford handbook of philosophy of language* (pp. 633–666). Oxford University Press.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1), 3-71. doi: 10.1016/0010-0277(88)90031-5

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1-58.

Goodhill, G. J. (2007). Contributions of theoetical modeling to the understanding of neural map development. *Neuron*, *56*, 301-311.

Häussler, A. F., & von der Malsburg, C. (1983). Development of retinotopic projections: An analytical treatment. *J. Theoretical Neurobiology*, *2*, 47–73. Retrieved from `https://vfs.fias.science/d/3cfce0fe5a/files/?p=/Retina.pdf`

Hinton, G. E. (1981). A Parallel Computation that Assigns Canonical Object-Based Frames of Reference. In *International joint conference on artificial intelligence* (pp. 683–685).

Kree, R., & Zippelius, A. (1988). Recognition of topological features of graphs and images in neural networks. *J. Phys. A*, *21*, 813-818.

Minsky, M. (1974, June). *A framework for representing knowledge* (Tech. Rep. No. 306). MIT AI Laboratory.

Olshausen, B., Anderson, C., & Van Essen, D. (1995). A multiscale dynamic routing circuit for forming size- and position-invariant object representations. *Journal of Computational Neuroscience*, *2*, 45-62.

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. New Jersey: Erlbaum.

Shettleworth, S. (2010). *Cognition, evolution, and behavior (2nd ed.)*. Oxford: Oxford University Press.

von der Malsburg, C. (1988). Pattern recognition by labeled graph matching. *Neural Networks*, *1*, 141–148.

von der Malsburg, C., Stadelmann, T., & Grewe, B. (2022). *A theory of natural intelligence.* doi: 10.48550/ARXIV.2205.00002

Willshaw, D. J., & von der Malsburg, C. (1979). A marker induction mechanism for the establishment of ordered neural mappings; its application to the retinotectal problem. *Philosophical Transactions of the Royal Society of London, Series B*, *287*, 203–243.

Wolfrum, P., Wolff, C., Lücke, J., & von der Malsburg, C. (2008). A recurrent dynamic model for correspondence-based face recognition. *Journal of Vision*, *8*(7), 34. doi: 10.1167/8.7.34

Wolpert, D. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, *8*, 1341-1390.

frontiers

Is it enough to optimize CNN architectures on ImageNet?

Lukas Tuggener[1,2], **Jürgen Schmidhuber**[2,3,4], **Thilo Stadelmann**[1,5]

[1] *ZHAW Zurich University of Applied Sciences, Centre for Artificial Intelligence, Winterthur, Switzerland,* [2] *University of Lugano, Switzerland,* [3] *The Swiss AI Lab IDSIA, Switzerland,* [4] *AI Initiative, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia,* [5] *Fellow of the European Centre for Living Technology, Venice, Italy*

Correspondence*:
Lukas Tuggener
tugg@zhaw.ch

ABSTRACT

Classification performance based on ImageNet is the de-facto standard metric for CNN development. In this work we challenge the notion that CNN architecture design solely based on ImageNet leads to generally effective convolutional neural network (CNN) architectures that perform well on a diverse set of datasets and application domains. To this end, we investigate and ultimately improve ImageNet as a basis for deriving such architectures. We conduct an extensive empirical study for which we train 500 CNN architectures, sampled from the broad AnyNetX design space, on ImageNet as well as 8 additional well known image classification benchmark datasets from a diverse array of application domains. We observe that the performances of the architectures are highly *dataset dependent*. Some datasets even exhibit a negative error correlation with ImageNet across all architectures. We show how to significantly increase these correlations by *utilizing ImageNet subsets restricted to fewer classes*. These contributions can have a profound impact on the way we design future CNN architectures and help alleviate the tilt we see currently in our community with respect to over-reliance on one dataset.

Keywords: Deep Learning, CNN architecture design, ImageNet, empirical study

1 INTRODUCTION

Deep convolutional neural networks (CNNs) are the core building block for most modern visual recognition systems and lead to major breakthroughs in many domains of computer perception in the past several years. Therefore, the community has been searching the high dimensional space of possible network architectures for models with desirable properties. Important milestones such as DanNet Ciresan et al. (2012), AlexNet Krizhevsky et al. (2012), VGG Simonyan and Zisserman (2015), HighwayNet Srivastava et al. (2015), and ResNet He et al. (2016) (a HighwayNet with open gates) can be seen as update steps in this stochastic optimization problem and stand testament that the manual architecture search works. It is of great importance that the right metrics are used during the search for new neural network architectures. Only when we measure performance with a truly meaningful metric is it certain that a new high-scoring architecture is also fundamentally better. So far, the metric of choice in the community has generally been the performance on the most well-known benchmarking dataset—ImageNet Russakovsky et al. (2014).

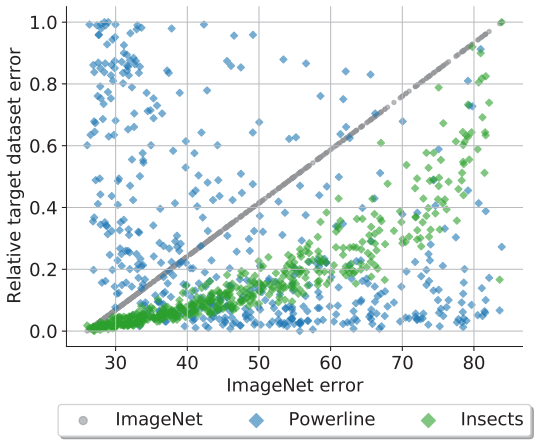


Figure 1. Is a CNN *architecture* that performs well on ImageNet automatically a good choice for a different vision dataset? This plot suggests otherwise: It displays the relative test errors of 500 randomly sampled CNN architectures on three datasets (ImageNet, Powerline, and Insects) plotted against the test error of the same architectures on ImageNet. The architectures have been trained from scratch on all three datasets. Architectures with low errors on ImageNet also perform well on Insects, on Powerline the opposite is the case.

More specifically, it would be desirable to construct such a metric from a solid theoretical understanding of deep CNNs. Due to the absence of a solid theoretical basis novel neural network designs are tested in an empirical fashion. Traditionally, model performance has been judged using accuracy point estimates Krizhevsky et al. (2012); Zeiler and Fergus (2014); Simonyan and Zisserman (2015). This simple measure ignores important aspects such as model complexity and speed. Newer work addresses this issue by reporting a curve of the accuracy at different complexity settings of the model, highlighting how well a design deals with the accuracy versus complexity tradeoff Xie et al. (2017); Zoph et al. (2018).

Very recent work strives to improve the quality of the empiric evaluation even further. There have been attempts to use extensive empirical studies to discover general rules on neural network design Hestness et al. (2017); Rosenfeld et al. (2020); Kaplan et al. (2020); Tuggener et al. (2020), instead of simply showing the merits of a single neural network architecture. Another line of research aims to improve empiricism by sampling whole populations of models and comparing error distributions instead of individual scalar errors Radosavovic et al. (2019).

We acknowledge the importance of the above-mentioned improvements in the empirical methods used to test neural networks, but identify a weak spot that runs trough the above-mentioned work: the heavy reliance on ImageNet Russakovsky et al. (2014) (and to some extent the very similar Cifar100 Krizhevsky et al. (2009)). In 2011, Torralba and Efros already pointed out that visual recognition datasets that were built to represent the visual world tend to become a small world in themselves Torralba and Efros (2011). Objects are no longer in the dataset because they are important, they are important because they are in the dataset. *In this paper, we investigate how well ImageNet represents a diverse set of visual classification datasets—and present methods to improve said representation, such that CNN architectures optimized on ImageNet become more effective on visual classification beyond ImageNet.* Specifically, our contributions are: (a) an extensive empirical study examining the fitness of ImageNet as a basis for deriving generally effective CNN architectures; (b) we show how class-wise subsampled versions of ImageNet in conjunction with the original datasets yield a 2.5-fold improvement in average error correlations with other datasets (c)

we identify cumulative block depth and width as the architecture parameters most sensitive to changing datasets.

As a tool for this investigation we introduce the notion of architecture and performance relationship (APR). The performance of a CNN architecture does not exist in a vacuum, it is only defined in relation to the dataset on which it is used. This dependency is what we call APR induced by a dataset. We study the change in APRs between datasets by sampling 500 neural network architectures and training all of them on a set of datasets[1]. We then compare errors of the same architectures across datasets, revealing the changes in APR (see Figure 1). This approach allows us to study the APRs induced by different datasets on a whole population of diverse network designs rather than just a family of similar architectures such as the ResNets He et al. (2016) or MobileNets Howard et al. (2017).

All of our code, sampled architectures, complete training run data, and additional figures are available at `https://github.com/tuggeluk/pycls/tree/ImageNet_as_basis`.

2 RELATED WORK

Neural network design. With the introduction of the first deep CNNs Ciresan et al. (2012); Krizhevsky et al. (2012) the design of neural networks immediately became an active research area. In the following years many improved architectures where introduced, such as VGG Simonyan and Zisserman (2015), Inception Szegedy et al. (2015), HighwayNet Srivastava et al. (2015), ResNet He et al. (2016) (a HighwayNet with open gates), ResNeXt Xie et al. (2017), or MobileNet Howard et al. (2017). These architectures are the result of manual search aimed at finding new design principles that improve performance, for example increased network depth and skip connections. More recently, reinforcement learning Zoph et al. (2018), evolutionary algorithms Real et al. (2019) or gradient descent Liu et al. (2019) have been successfully used to find suitable network architectures automatically. Our work relates to manual and automatic architecture design because it adds perspective on how stable results based on one or a few datasets are.

Empirical studies. In the absence of a solid theoretical understanding, large-scale empirical studies are the best tool at our disposal to gain insight into the nature of deep neural networks. These studies can aid network design Greff et al. (2017); Collins et al. (2017); Novak et al. (2018) or be employed to show the merits of different approaches, for example that the classic LSTM Hochreiter and Schmidhuber (1997) architecture can outperform more modern models Melis et al. (2018), when it is properly regularised. More recently, empirical studies have been used to infer more general rules on the behaviour of neural networks such as a power-law describing the relationship between generalization error and dataset size Hestness et al. (2017) or scaling laws for neural language models Kaplan et al. (2020).

Generalization in neural networks. Despite their vast size have deep neural networks shown in practice that they can generalize extraordinarily well to unseen data stemming from the same distribution as the training data. Why neural networks generalize so well is still an open and very active research area Kawaguchi et al. (2017); Dinh et al. (2017); Zhang et al. (2017). This work is not concerned with the generalization of a trained network to new data, but with the generalization of the architecture design progress itself. Does an architecture designed for a certain dataset, e.g. natural photo classification using ImageNet, work just as well for medical imaging? There has been work investigating the generalization to a newly collected test set, but in this case the test set was designed to be of the same distribution as the original training data Recht et al. (2019).

[1] Since we only sample models in the complexity regime of 340 mega flops (MF) to 400MF (ResNet-152 has 11.5GF) we could complete the necessary 7500 model trainings within a moderate 85 GPU days on Tesla V100-SXM2-32GB GPUs.

Neural network transferability It is known that the best architecture for ImageNet is not necessarily the best base architecture for other applications such as semantic segmentation Long et al. (2015) or object detection Chen et al. (2019). Researchers who computed a taxonomy of multiple visions tasks identified that the simmilarities between tasks did not depend on the used architecture Zamir et al. (2019). Research that investigates the relation between model performance on ImageNet and new classification datasets in the context of transfer learning Razavian et al. (2014); Donahue et al. (2014) suggests that there is a strong correlation which is also heavily dependent on the training regime used Kornblith et al. (2019). Our work differs form the ones mentioned above in that we are not interested in the transfer of learned features but transfer of the architecture designs and therefore we train our networks from scratch on each dataset. Moreover do we not only test transferability on a few select architectures but on a whole network space.

Neural network design space analysis. Radosavovic et al. Radosavovic et al. (2019) introduced network design spaces for visual recognition. They define a design space as a set of architectures defined in a parametric form with a fixed base structure and architectural hyperparameters that can be varied, similar to the search space definition in neural architecture search Zoph et al. (2018); Real et al. (2019); Liu et al. (2019). The error distribution of a given design space can be computed by randomly sampling model instances from it and computing their training error. We use a similar methodology but instead of comparing different design spaces, we compare the results of the same design space on different datasets.

3 DATASETS

To enable cross dataset comparison of APRs we assembled a corpus of datasets. We chose datasets according to the following principles: (a) include datasets from a wide spectrum of application areas, such that generalization is tested on a diverse set of datasets; (b) only use datasets that are publicly available to anyone to ensure easy reproducibility of our work. Figure 2a shows examples and Table 1 lists meta-data of the chosen datasets. More detailed dataset specific information is given in the remainder of this chapter.

Concrete Özgenel and Sorguç (2018) contains 40 thousand image snippets produced from 458 high-resolution images that have been captured from various concrete buildings on a single campus. It contains two classes, positive (which contains cracks in the concrete) and negative (with images that show intact concrete). With 20 thousand images in both classes the dataset is perfectly balanced.

MLC2008 Shihavuddin et al. (2013) contains 43 thousand image snippets taken form the MLC dataset Beijbom et al. (2012), which is a subset of the images collected at the Moorea Coral Reef Long Term Ecological Research site. It contains images from three reef habitats and has nine classes. The class distribution is very skewed with crustose coralline algae (CCA) being the most common by far (see Figure 11a in Appendix 6.1).

ImageNet Russakovsky et al. (2014) (ILSVRC 2012) is a large scale dataset containing 1.3 million photographs sourced from flickr and other search engines. It contains 1000 classes and is well balanced with almost all classes having exactly 1300 training and 50 validation samples.

HAM10000 Tschandl et al. (2018) is comprised of 10 thousand dermatoscopic images, collected from different populations and by varied modalities. It is a representative collection of all important categories of pigmented lesions that are categorized into seven classes. It is imbalanced with an extreme dominance of the melanocytic nevi (nv) class (see Figure 11a in Appendix 6.1).

Powerline Yetgin et al. (2017) contains images taken in different seasons as well as weather conditions from 21 different regions in Turkey. It has two classes, positive (that contain powerlines) and negative (which do not). The dataset contains 8000 images and is balanced with 4000 samples per classes.

Insects Hansen et al. (2019) contains 63 thousand images of 291 insect species. The images have been taken of the collection of British carabids from the Natural History Museum London. The dataset is not completely balanced but the majority of classes have 100 to 400 examples.

Intel Image Classification Bansal (2018) dataset ("natural") is a natural scene classification dataset containing 25 thousand images and 6 classes. It is very well balanced with all classes having between 2.1 thousand and 2.5 thousand samples in the training set.

Cifar10 and Cifar100 Krizhevsky et al. (2009) both consist of 60 thousand images. The images are sourced form the 80 million tiny images dataset Torralba et al. (2008) and are therefore of similar nature (photographs of common objects) as the images found in ImageNet, bar the much smaller resolution. Cifar10 has 10 classes with 6000 images per class, Cifar100 consists of 600 images in 100 classes, making both datasets perfectly balanced.

4 EXPERIMENTS AND RESULTS

4.1 Experimental setup

We sample our architectures form the very general AnyNetX Radosavovic et al. (2020) parametric network space. The networks in AnyNetX consist of a stem, a body, and a head. The body performs the majority of the computation, stem and head are kept fixed across all sampled models. The body consists of four stages, each stage i starts with a 1×1 convolution with stride s_i, the remainder is a sequence of d_i identical blocks. The blocks are standard residual bottleneck blocks with group convolution Xie et al. (2017), with a total block width w_i, bottleneck ratio b_i and a group width g_i (into how many parallel convolutions the total width is grouped into). Within a stage, all the block parameters are shared. See Figure 2b for a comprehensive schematic. All models use batch normalisation.

The AnyNetX design space has a total of 16 degrees of freedom, having 4 stages with 4 parameters each. We obtain our model instances by performing log-uniform sampling of $d_i \leq 16$, $w_i \leq 1024$ and divisible by 8, $b_i \in 1, 2, 4$, and $g_i \in 1, 2, ..., 32$. The stride s_i is fixed with a stride of 1 for the first stage and a stride of 2 for the rest. We repeatedly draw samples until we have obtained a total of 500 architectures in our target complexity regime of 360 mega flops (MF) to 400 MF. We chose a narrow band of complexities to allow for fair comparisons of architectures with minimal performance variation due to model size. We use a very basic training regime, input augmentation consists of only flipping, cropping and mean plus variance normalisation, based on each datasets statistics. For training we use SGD with momentum and weight decay.

The same 500 models are trained on each dataset until the loss is reasonably saturated. The exact number of epochs has been determined in preliminary experiments and depends on the dataset (see Table 2). For extensive ablation studies ensuring the empirical stability of our experiments with respect to Cifar10 performance, training duration, training variability, top-1 to top-5 error comparisons, overfitting and class distribution see chapters 6.1.1 to 6.1.6 in Appendix 6.1. Supplementary results on the effect of pretraining and the structure of the best performing architectures can be found in chapters 6.2.1 and 6.2.2 in Appendix 6.2.

4.2 Experimental results

We analyze the architecture-performance relationship (APRs) in two ways. For every target dataset (datsets which are not ImageNet) we plot the test error of every sampled architecture against the test error of the same architecture (trained and tested) on ImageNet, visualizing the relationship of the target dataset's APR with the APR on ImageNet. Second, we compute Spearman's ρ rank correlation coefficient Freedman et al. (2007). It is a nonparametric measure for the strength of the relation between two variables (here the error on the target datasets with the error of the same architecture on ImageNet). Spearman's ρ is defined on $[-1, 1]$, where 0 indicates no relationship and -1 or 1 indicates that the relationship between the two variables can be fully described using only a monotonic function.

Figure 3 contains the described scatterplots with the corresponding correlation coefficients in the title. The datasets plotted in the top two rows show a strong (Insects) or medium (MLC2008, HAM10000, Cifar100) error correlation with ImageNet. This confirms that many classification tasks have an APR similar to the one induced by ImageNet, which makes ImageNet performance a decent architecture selection indicator for these datasets. The accuracies on Concrete are almost saturated with errors between 0 and 0.5, it is plausible that the variations in performance are due to random effects rather than any properties of the architectures or the dataset, especially so since the errors are independent of their corresponding ImageNet counterparts. Therefore we refrain from drawing any further conclusions from the experiments on Concrete. This has implications for practical settings, where in such cases suitable architectures should be chosen according to computational and model complexity considerations rather than ImageNet performance, and reinforces the idea that practical problems may lie well outside of the ImageNet visual world Stadelmann et al. (2018). The most important insight from Figure 3, however, is that some datasets have a slight (Cifar10) or even strong (Powerline, Natural) *negative error correlation* with ImageNet. Architectures which perform well on ImageNet tend perform sub-par on these datasets. A visual inspection shows that some of the very best architectures on ImageNet perform extraordinarily poor on these three datasets. We can conclude that the *APRs can vary wildly between datasets and high performing architectures on ImageNet do not necessarily work well on other datasets*.

An analysis of the correlations between all datasets (see Figure 14 in Appendix 6.2) reveals that Powerline and Natural not only have low correlation with ImageNet but also with most of the other datasets making these two truly particular datasets. Interestingly is the correlation between Powerline and Naural relatively high, which suggests that there is a common trait that makes these two datasets behave differently. MLC 2008, HAM10000 and Cifar100 have a correlation of 0.69 with each other which indicates that they induce a very similar APR. This APR seems to be fairly universal since MLC 2008, HAM10000 and Cifar100 have a moderate to high correlation with all other datasets.

4.3 Impact of the Number of Classes

Having established that APR varies heavily between datasets, leaves us width the questions if it is possible to identify properties of the datasets themselves that influences its APR and if it is possible to control these factors to reduce the APR differences.

ImageNet has by far the largest number of classes among all the datasets. Insects, which is the dataset with the second highest class count, also shows the strongest similarity in APR to ImageNet. This suggests that the number of classes might be an important property of a dataset with respect to APR. We test this hypothesis by running an additional set of experiments on subsampled versions of ImageNet. We create new datasets by randomly choosing a varying number of classes from ImageNet and deleting the rest of

the dataset (see Section 3. in the supplementary material for chosen classes). This allows us to isolate the impact of the number of classes while keeping all other aspects of the data itself identical. We create four subsampled ImageNet versions with 100, 10, 5, and 2 classes, which we call ImageNet-100, ImageNet-10, ImageNet-5, and ImageNet-2, respectively. We refer to the resulting group of datasets (including the original ImageNet) as the ImageNet-X family. The training regime for ImageNet-100 is kept identical to the one of ImageNet, for the other three datasets we switch to top-1 error and train for 40 epochs, to account for the smaller dataset size. (see section 4.3.1 in Appendix 6.1 for a control experiment that disentangles the effects of reduced dataset size and reduced number of classes)

Figure 4 shows the errors on the subsampled versions plotted against the errors on original ImageNet. APR on ImageNet-100 shows an extremely strong correlation with APR on ImageNet. This correlation significantly weakens as the class count gets smaller. ImageNet-2 is on the opposite end has errors which are practically independent from the ones on ImageNet. *This confirms our hypothesis that the number of classes is a dataset property with significant effect on the architecture to performance relationship.*

We have observed that the number of classes has a profound effect on the APR associated with ImageNet-X members. It is unlikely that simply varying the number of classes in this dataset is able to replicate the diversity of APRs present in an array of different datasets. However, it is reasonable to assume that a dataset's APR is better represented by the ImageNet-X member closest in terms of class count, instead of ImageNet. We thus recreate Figure 3 with the twist of not plotting the target dataset errors against ImageNet, but against the ImageNet-X variant closest in class count (see Figure 5). We observe gain in correlation across all datasets, in the cases of MLC2008 or Cifar10 a quite extreme one. The datasets which have a strong negative correlation with ImageNet (Powerline, Natural) have slightly (Natural) or even moderately (Powerline) positive correlation to their ImageNet-X counterparts. A visual inspection shows that the best models on Imagenet-X also yield excellent results on Powerline and Natural, which was not the case for ImageNet. Table 3 shows the error correlations of all target datasets with ImageNet as well as with their ImageNet-X counterpart. *The move from ImageNet to ImageNet-X more than doubles the average correlation (from 0.19 to 0.507), indicating that the ImageNet-X family of datasets is capable to represent a much wider variety of APRs than ImageNet alone.*

4.3.1 Disentangling the Effects of Class Count and Dataset Size

We showed how sub-sampled versions of ImageNet matching the number of classes of the target dataset tend to represent the APR of said target dataset far better. A side effect of downsampling ImageNet to a specific number of classes is that the total number of images present in the dataset also shrinks. This raises the question if the increase in error correlation is actually due to the reduced dataset size rather than to the matching class count. We disentangle these effects by introducing another downsampled version of ImageNet, Imagenet-1000-10. It retains all 1000 classes but only 10 examples per class resulting in a datastet with the same number of classes as ImageNet but with the total number of images of ImageNet-10. We train our population of architectures on ImageNet-1000-10 and show the error relationship of Cifar10, Natural, and Powerline with ImageNet-1000-10 (as well as with ImageNet and ImageNet-10 as a reminder) in Figure 6. The plots show that there are some correlation gains by using ImageNet-1000-10 over ImageNet, but the effect is far lower compared to ImageNet-10. *This shows that downsampling size has a minor positive effect but the majority of the gain in APR similarity achieved trough class downsampling actually stems from the reduced the class number.*

4.4 Identifying Drivers of Difference between Datasets

The block width and depth parameters of the top 15 architectures for ImageNet (see Figure 13 in Appendix 6.2) follow a clear structure: they consistently start with low values for both block depth and width in the first stage, then the values steadily increase across the stages for both parameters. The error relationships observed in Figure 3 are consistent with how well these patterns are replicated by the other datasets. Insects shows a very similar pattern, MLC2008 and HAM10000 have the same trends but more noise. Powerline and Natural clearly break from this structure, having a flat or decreasing structure in the block width and showing a quite clear preference for a small block depth in the final stage. Cifar10 and Cifar100 are interesting cases, they have the same behaviour as ImageNet with respect to block width but a very different one when it comes to block depth.

We thus investigate the effect of the cumulative block depth (summation of the depth parameter for all four stages, yielding the total depth of the architecture) across the whole population of architectures by plotting the cumulative block depth against the test error for the six above-mentioned datasets. Additionally, we compute the corresponding correlation coefficients. Figure 7a shows that the best models for ImageNet have a cumulative depth of at least 10. Otherwise there is no apparent dependency between the ImageNet errors and cumulative block depth. The errors of Insects do not seem to be related to the cumulative block depth at all. HAM10000 has a slight right-leaning spread leading to a moderate correlation, but the visual inspection shows no strong pattern. The errors on Powerline, Natural, and Cifar100 on the other hand have a strong dependency with the cumulative block depth. The error increases with network depth for all three datasets. with the best models all having a cumulative depth smaller than 10.

We also plot the cumulative block widths against the errors and compute the corresponding correlation coefficients for the same six datasets (see Figure 7b). We observe that the ImageNet errors are negatively correlated with the cumulative block width, and visual inspection shows that a cumulative block width of at least 250 is required to achieve a decent performance. The errors on Insects and HAM10000 replicate this pattern to a lesser extent, analogous to the top 15 architectures. Powerline and Natural have no significant error dependency with the cumulative block width, but Cifar100 has an extremely strong negative error dependency with the cumulative block width, showing that it is possible for a dataset to replicate the behaviour on ImageNet in one parameter but not the other. In the case of Cifar100 and ImageNet, low similarity in block depth and high similarity in block width yield a medium overall similarity of ARPs on Cifar100 and Imagenet. This is consistent with the overall relationship of the two datasets displayed in Figure 3.

Combining this result with the outcome of the last section, we study the interaction between the number of classes, the cumulated block depth and the cumulative block width. Table 4 contains the correlations between cumulative block depth/width and the errors on all members of ImageNet-X. With decreasing number of classes, the correlation coefficients increase for cumulative block depth and cumulative block width. Although the effect on cumulative block depth is stronger, there is a significant impact on both parameters. We therefore can conclude that *both optimal cumulative block depth and cumulative block with can drastically change based on the dataset choice and that both are simultaneously influenced by the class count.*

5 DISCUSSION AND CONCLUSIONS

ImageNet is not a perfect proxy. We have set out to explore how well other visual classification datasets are represented by ImageNet. Unsurprisingly there are differences between the APRs induced by the

datasets. More surprising and worrying, however, is that for some datasets ImageNet not only is an imperfect proxy but a very bad one. The negative error correlations with Natural, Powerline and Cifar10 indicates that architecture search based on ImageNet performance is worse than random search for these datasets.

Varying the number of classes is a cheap and effective remedy. It is striking how much more accurately the ImageNet-X family is able to represent the diversity in APRs present in our dataset collection, compared to just ImageNet by itself. It has become commonplace to test new architectures in multiple complexity regimes He et al. (2016); Howard et al. (2017), we argue for augmenting this testing regime with an additional dimension for class count. This simple and easy to implement extension would greatly extend the informative value of future studies on neural network architectures.

Visual variability is less important than anticipated. In the introduction we critiqued the over-reliance on ImageNet based on the limits of ”visual world” it represents, since it only contains natural images and is mostly focused on animals and common objects. However, our results show that datasets with visually very different content such as Insects and HAM10000 have a high APR correlation with ImageNet. For Natural and Cifar10, which contain natural images, the opposite is the case. This shows that the visual domain of a dataset is not the central deciding factor for choosing the correct CNN architecture.

Future directions. A future similar study should shed light on how well the breadth of other domains such as object detection, segmentation or speech classification are represented by their essential datasets. If the representation is also insufficient it could be verified if the symptoms are similar and the varying the number of classes also helps covering more dataset variability in these domains.
A labeled dataset will always be a biased description of the visual world, due to having a fixed number of classes and being built with some systematic image collection process. Self-supervised learning of visual representations Jing and Tian (2019) could serve as remedy for this issue. Self-supervised architectures could be fed with a stream completely unrelated images, collected from an arbitrary number of sources in a randomized way. A comparison of visual features learned in this way could yield a more meaningful measure of the quality of CNN architectures.

Limitations As with any experimental analysis of a highly complex process such as training a CNN it is virtually impossible to consider every scenario. We list below three dimensions along which our experiments are limited together with measures we took to minimize the impact of these limitations.

Data scope: We criticize ImageNet for only representing a fraction of the “visual world”. We are aware that our dataset collection does not span the entire “visual world” either but went to great lengths to maximise the scope of our dataset collection by purposefully choosing datasets from different domains, which are visually distinct.

Architecture scope: We sample our architectures from the large AnyNetX network space. It contains the CNN building blocks to span basic designs such as AlexNet or VGG as well as the whole ResNet, ResNeXt and RegNet families. We acknowledge that there are popular CNN components not covered, however, Radosavovic et al. Radosavovic et al. (2020) present ablation studies showing that network designs sourced from high performing regions in the AnyNetX space also perform highly when swapping in different originally missing components such as depthwise convolutions Chollet (2017), swish activation functions Ramachandran et al. (2018) or the squeeze-and-excitation Hu et al. (2018) operations.

Training scope: When considering data augmentation and optimizer settings there are almost endless possibilities to tune the training process. We opted for a very basic setup with no bells an whistles in general. For certain such aspects of the training, which we assumed might skew the results of our study

Table 1. Meta data of the used datasets.

DATASET	NO. IMAGES	NO. CLASSES	IMG. SIZE
CONCRETE	40K	2	227×227
MLC2008	43K	9	312×312
IMAGENET	1.3M	1000	256×256
HAM10000	10K	7	296×296
POWERLINE	8K	2	128×128
INSECTS	63K	291	296×296
NATURAL	25K	6	150×150
CIFAR10	60K	10	32×32
CIFAR100	60K	100	32×32

Table 2. Dataset-specific experimental settings.

DATASET	NO. TRAINING EPOCHS	EVAL. ERROR
CONCRETE	20	TOP-1
MLC2008	20	TOP-1
IMAGENET	10	TOP-5
HAM10000	30	TOP-1
POWERLINE	20	TOP-1
INSECTS	20	TOP-5
NATURAL	20	TOP-1
CIFAR10	30	TOP-1
CIFAR100	30	TOP-5

Table 3. Comparison of error correlations between target datasets and ImageNet as well as the closest ImageNet-X member.

DATASET	ρ -IMAGENET	ρ -IMAGENET-X	DIFFERENCE
CONCRETE	0.001	0.106	0.105
MLC2008	0.476	0.811	0.335
HAM10000	0.517	0.608	0.091
POWERLINE	-0.436	0.294	0.73
INSECTS	0.967	0.95	-0.017
NATURAL	-0.38	0.186	0.566
CIFAR10	-0.104	0.45	0.554
CIFAR100	0.476	0.595	0.119
AVERAGE	0.19	0.507	0.317

(such as training duration, dataset prepossessing etc.), we have conducted extensive ablation studies to ensure that this is not the case (see sec. 6.1.2 and 6.1.6 in Appendix 6.1).

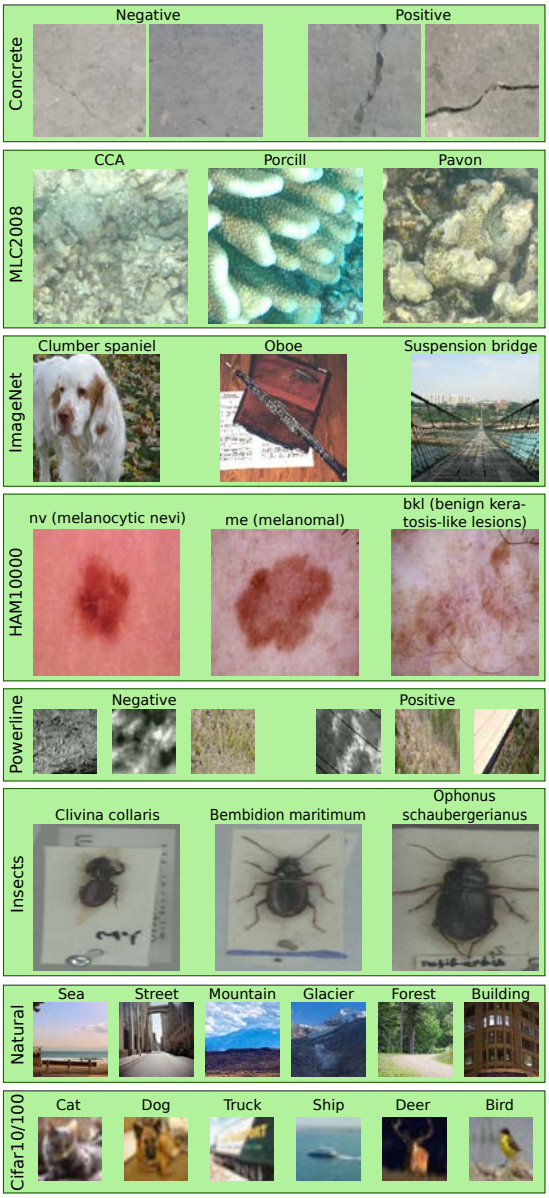
Acknowledgments

This work has been financially supported by grants 25948.1 PFES-ES “Ada” (CTI), 34301.1 IP-ICT “RealScore” (Innosuisse) and ERC Advanced Grant AlgoRNN nr. 742870. Open access funding provided

Table 4. Correlation of observed error rates with the cumulative block depth and width parameters for all ImageNet-X datasets.

DATASET	C. BLOCK DEPTH	C. BLOCK WIDTH
IMAGENET	-0.205	-0.511
IMAGENET-100	-0.022	-0.558
IMAGENET-10	0.249	-0.457
IMAGENET-5	0.51	-0.338
IMAGENET-2	0.425	-0.179

by ZHAW Zurich University of Applied Sciences. We are grateful to Frank P. Schilling for his valuable inputs.



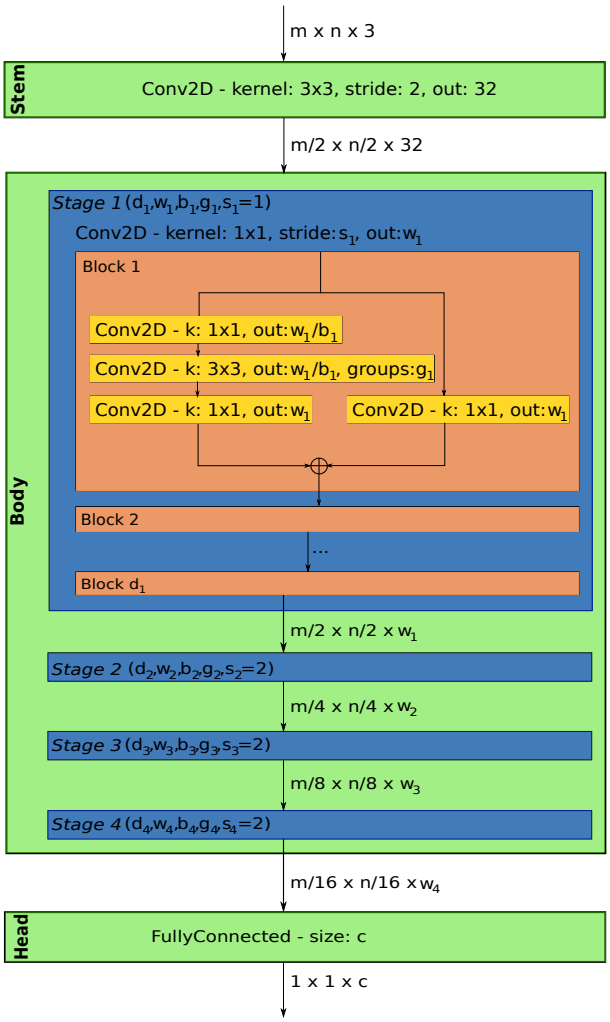


Figure 2a. **Figure 2b.**

Figure 2. **(A)** Example images from each dataset. Images of Cifar10/100 are magnified fourfold, the rest are shown in their original resolution (best viewed by zooming into the digital document). **(B)** The structure of models in the AnyNetX design space, with a fixed stem and a head, consisting of one fully-connected layer of size c, (where c is the number of classes). Each stage i of the body is parametrised by d_i, w_i, b_i, g_i, the strides of the stages are fixed with $s_1 = 1$ and $s_i = 2$ for the remainder.

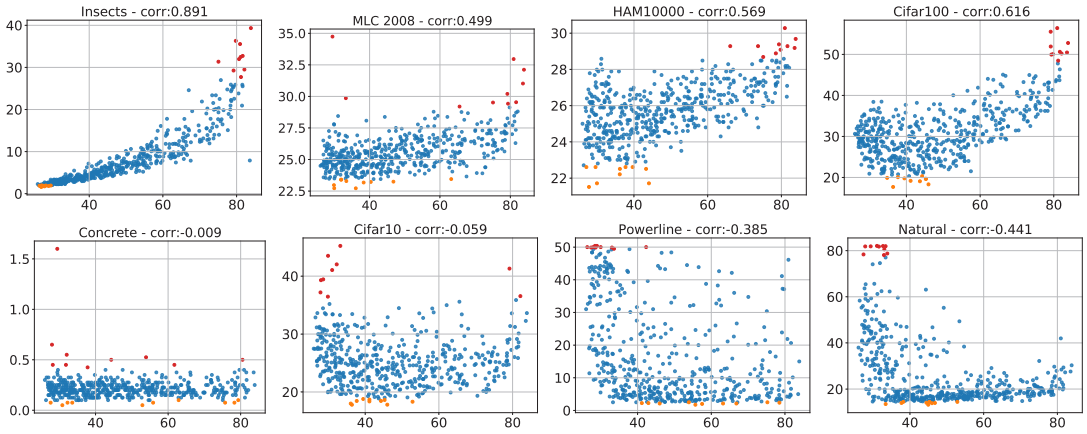


Figure 3. Test errors of all 500 sampled architectures on target datasets (y-axis) plotted against the test errors of the same architectures (trained and tested) on ImageNet (x-axis). The top 10 performances on the target datasets are plotted in orange and the worst 10 performances in red.

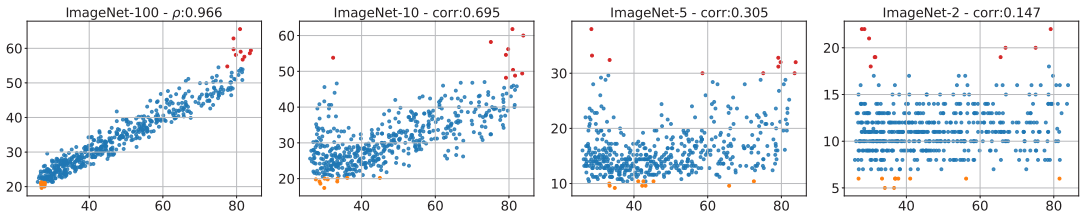


Figure 4. Error of all 500 sampled architectures on subsampled (by number of classes) versions of ImageNet (y-axis) plotted against the error of the same architectures on regular ImageNet (x-axis). The top 10 performances on the target dataset are plotted in orange and the worst 10 performances in red.

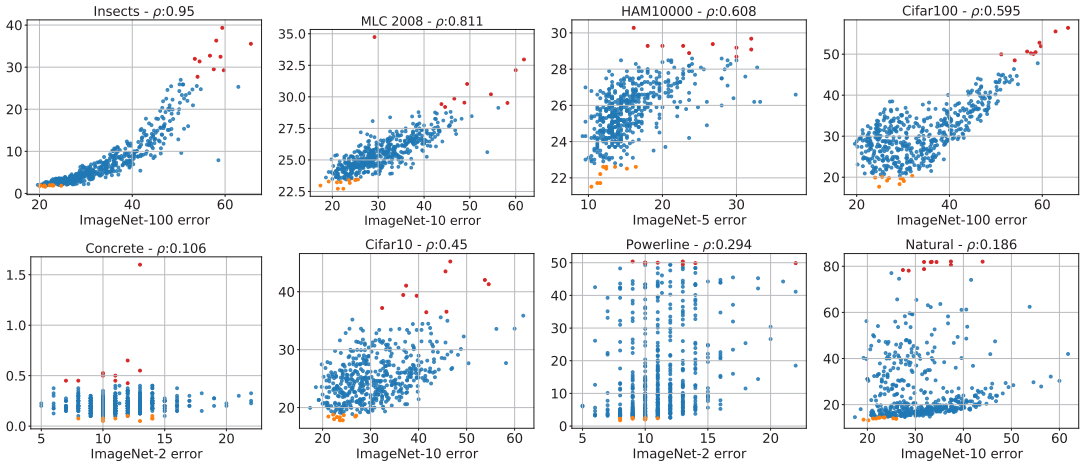


Figure 5. Test errors of all 500 sampled architectures on target datasets (y-axis) plotted against the test errors of the same architectures on the ImageNet-X (x-axis). The top 10 performances on the target dataset are orange, the worst 10 performances red.

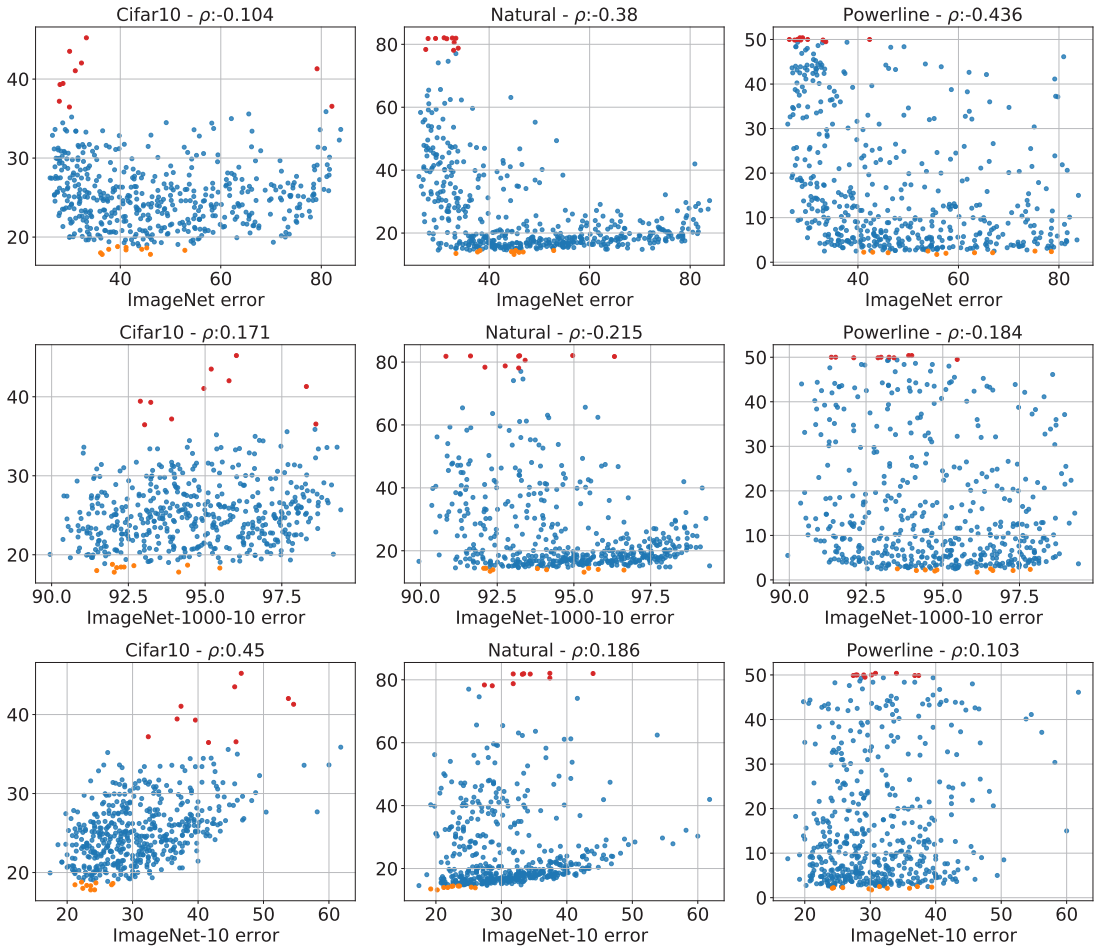


Figure 6. The errors of all 500 architectures on Cifar10, Natural, and Powerline plotted against the errors on ImageNet (top row), ImageNet-1000-10 (middle row) and ImageNet-10 (bottom row). We observe that class-wise downsampling has the largest positive effect on error correlation.

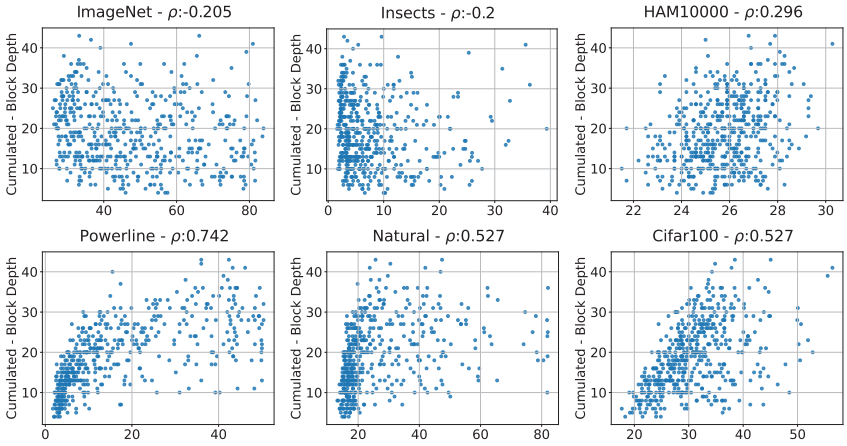


Figure 7a. depths

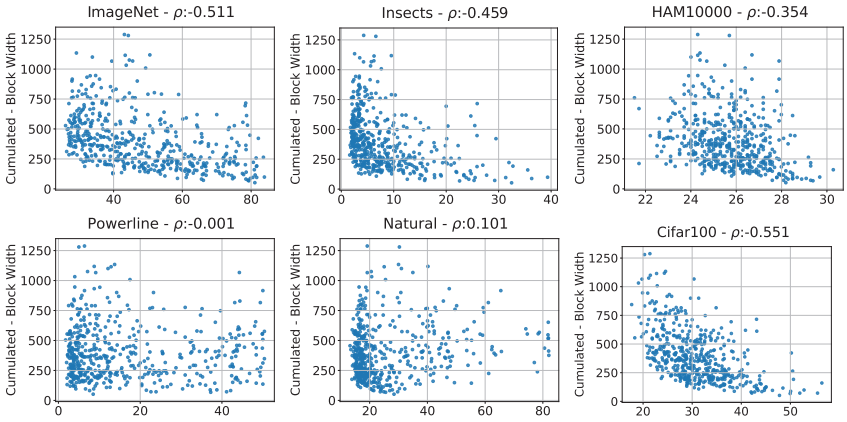


Figure 7b. widths

Figure 7. Errors of all 500 sampled architectures on ImageNet, Insects, HAM10000, Powerline, Natural, and Cifar100 (x-axis) plotted against the cumulative block **(A)** *depths* and **(B)** *depths* (y-axis).

REFERENCES

Bansal, P. (2018). Intel image classification

Beijbom, O., Edmunds, P. J., Kline, D. I., Mitchell, B. G., and Kriegman, D. J. (2012). Automated annotation of coral reef survey images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society), 1170–1177

Chen, Y., Yang, T., Zhang, X., Meng, G., Xiao, X., and Sun, J. (2019). Detnas: Backbone search for object detection. In *32th Annual Conference on Neural Information Processing Systems*, eds. H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett. 6638–6648

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society), 1800–1807. doi:10.1109/CVPR.2017.195

Ciresan, D. C., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society), 3642–3649

Collins, J., Sohl-Dickstein, J., and Sussillo, D. (2017). Capacity and trainability in recurrent neural networks. In *5th International Conference on Learning Representations* (OpenReview.net)

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. (2017). Sharp minima can generalize for deep nets. In *34th International Conference on Machine Learning*, eds. D. Precup and Y. W. Teh (PMLR), 1019–1028

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., et al. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *31th International Conference on Machine Learning* (JMLR.org), 647–655

Freedman, D., Pisani, R., and Purves, R. (2007). Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Trans. Neural Networks Learn. Syst.* 28, 2222–2232

[Dataset] Hansen, O. L. P., Svenning, J.-C., Olsen, K., Dupont, S., Garner, B. H., Iosifidis, A., et al. (2019). Image data used for publication "Species-level image classification with convolutional neural network enable insect identification from habitus images "

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society), 770–778

Hestness, J., Narang, S., Ardalani, N., Diamos, G. F., Jun, H., Kianinejad, H., et al. (2017). Deep learning scaling is predictable, empirically. *CoRR* abs/1712.00409

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR* abs/1704.04861

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society), 7132–7141. doi:10.1109/CVPR.2018.00745

Jing, L. and Tian, Y. (2019). Self-supervised visual feature learning with deep neural networks: A survey. *CoRR* abs/1902.06162

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., et al. (2020). Scaling laws for neural language models. *CoRR* abs/2001.08361

Kawaguchi, K., Kaelbling, L. P., and Bengio, Y. (2017). Generalization in deep learning. *CoRR* abs/1710.05468

Kornblith, S., Shlens, J., and Le, Q. V. (2019). Do better imagenet models transfer better? In *2019 IEEE Conference on Computer Vision and Pattern Recognition* (Computer Vision Foundation / IEEE), 2661–2671

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *26th Annual Conference on Neural Information Processing Systems*, eds. P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. 1106–1114

Liu, H., Simonyan, K., and Yang, Y. (2019). DARTS: differentiable architecture search. In *7th International Conference on Learning Representations* (OpenReview.net)

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society), 3431–3440. doi:10.1109/CVPR.2015.7298965

Melis, G., Dyer, C., and Blunsom, P. (2018). On the state of the art of evaluation in neural language models. In *6th International Conference on Learning Representations,* (OpenReview.net)

Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. (2018). Sensitivity and generalization in neural networks: an empirical study. In *6th International Conference on Learning Representations* (OpenReview.net)

Özgenel, Ç. F. and Sorguç, A. G. (2018). Performance comparison of pretrained convolutional neural networks on crack detection in buildings. In *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction* (IAARC Publications), 1–8

Radosavovic, I., Johnson, J., Xie, S., Lo, W., and Dollár, P. (2019). On network design spaces for visual recognition. In *International Conference on Computer Vision* (IEEE), 1882–1890

Radosavovic, I., Kosaraju, R. P., Girshick, R. B., He, K., and Dollár, P. (2020). Designing network design spaces. In *2020 IEEE Conference on Computer Vision and Pattern Recognition*. 10425–10433

Ramachandran, P., Zoph, B., and Le, Q. V. (2018). Searching for activation functions. In *6th International Conference on Learning Representations Workshop Track Proceedings* (OpenReview.net)

Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society), 512–519

Real, E., Aggarwal, A., Huang, Y., and Le, Q. V. (2019). Regularized evolution for image classifier architecture search. In *The Thirty-Third AAAI Conference on Artificial Intelligence* (AAAI Press), 4780–4789

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do imagenet classifiers generalize to imagenet? In *36th International Conference on Machine Learning*, eds. K. Chaudhuri and R. Salakhutdinov (PMLR), 5389–5400

Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., and Shavit, N. (2020). A constructive prediction of the generalization error across scales. In *8th International Conference on Learning Representations* (OpenReview.net)

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2014). Imagenet large scale visual recognition challenge. *CoRR* abs/1409.0575

Shihavuddin, A. S. M., Gracias, N., García, R., Gleason, A. C. R., and Gintert, B. (2013). Image-based coral reef classification and thematic mapping. *Remote. Sens.* 5, 1809–1841

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations*, eds. Y. Bengio and Y. LeCun

Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Highway networks. *CoRR* abs/1505.00387

Stadelmann, T., Amirian, M., Arabaci, I., Arnold, M., Duivesteijn, G. F., Elezi, I., et al. (2018). Deep learning in the wild. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition* (Springer), 17–38

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., et al. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society), 1–9

Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. In *2011 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society), 1521–1528

Torralba, A., Fergus, R., and Freeman, W. T. (2008). 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 1958–1970

Tschandl, P., Rosendahl, C., and Kittler, H. (2018). The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *CoRR* abs/1803.10417

Tuggener, L., Amirian, M., Benites, F., von Däniken, P., Gupta, P., Schilling, F.-P., et al. (2020). Design patterns for resource-constrained automated deep-learning methods. *AI* 1, 510–538

Xie, S., Girshick, R. B., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society), 5987–5995

Yetgin, Ö. E., Gerek, Ö. N., and Nezih, Ö. (2017). Ground truth of powerline dataset (infrared-ir and visible light-vl). *Mendeley Data* 8

Zamir, A. R., Sax, A., Shen, W. B., Guibas, L. J., Malik, J., and Savarese, S. (2019). Taskonomy: Disentangling task transfer learning. In *International Joint Conference on Artificial Intelligence 2019*, ed. S. Kraus (ijcai.org), 6241–6245. doi:10.24963/ijcai.2019/871

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *13th European Conference on Computer Vision, Proceedings, Part I*, eds. D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Springer), 818–833

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations* (OpenReview.net)

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society), 8697–8710

6 APPENDICES

6.1 Verifying the numerical robustness of our study

This Chapter we present additional studies designed to test for possible flaws or vulnerabilities in our experiments. We conduct these to further strengthen the empirical robustness of our results.

6.1.1 Stability of Empirical Results on Cifar10

The top-1 errors of our sampled architectures on Cifar10 lie roughly between 18 and 40, which is fairly poor, not only compared to the state of the art but also compared to performance that can be achieved with fairly simple models. This calls into question if our Cifar10 results are flawed in a way that might have lead us to wrong conclusions. We address this by running additional tests on Cifar10 and evaluate their impact on our main results. We get a goalpost for what performance would be considered good with our style of neural network and training setup by running the baseline code for Cifar10 published by Radosavovic et al.

Radosavovic et al. (2020). Table 5 shows that these baseline configurations achieve much lower error rates. We aim to improve the error results on Cifar10 in two ways: First we train our architecture population with standard settings for 200 epochs instead of 30, second we replaced the standard network stem with one that is specifically built for Cifar10, featuring less stride and no pooling. Figure 8 shows scatterplots of the errors from all 500 architectures on Cifar10 against the errors on ImageNet and ImageNet-10. We can see that both new training methods manage to significantly improve the performance with a minimum top-1 error below 10 in both cases. More importantly can we observe that both new training methods have, despite lower overall error, a very similar error relationship to ImageNet. The error correlation is even slightly lower than with our original training (replicated in Figure 8 left row). We can also see that in all three cases the error relationship can be significantly strengthened by replacing ImageNet with ImageNet-10, *this shows that tuning for individual performance on a dataset does not significantly impact the error relationships between datasets which further strengthens our core claim.*

6.1.2 Verifying Training Duration

Since we have a limited amount of computational resources and needed to train a vast number of networks we opted to train the networks up to the number of epochs where they started to saturate significantly in our pre-studies. As we have seen in section 6.1.1 can the network performance still improve quite a bit if it is trained for much longer. Even though the improved performances on Cifar10 did not yield any results contradicting the findings of our study, we still deemed it necessary to closer inspect what happened in the later stages of training and thus performed a sanity check for Cifar10 as well as the other two datasets that show a negative error correlation with ImageNet—Powerline and Natural. Figure 9a shows the Cifar10 test error curves of 20 randomly selected architectures over 200 epochs. On the left side we see the same curves zoomed in to epochs 30 to 200. We see that the error decreases steadily for all architectures, the ranking among architectures barely changes past epoch 30. The relative performance between architectures and not absolute error rates are relevant for our evaluations, we can therefore conclude that the errors at epoch 30 are an accurate enough description of an architecture's power.

For Powerline and Natural, we select the five best and five worst architectures respectively and continue training them for a total of five times the regular duration. Figure 9b shows the resulting error curves. Both datasets exhibit minimal changes in the errors of the top models. On Natural we observe clear improvements on the bottom five models but similar to Cifar10 there are very little changes in terms of relative performance. Powerline exhibits one clear cross-over but for the remainder of the bottom five models the ranking also stays intact. *Overall we can conclude that longer training does not have a significant effect on the APR of our datasets.*

6.1.3 Impact of Training Variability

The random initialization of the model weights has an effect on the performance of a CNN. In an empirical study it would therefore be preferable to train each model multiple times to minimize this variability. We opted to increase the size of our population as high as our computational resources allow, this way we get a large number of measurements to control random effects as well as an error estimate of a large set of architectures. However, we still wanted to determine how much of the total variability is caused by training noise and how much is due to changing the architectures. We estimate this by selecting two of the sampled CNN designs, number 147 performing slightly above average with an error of $e_{147} = 11.9$ and number 122 performing slightly below average with $e_{122} = 14.5$. The quantiles of the error distribution from all 500 architectures are $q_{0.25} = 11.53$, $q_{0.5} = 13.02$ and $q_{0.75} = 15.46$ with an overall mean of $\mu = 13.9$. We then train the architectures 147 and 122 each 250 times. Figure 10 shows the error distributions of both selected

architectures as well as the overall distribution obtained from training each of the 500 architectures once. There is of course some variability within both architectures but both individual architectures produce very narrow densities and show essentially no overlap. *We can therefore conclude that the effect of choosing an architecture is much greater than the variability caused by random training effects.*

6.1.4 Relationship of Top-1 with Top-5 Error on ImageNet, Insects and Cifar100

We opted to use top-5 error since it is the most widely reported metric for ImageNet and the top-5 numbers are therefore easy to interpret on that dataset. Many of our datasets have a significantly lower number of classes such that top-5 error makes little sense and we opted to use top-1 for those. This raises the question if comparing top-1 with top-5 errors introduces unwanted perturbations into our analysis. We therefore compare the top-1 and top-5 errors for the three datasets on which we use top-1 error (see Figure 11a). We see that the two metrics have an almost linear relationship for the ImageNet and Cifar100 datasets. More importantly are the top-1 to top-5 error graphs monotonically ascending for all three datasets, such that the ordering of architectures does not change when swapping between the two metrics. *Since we are interested in the relative performances of our sampled architectures changing between top-1 and top-5 error does not impact our analysis.*

6.1.5 Overfitting of High-Capacity Architectures

The best architectures on Powerline, Natural and Cifar100 have a very small cumulated depth, so it is only natural to ask if the deeper architectures perform poorly due to overfitting. We address this concern by plotting the *training errors* of Powerline, Natural, and Cifar100 against the cumulative block depths (see Figure 11b). The training errors are strongly correlated with the cumulative block depth, just like the test errors. *Plots of the cumulated block depth show almost the same structure for training and test errors. We can therefore exclude overfitting as a reason why the shallower networks perform better on Powerline, Natural, and Cifar100.*

6.1.6 Impact of Class Distribution

MLC2008 and HAM1000 have a strong class imbalance. They both have one class which makes up a large amount of the dataset. In order to study the impact of an imbalanced class distribution, we created two new more balanced datasets out of the existing data the following way: we reduced the number of samples in the overrepresented class such that it has the same amount of samples as the second most common class. We call these datasets MLC2008-balanced and HAM10000-balanced. Their new class distributions can be seen in Figure 11a. We train our architecture population on MLC2008-balanced and HAM10000-balanced leaving the training configuration otherwise unaltered. Figure 11b shows the errors on the balanced datasets versus the errors on the unbalanced counterparts.

For both HAM10000 and and MLC2008, there is a strong correlation between the errors on the balanced and unbalanced datasets. *We can therefore conclude that class imbalance is not a determining factor for the APRs of HAM10000 or MLC2008.*

6.2 Additional ablation studies

6.2.1 Impact of Pretraining

The main objective of our study is to identify how well different CNN designs perform on varying datasets and if the best architectures are consistent across the datasets. For this reason we train all of our networks from scratch on each dataset. However, we cannot ignore that pretraining on ImageNet is a huge

Table 5. Top-1 error of reference network implementations Radosavovic et al. (2020) for Cifar10.

MODEL	RESNET-56	RESNET-110	ANYNET-56	ANYNET-110
ERROR	5.91	5.23	5.68	5.59

factor in practice and we therefore study its impact on our evaluations. To this end have we train all of our sampled architectures again on each dataset but this time we initialize their weights with ImageNet pretraining (we omit Concrete, which has very low errors even without pretraining). Figure 12 shows the errors of each dataset without (blue) and with (green) pretraining plotted against the ImageNet errors. The data shows a distinct trend: the overall performance improvement due to pretraining dictates how much stronger the ImageNet-correlation of the pretrained errors is compared to the errors without pretraining. For Cifar10 and Cifar100 where the performance gain with pretraining is low to moderate the error correlations do not drastically change. On the other end of the spectrum are Natural and Powerline, where pretraining leads to drastically lower errors. This in turn leads to much higher error correlation with ImageNet(the Powerline correlation can not grow significantly above 0 because the overall errors are so small across all architectures). *We can conclude that our findings are still valid when pretraining is used, but their effects can be masked when pretraining is the most important factor contributing to the overall final performance.*

6.2.2 Structure of Top Performing Architectures

Figure 13 shows the configuration of the top performing architecture in blue, as well as the mean and standard deviation of the top 15 configurations for every dataset. We observe that the top 15 architectures have very high variance in both bottleneck ratio and group width.

Block width on the other hand shows a clear pattern: almost all high-performing architectures start with a very small block width that increases across the stages. Only Powerline and Natural do not show this pattern. In block depth, we observe a similar pattern with a bit more noise. For block depth, Powerline, Natural, Cifar10 and Cifar100, no such trend of increased parameter values towards the later stages is observed. *This reinforces the idea that block width and block depth greatly impact an architectures performance and their optimal choices are dataset dependent.*

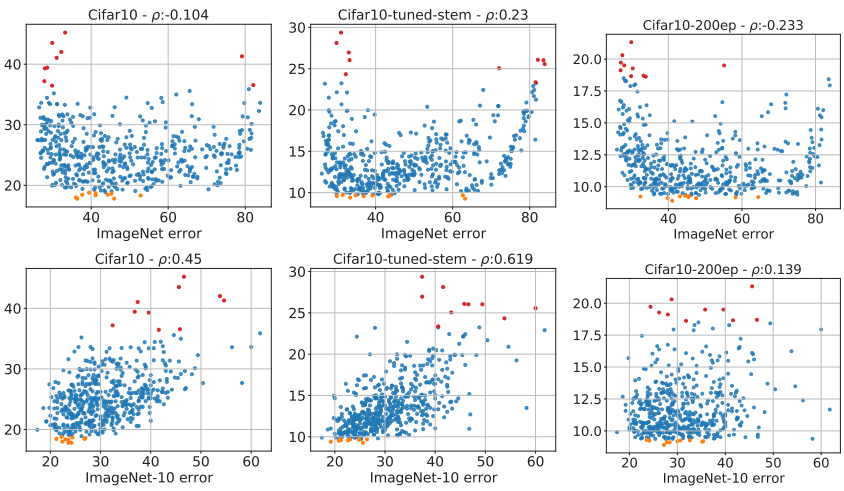


Figure 8. The Cifar10 test errors of all 500 architectures plotted against ImageNet (top row) and ImageNet-10 (bottom row), shown for our original Cifar10 training (left column), training with a Cifar10 specific stem in the architecture (middle column), and training for 200 epochs, which is roughly 6 times longer (right column). The plots show that the error correlation with ImageNet-10 is much larger in all three cases, confirming that optimizing for individual Cifar10 performance does not alter our core result.

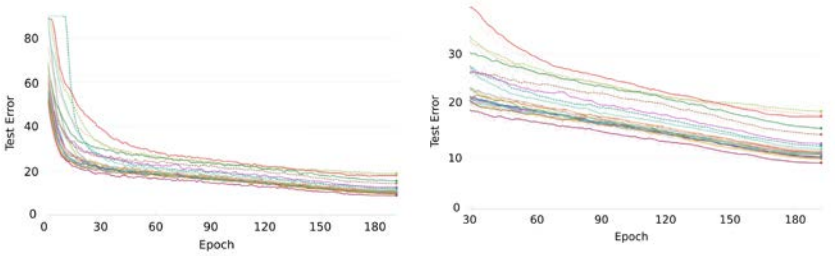


Figure 9a.

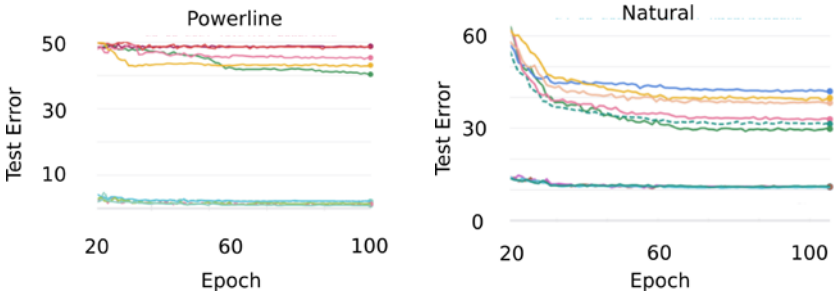


Figure 9b.

Figure 9. **(A)** Cifar10 test error curves of 20 randomly sampled architectures trained over 200 epochs (left). The same error curves but cut to epochs 30 to 200. **(B)** Test error curves of the five best and five worst models on Powerline and Natural, respectively, when training is continued to epoch 100

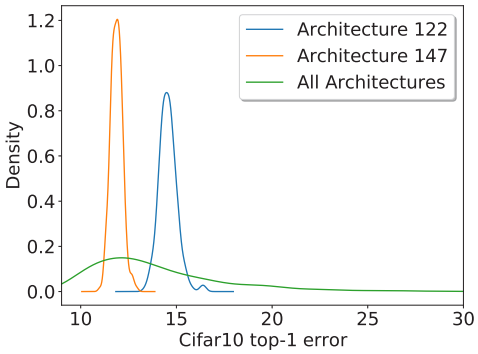


Figure 10. Error distributions on Cifar10 of two architectures (122, 147) both trained from scratch 250 times as well as the Cifar10 error distribution of all 500 architectures. The plot shows that the variability caused by changing architecture is much larger than the one caused by random training effects.

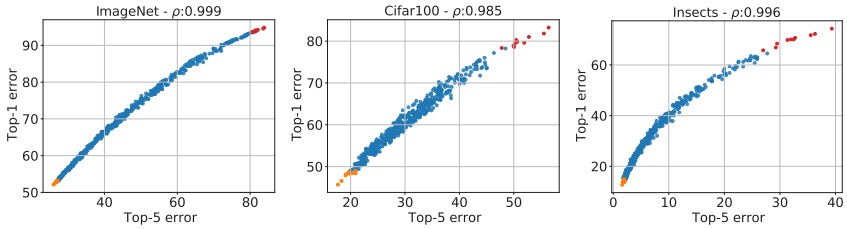


Figure 11a.

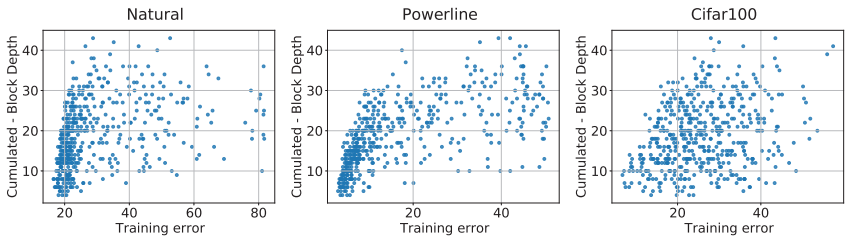


Figure 11b.

Figure 11. **(A)** Top-1 error plotted against top-5 error of all 500 architectures on ImageNet, Cifar100, and Insects. The plots reveal that on all three datasets the errors have a very close relationship: it is not perfectly linear but is monotonically ascending **(B)** Training errors of the sampled architectures (x-axis) plotted against the cumulated block *depth* for the 3 datasets that have the lowest test errors on shallow architectures. We observe that for all three datasets shallow architectures also have the lowest training errors. Therefore overfitting is not the cause of this behaviour.

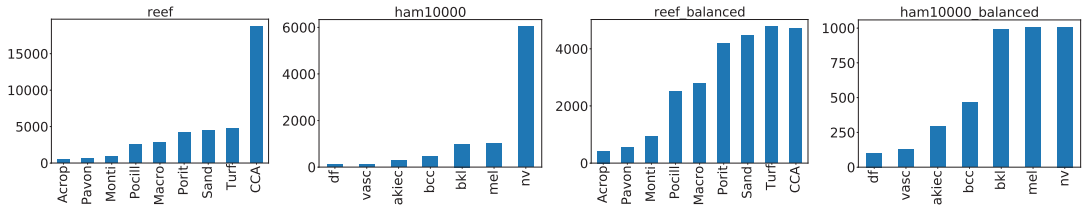


Figure 11a.

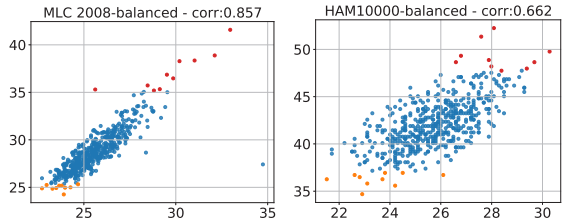


Figure 11b.

Figure 11. **(A)** Class distributions of MLC2008, HAM10000, and their balanced versions. **(B)** Errors of all 500 sampled architectures on MLC2008-balanced and HAM1000-balanced (y-axis) plotted against the errors of their unbalanced counterparts (x-axis). The top 10 performances on the target dataset are plotted in orange, the worst 10 performances in red. We observe a clear positive correlation for both datasets, hence we conclude that the dataset imbalance has a limited impact on the APRs.

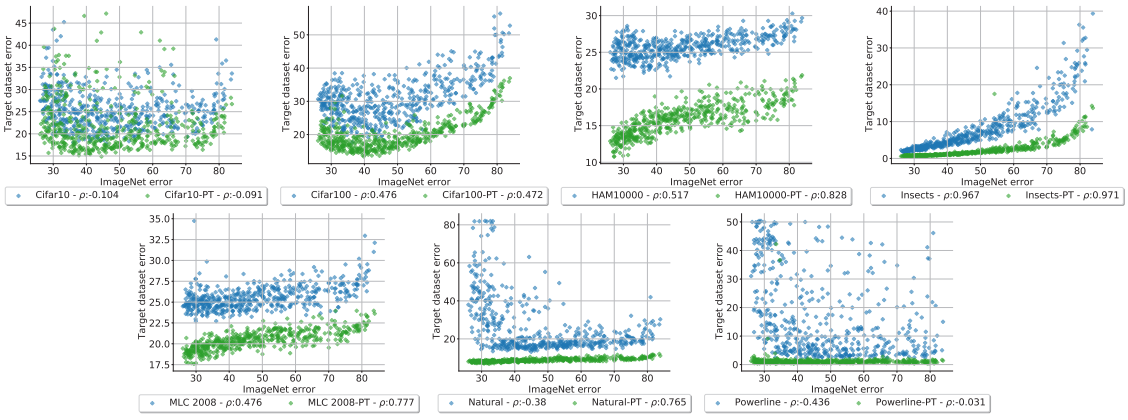


Figure 12. Errors form all 500 architectures trained from scratch (blue) as well as the same architectures pretrained on ImageNet (green), plotted against the respective ImageNet errors. We observe that the error correlation with ImageNet increases relative to the performance gain due to pretraining.

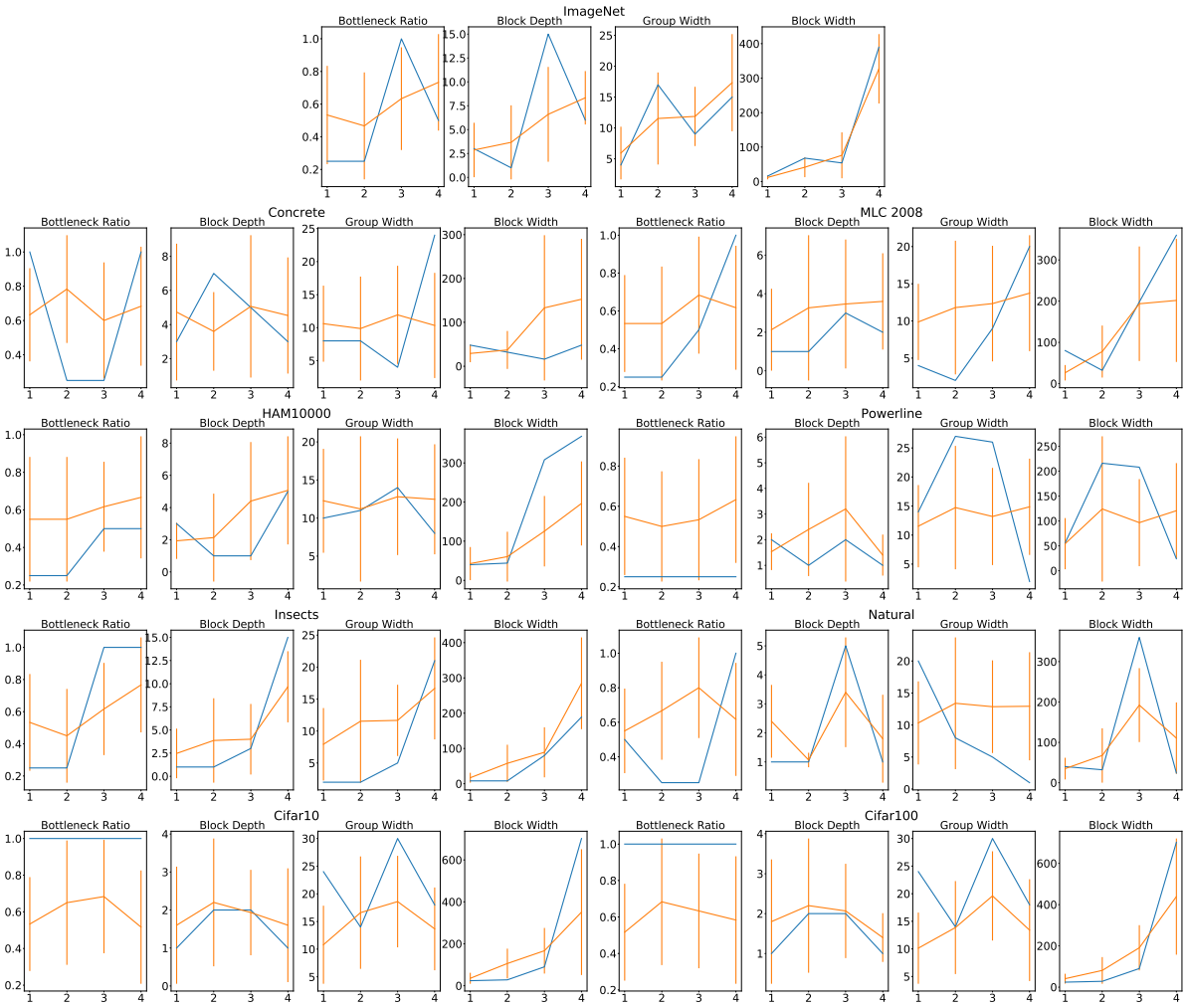


Figure 13. Configurations of the top-performing architectures, with the four stages depicted on the x-axis and the parameter values on the y-axis. The best architectures are shown in blue, the mean of the top 15 architectures is depicted in orange with with a vertical indication of one standard deviation.

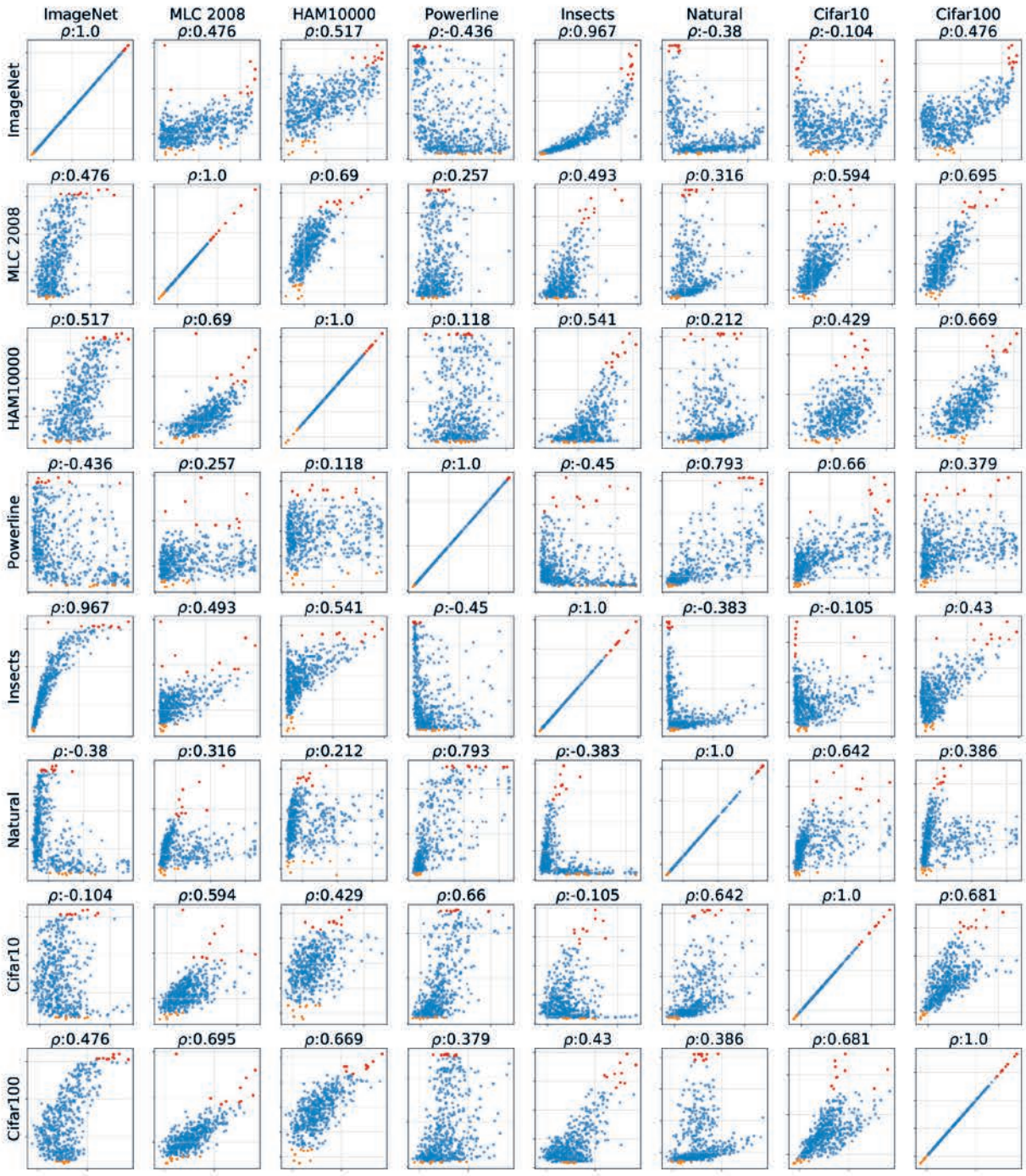


Figure 14. Matrix of error scatterplots of all datasets except Concrete (The first row replicates plots shown in Figure 3).

4 Research Output of the Explainable Artificial Intelligence Group

4 Research Output of the Explainable Artificial Intelligence Group

The XAI group, led by Jasmina Bogojeska, conducts research in machine learning and deep learning methodology, particularly explainable and multimodal AI, to address complex decision-making and knowledge discovery tasks in different domains. The ultimate goal is to enable successful application of these methods in products in practice hand-in-hand with domain experts and users.

We are primarily interested in the challenging problem of developing AI-powered systems able to properly utilise multimodal data, provide human intelligible information about their outputs and engage seamlessly with users. This is essential for the successful practical application and adoption of AI in many domains. Reinforcement learning and causal inference are two additional areas of interest very relevant for sequential decision making and knowledge discovery with human in the loop. Finally, while we are up for taking on challenging problems in various domains, we are particularly interested in advancing, improving, and digitising the healthcare domain. We envision this by building explainable AI-powered products for disease diagnosis and treatment as well as developing transparent AI-powered approaches to advance the understanding of health and disease, all done collaboratively with users (experts and patients) in the loop, safely and responsibly.

The group was newly created end of 2022 and currently has one senior lecturer (Dr. Jasmina Bogojeska) who joined ZHAW in September 2022 working on building it up. It is set to grow further in the coming years.

We plan to start the journey of the group by acquiring projects in the healthcare domain. The availability of large, complex medical and biological datasets coupled with the technological and methodological advances in AI in recent years provides the opportunity to advance the understanding of health and disease and pave the way to revolutionize healthcare. We envision working on novel trustworthy AI methodologies able to leverage the rich information from complex, multimodal, multi-source medical data and able to engage seamlessly with their users. Such methods have the potential to improve patient care via early, precise, personalized disease diagnosis, treatment, and prevention. We plan to follow this research direction and collaborate with partners from hospitals, medical research institutions and industry with the final goal of unlocking the power of AI methodology functioning collaboratively with the users in clinical practice for optimizing patient care and disease prevention.

The group had one publication in 2022 on deep learning for NLP where the usefulness of pre-trained transformer document embeddings in active learning for multilabel classification was investigated. Furthermore, Dr. Bogojeska's prior work on deep learning for NLP, e.g., data augmentation for low resource domain specific text classification, was also published in two prestigious conferences in 2022.

We thank CAI and ZHAW for this great opportunity and the continuous support!

The XAI 2022 team

Jasmina Bogojeska (along with one Master student)

Evaluating Pre-Trained Sentence-BERT with Class Embeddings in Active Learning for Multi-Label Text Classification

Lukas Wertz
University of Stuttgart
lukas.wertz@ims.uni-stuttgart.de

Jasmina Bogojeska
Zurich University of Applied Sciences
bogo@zhaw.ch

Katsiaryna Mirylenka
IBM Research – Zurich
kmi@zurich.ibm.com

Jonas Kuhn
University of Stuttgart
jonas.kuhn@ims.uni-stuttgart.de

Abstract

The Transformer Language Model is a powerful tool that has been shown to excel at various NLP tasks and has become the de-facto standard solution thanks to its versatility. In this study, we employ pre-trained transformer document embeddings in an Active Learning task to group samples with the same labels in the embedding space on domain-specific corpora. We find that the calculated class embeddings are not close to the respective samples and consequently do not partition the embedding space in a meaningful way. In addition, using the class embeddings as an Active Learning strategy yields reduced results compared to all baselines.

1 Introduction

While text classification models have become more and more powerful, the need for sufficient data to train ever growing neural networks is also increasing massively. When dealing with domain-specifc data, such as legal or medical in particular, finding a fitting dataset with detailed annotations can be exceedingly difficult. Creating such a dataset is likely to be a massive undertaking due to the difficult annotation process which often requires domain experts to work through enormous amounts of data. Active Learning serves as a way to speed up this process by selecting informative samples to be annotated. However, Active Learning strategies are often very specific to target domains (Wertz et al., 2022) and strategies tailored specifically for pre-trained transformer language models are often experimental and not thoroughly explored (Zhan et al., 2022).

In this work, we present an Active Learning strategy that employs class embeddings which are generated from pre-trained sentence embeddings to predict the classes of unlabeled samples. While the intuition of the approach is sound, we find that the class embeddings do not generalize from the samples they were calculated on. Our experiment focuses on powerful pre-trained, transformer sentence-embeddings which are prevalent in both research and industrial application. We demonstrate that such embeddings struggle to find good separations between the multi-class, multi-label texts in the training set on two domain-specific datasets. Our work details the class embedding approach, illustrates the reduced performance on two domain-specific, multi-label datasets and analyses the vector space of the samples to gain an understanding of the methods failure.

2 Related Work

The effectiveness of AL for Text Classification has been subject to extensive research (Tong and Koller, 2001), (Goudjil et al., 2018) with specific solutions for deep models (Schröder and Niekler, 2020), (An et al., 2018) and multi-label settings (Reyes et al., 2018) (Yang et al., 2009). Our approach targets Active Learning for Deep Learning which poses new challenges (Schröder and Niekler, 2020) and is still a topic in need of exploration (Ein-Dor et al., 2020). Generating embeddings from words has been performed with trained vector models (Church, 2017) (Pennington et al., 2014) but has been moved to the contextual embedded information within large transformer language models such as BERT (Devlin et al., 2018). Extracting embeddings across word boundaries from BERT can be done in several ways, such as a grid-based approach (Denk and Reisswig, 2019), a "siamese" dual network architecture (Reimers and Gurevych, 2019) or unsupervised techniques (Zhang et al., 2020).

3 Class Embeddings

3.1 Intuition

In any text classification task, the aim is to identify the belonging of a text T to a range of pre-defined classes C. Using pre-trained language models, a

text classification model M decides the class $c \epsilon C$ using only the tokenized text as input, leveraging the powerful pre-trained weights of the underlying transformer network as information. We can thus assume that the surface tokens are the critical information that determine, what class T is assigned.
One option to represent text in a continuous vector space is via *embeddings* - vectors that are conditioned to correspond to pieces of text. We convert T into the vector space via embeddings (T_e). Intuitively, one would assume that T_e which belong to the same c are also closer together in the vector space. After all, if c is mainly decided based on the surface tokens, it follows that there should be either syntactical or semantical similarity between two T both belonging to c. While semantical similarity is much harder to capture than the surface realisation of language, current text embedding techniques have shown to also be sensible to word meaning (Wiedemann et al., 2019).
In conclusion, we expect T that belong to the same class to be closer together in a fitting vector space representation because their text should show similarities. Consequently, we assume that if a new text T^* is mapped into the same vector space, it is more likely to belong to the same classes as its neighbours. As such, the centroid of a set of T_e can be used to predict the class of said T^*.

3.2 Active Learning with Class Embeddings

$$C_e = \{mean(T_e) | T \epsilon D \text{ and } T \text{ belongs to } c\} \quad (1)$$

Active Learning is a cyclic, supervised learning mechanism that seeks to reduce annotation effort by strategically selecting informative samples to be labeled by a human annotator and then given to the model for training. Given an annotated training set D and an unlabeled set U, the main loop of Active Learning can be summarized in three repeating steps:
1. Train classification model M on available data D.
2. Select informative samples from U and pass them to the annotator.
3. Annotate the samples and add them to D.
Given an annotated set D, our approach calculates **Class Embeddings** C_e for each class c by first collecting all T that belong to c and then using an embedding technique to map T into the vector space. The corresponding $c_e \epsilon C_e$ are determined by calculating the centroid of all T_e belonging to c (Equation (1)).

	train	dev	test	Macro F1
eurlex	10.294	1.901	1.905	0.93
arXiv	13.174	13.414	13.131	0.79

Table 1: Split sizes and Macro F1 on the full *eurlex* and *arXiv* datasets.

In the Active Learning setting, we calculate C_e given the current D and then select k samples which are close to the c_e of classes that are less frequent in the training set. The idea is, that finding samples of less represented classes will improve classifier accuracy on that class and consequently, will improve Macro F1. We update and evaluate M after k samples have been selected and repeat this process until an annotation budget is exhausted. The full procedure is detailed in Algorithm 1.

Algorithm 1 Active Learning with Class Embeddings

procedure CE(labeled set D, unlabeled set U, model M, budget b, sample size k)
  while budget > 0 **do**
    train M on D
    $C_e \leftarrow$ Class Embeddings on D
    $k^* \leftarrow k$
    while $k^* > 0$ **do**
      $c_{min} \leftarrow$ least frequent class in D
      $T \leftarrow T \epsilon U, T$ closest to c_e of c_{min}
      annotate T
      $D \leftarrow D \cup T$
      $k^* \leftarrow k^* - 1$
      $b \leftarrow b - 1$

4 Experiment

4.1 Datasets

We use modified versions of the Eurlex57K (referred to as *eurlex*) (Chalkidis et al., 2019) corpus containing excerpts from European law as well as a collection of abstracts from scientific publication site *arXiv* (`https://www.kaggle.com/Cornell-University/arxiv`). Both datasets are annotated with several hundred classes and are intended for large-scale, multi-label text classification, meaning that a sample can belong to any number of classes instead of only one. We reduce the number of classes to 5 frequent and 5 rare labels to create a reduced version of the corpus, keeping the multi-label nature intact. Macro F1 when using the full dataset is found in Table 1.

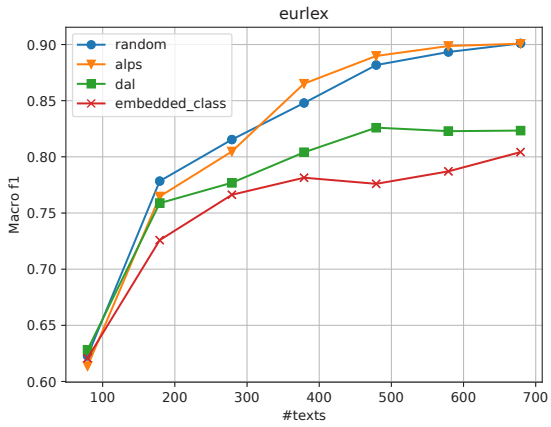


Figure 1: Macro F1 on the *eurlex* dataset of Active Learning for training set sizes 100 to 600 samples compared to random selection and two Active Learning baselines.

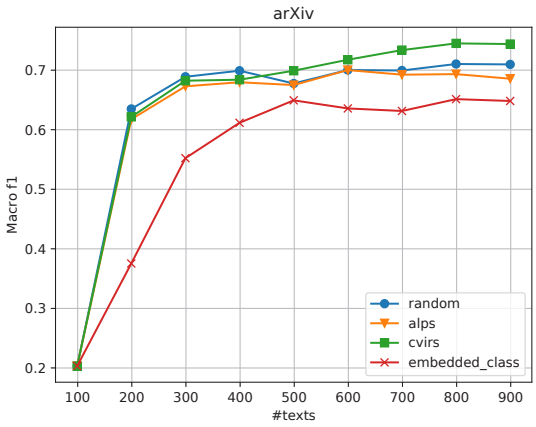


Figure 2: Macro F1 on the *arXiv* dataset of Active Learning for training set sizes 100 to 600 samples compared to random selection and two Active Learning baselines.

4.2 Setup

We use BERT (Devlin et al., 2018)* for text classification with a single, feed-forward output layer. We train the model for 15 epochs with early stopping, a batch size of 16 and an adaptive learning rate (ADAM). We evaluate all experiments using the multi-class measures Macro F1† (averaging F1 for each class, thus, treating each class as equally important, which is beneficial in the unbalanced class settings).

For document embeddings, we employ pre-trained Sentence-Bert (Reimers and Gurevych, 2019) embeddings‡ which maps a document into a 380 element vector.

We simulate Active Learning by using a subset of the corpus as "labeled" set and reserving the rest as the "unlabeled" set, using the oracle annotations once a sample is queried from the "unlabeled" set. We start with a labeled set of 100 randomly selected samples and query 50 samples in each Active Learning step until the annotation budget of 600 samples is exhausted.

All experiments are run on a NVIDIA RTX 6000 GPU.

4.3 Results

Figures 1 and 2 show the results of Active Learning on the *eurlex* and *arXiv* datasets respectively. We compare the class embedding approach (Section 3.2) against three Active Learning baselines (DAL - (Gissin and Shalev-Shwartz, 2019), ALPS - (Yuan et al., 2020), CVIRS - (Reyes et al., 2018)) as well as Active Learning by random sampling. Out of the Active Learning strategies, we report the two best performing approaches for each dataset. We find that the class embeddings perform significantly worse than all baselines by a margin of up to 0.15 compared to random selection. Class Embeddings appear to hinder the Active Learning process as they even perform worse than Active Learning strategies which already have reduced performance compared to random selection, i.e. the *DAL* baseline on the *eurlex* dataset.

5 Analysis

5.1 Proximity to unlabeled samples

One important assumption presented in Section 3.1 is, that an unlabeled§ sample $T^* \epsilon U$ will be close in the embedding space to the class embeddings $c_e \epsilon C_e$ of the classes $c \epsilon C$ it belongs to. We test this assumption by analysing how many T^* that belong to c are actually closest to the corresponding class embedding by querying the closest 100 T^* for every c_e. Table 2 shows, that on the *eurlex* dataset for a small labeled set with 100 samples, almost no T^* are near a c_e of a class they belong to. We also see that this is not an effect of the labeled set being too small as increases in the size of D (even to around 50% of the full training set) do not

*Using the "bert-base-uncased" model from *huggingface* https://huggingface.co

†We also evaluated Micro F1 but found that the two behaved similarly.

‡Using the "all-mpnet-base-v2" downloadable from https://www.sbert.net

§Here, *unlabeled* simply denotes that the sample does not come from the training set of the model (Section 4.2).

size of D	class 1	class 2	class 3	class 4	class 5	class 6	class 7	class 8	class 9	class 10
100	1	0	0	0	0	0	7	2	0	0
200	2	0	0	0	0	0	7	2	0	0
500	1	0	0	0	0	0	7	2	0	0
1500	2	0	0	0	0	0	7	1	0	0

Table 2: Number of samples in the unlabeled set U of the **eurlex** dataset with class *j* found within the closest 100 samples of the centroid of class *j* using pre-trained Sentence-BERT. We experiment with varying sizes of the labeled set D.

significantly change the results. Effectively, this means that the computed c_e are not close to new samples of the same class and that our assumption is incorrect. This observation holds for the *arxiv* dataset as well. (See Appendix for the full results table).

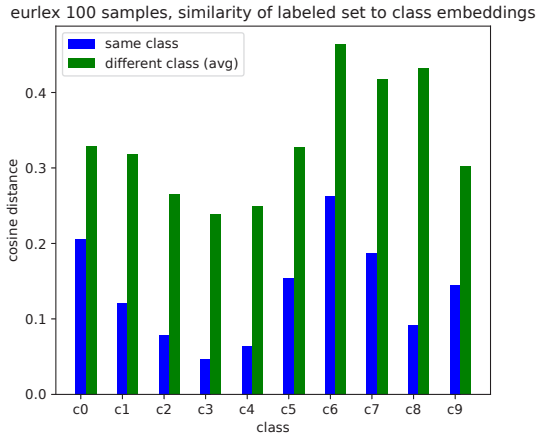


Figure 3: Average cosine distance between labeled samples and corresponding class embedding of the same class (blue, left) and averaged class embeddings of all other classes (green, right).

5.2 Examination of the labeled set

One explanation for the behaviour on unlabeled samples is, that the class embeddings are not well-positioned. For example, when calculating C_e we do not account for outliers which might cause a shift in the centroid. Alternatively, class embeddings might all be very close to each other, resulting in a partitioning that is not very meaningful. We run a sanity check in Figure 3 and Figure 4 and look at the average distance between samples in the labeled set $T \epsilon D$ and the computed class embeddings for a size of 100 samples[¶]. We find that on average, samples are closer to the c_e of classes

[¶]We also experiment with higher numbers but find no significant differences.

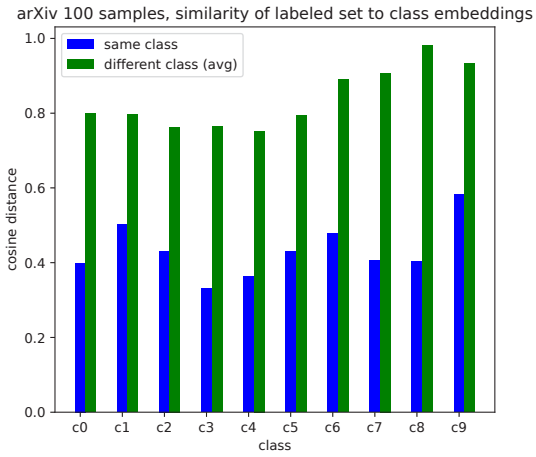


Figure 4: Average cosine distance between labeled samples and corresponding class embedding of the same class (blue, left) and averaged class embeddings of all other classes (green, right).

they belong to by a margin of around 0.2 on the *eurlex* dataset and 0.4 on the *arXiv* dataset. Due to the multi-label nature of the datasets we expect certain overlap between classes. Overall, Figures 3 and 4 seem to indicate a good positioning of the class embeddings, which means that the training set samples are in fact found in the proximity of corresponding class embeddings. Figures 5 and 6 show the result of a Principal Component Analysis (PCA) on the two datasets respectively. We find that while there are some clusters, overall there is no clear separation of classes. This could be an indication, that the sentence-BERT embeddings (see Section 4.2) are too large or too diverse to effectively decompose into 2 dimensions. However, it is also possible that even in the high-dimensional space, separation of the different classes is already difficult.

On the *eurlex* dataset, Figure 3 confirms this suspicion somewhat since the distance margins are narrow overall. We find that for many classes, observations hold between Figure 3 and Figure 5.

For example, samples belonging to class 2 have a are generally very close to their corresponding class embedding while Figure 5 also shows a narrow cluster of class 2 samples. However, for some samples we observe conflicting information from the two Figures, for example class 3, which has the least average distance in Figure 3 but is very spaced out in the PCA in Figure 5.
In general, the analysis of the *arxiv* dataset in Figures 4 and 6 leads to analogous conclusions. The main difference is that while the average distances in Figure 4 are twice as long as for the *eurlex* dataset, the samples in Figure 6 seem even more clustered around a central point. In general, most of the centroids are very close together in the reduced space, making clear separation of classes difficult. Overall, we can conclude that the class embeddings provide only limited grouping for the dataset they were calculated on.
In addition, we find that the labels have semantic overlap to each other. In the *arXiv* dataset, frequent labels deal with various areas of Physics, while rare labels deal with Computer Science and Informatics. On the *eurlex* dataset, frequent labels deal with Fruit, import and export while rare labels are more diverse. (Full Table is found in the appendix). This could explain the proximity of centroids in the PCA analysis, especially for the *arxiv* dataset in Figure 6. On the *eurlex* dataset in Figure 5 however, centroids of different topics, e.g. *Gaming* (centroid 9) and *Export Refund* (centroid 1) are close to each other.

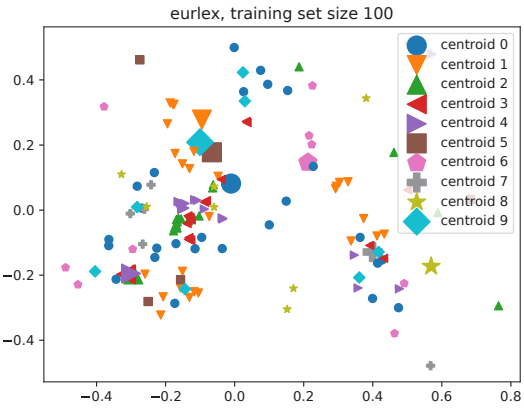


Figure 5: PCA with 2 components of the class embeddings and embedded samples in the training set with 100 samples. Shapes of the data points indicate class (samples with multiple classes are plotted multiple times) and enlarged data points mark centroids (i.e. class embeddings).

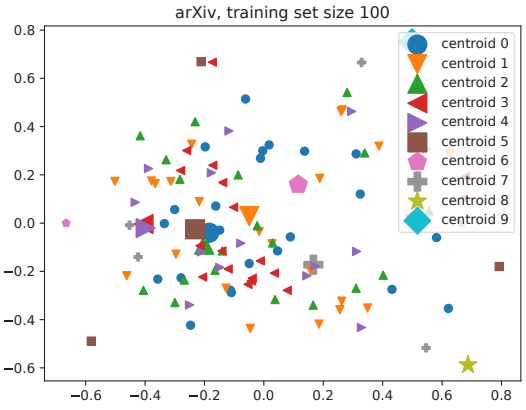


Figure 6: PCA with 2 components of the class embeddings and embedded samples in the training set with 100 samples. Shapes of the data points indicate class (samples with multiple classes are plotted multiple times) and enlarged data points mark centroids (i.e. class embeddings).

6 Conclusion & Future Work

We present Class Embeddings, which hinder the Active Learning (Section 4.3) since the classes of new samples can not be correctly predicted (Section 5.1). Despite reasonable assumptions about the effectiveness of pre-trained embeddings (Section 3.1) we find that class embeddings are not meaningful representatives of the dataset classes and that their ability to partition the dataset is limited (5.2). We encourage experimenting with this approach, as it is relatively inexpensive to compute. In addition to using common heuristics with BERT, such as averaging the word embeddings, fine-tuning the sentence-embeddings on the dataset might make a difference and result in higher quality Class Embeddings. Also, testing the approach on different datasets is crucial - in our work, improving upon random selection is difficult even for sophisticated Active Learning strategies. Finally, we would like to motivate more application-oriented research (e.g. Information Retrieval, Semantic Similarity rankings etc...) into the inner workings of pre-trained contextual embeddings in order to improve understanding of the information they encode.

Acknowledgments

This work was funded and supported by IBM.

References

Bang An, Wenjun Wu, and Huimin Han. 2018. Deep active learning for text classification. In *Proceedings of the 2nd International Conference on Vision, Image and Signal Processing*, pages 1–6.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

Kenneth Ward Church. 2017. Word2vec. *Natural Language Engineering*, 23(1):155–162.

Timo I Denk and Christian Reisswig. 2019. Bertgrid: Contextualized embedding for 2d document representation and understanding. *arXiv preprint arXiv:1909.04948*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Daniel Gissin and Shai Shalev-Shwartz. 2019. Discriminative active learning. *arXiv preprint arXiv:1907.06347*.

Mohamed Goudjil, Mouloud Koudil, Mouldi Bedda, and Noureddine Ghoggali. 2018. A novel active learning method using svm for text classification. *International Journal of Automation and Computing*, 15(3):290–298.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Oscar Reyes, Carlos Morell, and Sebastián Ventura. 2018. Effective active learning strategy for multi-label learning. *Neurocomputing*, 273:494–508.

Christopher Schröder and Andreas Niekler. 2020. A survey of active learning for text classification using deep neural networks. *arXiv preprint arXiv:2008.07267*.

Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66.

Lukas Wertz, Katsiaryna Mirylenka, Jonas Kuhn, and Jasmina Bogojeska. 2022. Investigating active learning sampling strategies for extreme multi label text classification. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4597–4605, Marseille, France. European Language Resources Association.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *ArXiv*, abs/1909.10430.

Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. 2009. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 917–926.

Michelle Yuan, Hsuan-Tien Lin, and Jordan L. Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. *CoRR*, abs/2010.09535.

Xueying Zhan, Qingzhong Wang, Kuan-hao Huang, Haoyi Xiong, Dejing Dou, and Antoni B. Chan. 2022. A comparative survey of deep active learning.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610, Online. Association for Computational Linguistics.

Appendix

size of D	class 1	class 2	class 3	class 4	class 5	class 6	class 7	class 8	class 9	class 10
100	1	0	0	0	0	0	7	2	0	0
200	2	0	0	0	0	0	7	2	0	0
500	1	0	0	0	0	0	7	2	0	0
1500	2	0	0	0	0	0	7	1	0	0

Table 3: Number of samples in the unlabeled set U of the **arXiv** dataset with class j found within the closest 100 samples of the centroid of class j using pre-trained Sentence-BERT. We experiment with varying sizes of the labeled set D.

	arXiv	*eurlex*
class 1	High-Energy-Physics	import
class 2	Statistical Mechanics	export refund
class 3	Quantum Physics	Pip Fruit
class 4	Superconductivty	Fruit Vegetable
class 5	Strongly Correlated Electrons	Citrus Fruit
class 6	Atomic and Molecular Clusters	Quantitative Restriction
class 7	Network Architecture	Germany
class 8	Formal Languages	Portugal
class 9	Human Computer Interaction	Ship's Flag
class 10	Other Computer Science	Gaming

Table 4: Descriptions of labels used in both datasets.Frequent labels are above center line, rare labels are below center line.

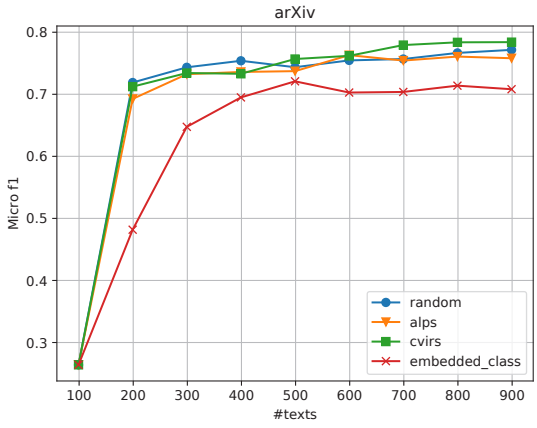


Figure 7: Micro F1 on the arXiv dataset.

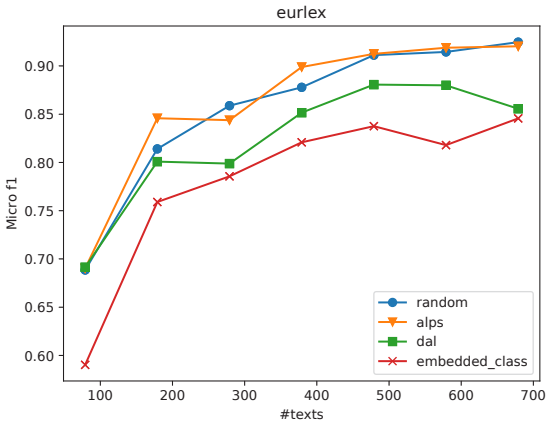


Figure 8: Micro F1 on the arXiv dataset.

5 Research Output of the Intelligent Vision Systems Group

5 Research Output of the Intelligent Vision Systems Group

The IVS group, led by Frank-Peter Schilling, conducts research in the domains of deep-learning based computer vision, Machine Learning Operations (MLOps), as well as in methods to create trustworthy and certifiable AI systems.

We are interested in computer vision using image or video data, for which we develop state of the art deep neural network architectures. We are particularly interested in recent developments including vision transformers, gauge equivariant neural networks and geometric deep learning. Domains of applications include, but are not limited to, industrial quality control, medical imaging and diagnosis (computed tomography), as well as earth (satellites) and sky (radio-astronomy) observation data. Our second main area of interest concerns MLOps, which describes best practices for building complete, production-ready and scalable Machine Learning systems. Finally, we are interested in methods to create safe, trustworthy and certifiable AI systems, which comply with current and future legislation.

The group was newly created in 2022 and currently consists of one senior lecturer (Prof. Dr. Frank-Peter Schilling), one senior scientist (Dr. Philipp Denzel) and one doctoral student (Daniel Barco). It is set to grow further in the next years.

We concluded the successful project *“DIR3CT: Deep Image Reconstruction through X-Ray Projection-based 3D Learning of Computed Tomography Volumes”* which was carried out jointly with a team from the Institute of Applied Mathematics and Physics (IAMP) in collaboration with Varian Medical Systems, the world market leader in clinical radiation therapy. We developed a novel deep-learning based approach to mitigate motion-induced artefacts in 3D cone-beam CT (CBCT) images acquired during patient treatment. The results led to a publication with the journal *Medical Physics* (under review), as well as a poster presented at the annual meeting of the *American Association of Physicists in Medicine AAPM* in Washington DC. The research collaboration with Varian is continued in the recently started follow-up project *“AC3T – AI powered CBCT for improved Combination Cancer Therapy”,* which has the goal to enable a novel, combined, adaptive cancer therapy due to significantly improved 3D and 4D low dose CBCT images based on AI-improved image reconstruction. It also involves two universities and one start-up in South Korea.

Together with CAI associate Dr. Elena Gavagnin (Institute of Business IT), we joined SKACH, the consortium of Swiss universities and research institutions participating in the international “big science” project *SKAO - Square Kilometre Array Observatory*, which will become the world’s largest radio telescope. We apply generative deep-learning models to the pipeline from astrophysical simulations to mock telescope observations. Further, we took a major role, together with the CVPC group, in the newly started project *“certAInty - A Certification Scheme for AI systems”,* and we led a project for Roche Diagnostics, *“OSR4H - Open Set Recognition for Hematology”,* joined by researchers from the Institute of Computational Life Sciences (ICLS). Finally, our summary article about the workshop *“1st International Symposium on the Science of Data Science: ISSDS 2021”,* which we organized in 2021, was published in the *Archives of Data Science.*

We thank our project partners and funding agencies for their support, without which these results would not have been possible!

The IVS 2022 team

Frank-Peter Schilling, Daniel Barco and Philipp Denzel (along with two associated faculty members



Foundations of Data Science: A Comprehensive Overview Formed at the 1st International Symposium on the Science of Data Science

Frank-Peter Schilling, Dandolo Flumini, Rudolf M. Füchslin, Elena Gavagnin, Armando Geller, Silvia Quarteroni and Thilo Stadelmann

Abstract We present a summary of the 1^{st} International Symposium on the Science of Data Science, organized in Summer 2021 as a satellite event of the 8^{th} Swiss Conference on Data Science held in Lucerne, Switzerland. We discuss what establishes the scientific core of the discipline of data science by introducing the corresponding research question, providing a concise overview of

Frank-Peter Schilling, Thilo Stadelmann
ZHAW Centre for Artificial Intelligence and ZHAW Datalab, Winterthur, Switzerland
✉ scik@zhaw.ch, stdm@zhaw.ch

Dandolo Flumini
ZHAW Institute of Applied Mathematics and Physics and ZHAW Datalab, Winterthur, Switzerland
✉ flum@zhaw.ch

Rudolf M. Füchslin
ZHAW Institute of Applied Mathematics and Physics and ZHAW Datalab, Winterthur, Switzerland and European Centre for Living Technology, Venice, Italy
✉ furu@zhaw.ch

Elena Gavagnin
ZHAW Institute of Business Information Technology and ZHAW Datalab, Winterthur, Switzerland
✉ gava@zhaw.ch

Armando Geller
Scensei (Switzerland) GmbH, Zurich, Switzerland
✉ armando@scensei.com

Silvia Quarteroni
Swiss Data Science Center (SDSC) and EPFL, Lausanne and Zurich, Switzerland
✉ silvia.quarteroni@datascience.ch

ARCHIVES OF DATA SCIENCE, SERIES A
(ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING
Vol. -, No. -, -

DOI 10.5445/KSP/XXXXXXXX/XX
ISSN 2363-9881

relevant related prior work, followed by a summary of the individual workshop contributions. Finally, we expand on the common views which were formed during the extensive workshop discussions.

1 Introduction

The discipline of artificial intelligence was coined at the Dartmouth Conference (McCorduck, 1979; Nilsson, 2009); the discipline of Data Science was allegedly coined at LinkedIn and Facebook (Stadelmann et al, 2019b). If truth can be extracted from this abridged statement, it is the fact that data science as today's emerging discipline (Brodie, 2019b) has been largely shaped outside the walls of academia (Stadelmann et al, 2013), i.e., outside a scientific environment, but rather in business-driven settings. The goal of the recent 1st Symposium on the Science of Data Science (Schilling et al, 2021) hence has been to discuss the canon of its underlying principles and techniques (models, methods) that are applicable across different use cases and fields of application, to answer the question what "science" underlies the discipline—if it actually is a discipline.

Put in simpler terms, the symposium revolved around the following hypothetical question: If, 15 years from now, one would compare the contents of the standard textbooks of statistics, computer science, AI and other "source disciplines" of data science on the one hand, with the contents of the then classic text book of data science (still to be written) on the other hand—what would be part of the data science textbook? What establishes the scientific core of data science that is not covered somewhere else? The symposium's goal thus was to launch an activity towards establishing a reference framework for data science.

The importance of this activity transcends common academic drives for order, rigour and scrutiny. First, by starting research labs, degree programs and whole departments, academia creates structures and molds careers that will stay for a long time. It is important that these developments are well-founded, non-redundant and long-lasting, and not just tailored to a sudden demand. Second, a lesson can be learned from projected similarities between data science's development on the one hand, and how computer science on the other hand emerged out of the fields of mathematics and electrical engineering in the 1950s in Germany (Gunzenhäuser, 1988). At first being little more than the application of principles of these two source disciplines, computer science used the space it was granted as a new discipline to grow into completely new areas that might

arguably not have been developed otherwise (see also the *Annals of the History of Computing*). Today, little of computer science's curricula overlap with mathematics or electrical engineering as a result of this emancipation that paved the way for much of what propelled (scientific and economic) progress in the last decades.

Similar to the way Denning (2005) argued for computer science, we thus think that data science has the potential to "[meet] every criterion for being a science, but it has a self-inflicted credibility problem"—the mainstream and media hype around it. In the remainder of this paper, we survey related work on the foundations of data science in Section 2; We then summarize the main contributions from the ISSDS'21 symposium in a synthesis-forming way in Section 3, pointing to a solution to the credibility problems. Last, but not least, we discuss the ensuing implications in Section 4, before the concluding remarks. This paper thus serves as a key and introduction to the individual contributions from the ISSDS'21 participants.

2 Related work

Ever since the term "data science" came into existence around 60 years ago, there has been a debate on what exactly constitutes data science, how it differentiates itself from statistics and computer science, and whether it deserves the word science in its name. Can it be viewed as an academic discipline on its own that represents more than the sum of its constituent disciplines? In the following, we address these questions by giving a brief historic account as well as a, necessarily incomplete, summary of the current debate.

2.1 Historic roots

The first use of the term "data science" as a new scientific field goes back to the early 1960s, when Peter Naur introduced the term (interchangeably with "datalogy") (Sveinsdottir and Frøkjær, 1988), while John Tukey (1962) described a new scientific field he called "data analysis". In 1974, the term "data science" appeared in Naur's book "Concise Survey of Computer Methods" (Naur (1974), p. 30):

> Data science is the science of dealing with data, once they have been established, while the relation of data to what they represent is delegated to other fields and sciences.

It builds upon the IFIP[1] definition of data as "a representation of facts or ideas in a formalised manner capable of being communicated or manipulated by some process" (Gould, I.H. (ed.), 1971). Naur had a conception of data science rooted in computer science, while Tukey used the term in reference to statistics, two perspectives which are also alluded to in David Hand's two kinds of *big data exercise* (Hand, 2016).

The discussions in the scientific community then continued through the 80s and 90s. In 1985, C.F. Jeff Wu (1986) used the term "data science" as an alternative name for statistics. Later, in his inaugural lecture "Statistics=Data Science?" at the University of Michigan (Wu, 1997), he summarized statistics as a trilogy of data collection, data modelling and analysis, together with problem solving and decision making. He highlighted the most relevant future directions as dealing with large and complex data (data mining), employing a data-driven, empirical approach, as well as the representation of knowledge, and finally suggests that it is time for statistics to make a bold move, namely to rename itself to data science.

Already in 1992, at a statistics symposium in Montpellier, France, the emergence of a data science as a new discipline was acknowledged (Escoufier et al, 1995):

> The authors propose ways to formalize *data analysis*. ... Such an approach gives birth to a new science with data at its core. Its nature, numerical, qualitative or symbolic, determines the type of operations possible with them. Their origin, whether exhaustive collection or sample, conditions the objective expected in their analysis. It seems justified to coin the term *data science* for this particular activity.

The first international conference which had the term "data science" in its name took place in 1996 in Kobe, Japan, where Chikio Hayashi (1998) argued for data science as a new, interdisciplinary concept with three phases: data design, collection, and analysis.

In his paper "Statistical Modeling: The Two Cultures", Leo Breiman (2001) discussed two approaches to extract value from data: (i) Predictive modeling, i.e. the ability to predict outcomes to future input data to a model, and (ii) inference, i.e. to extract some information about the underlying model which generates the data. Breiman argued that statistics as a discipline so far was almost exclusively focused on inference, and highlighted the importance of

[1] International Federation for Information Processing

predictive modeling (the prime example being machine learning) when using data to solve problems.

In 2001, William S. Cleveland (2001) introduced data science as an independent scientific discipline based on his proposal of:

> ... a plan to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called "data science".

His "action plan" discussed six technical areas by which to extend statistics, namely (i) multidisciplinary investigations, (ii) models and methods for data, (iii) computing with data, (iv) pedagogy, (v) tool evaluation and (vi) theory.

These six technical areas of data science introduced by Cleveland were updated and generalized by David Donoho (2017) into the "6 divisions of Greater Data Science" as follows:

(i) Data exploration and preparation (exploratory data analysis, data cleaning);
(ii) Data representation and transformation (databases, feature extraction);
(iii) Computing with data (computer languages like R and Python, cluster and cloud computing, workflows and packages);
(iv) Data modeling (both generative and predictive cultures, c.f. Breiman (2001));
(v) Data visualization and presentation (plotting tools, dashboards);
(vi) Science about data science.

About the last division, the science about data science, Donoho writes (Donoho (2017), p. 756):

> Data scientists are doing science about data science when they identify commonly-occurring analysis/processing workflows, for example using data about their frequency of occurrence in some scholarly or business domain; when they measure the effectiveness of standard workflows in terms of the human time, the computing resource, the analysis validity, or other performance metric, and when they uncover emergent phenomena in data analysis, for example new patterns arising in data analysis workflows, or disturbing artifacts in published analysis results. The scope here also includes foundational work to make future such science possible – such as encoding documentation of individual analyses and conclusions in a standard digital format for future harvesting and meta analysis.

Donoho gives meta- and cross-study analyses as examples for this "science about data science", and concludes that it will grow dramatically in significance in the future, in particular because of the paradigm of reproducible, open science.

2.2 Contemporary definitions

In his comprehensive overview, Longbing Cao (2017) presents two definitions of data science as well as of data products (p. 43:8):

> Definition 2.1 (Data Science): A high-level statement is: "data science is the science of data" or "data science is the study of data."
> Definition 2.2 (Data Science): From the disciplinary perspective, data science is a new interdisciplinary field that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments (including domains and other contextual aspects, such as organizational and social aspects) in order to transform data to insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology.
> Definition 2.3 (Data Products): A data product is a deliverable from data, or is enabled or driven by data, and can be a discovery, prediction, service, recommendation, decision-making insight, thinking, model, mode, paradigm, tool, or system. The ultimate data products of value are knowledge, intelligence, wisdom, and decision.

In (Stadelmann et al, 2019a), it is argued that when defining data science, either a top-down or a bottom-up approach can be followed: The top-down view understands the field as the study of approaches to generate value from data and building data products, while in the bottom-up view, data science is an interdisciplinary research field with a new, holistic way to deal with data, integrating competencies from computer science, statistics, AI, data mining, but also entrepreneurship. Their definition of data science reads (p. 18):

> Data science refers to a unique blend of principles and methods from analytics, engineering, entrepreneurship and communication that aims at generating value from the data itself.

In contrast, Ley and Bordas (2018) first coin data science as "statistics 2.0", a rebirth of statistics in the big data era which has imposed new challenges and opened new research directions. They continue along similar lines as above by stating that being interdisciplinary by nature (statistics, computer and computational sciences, mathematics), it usually combines top-down (model-driven) and bottom-up (data-driven) approaches.

However, Weihs and Ickstadt (2018) point out that the role of statistics and statistical modeling of observational data in data science is often underestimated compared with e.g. computer science.

Michael L. Brodie (2019b) gives the following definition of data science (p. 104):

> Data Science is a body of principles and techniques for applying data analytic methods to data at scale, including volume, velocity, and variety, to accelerate the investigation of phenomena represented by the data, by acquiring data, preparing and integrating it, possibly integrated with existing data, to discover correlations in the data, with measures of likelihood and within

> error bounds. Results are interpreted with respect to some predefined (theoretical, deductive, top-down) or emergent (fact-based, inductive, bottom-up) specification of the properties of the phenomena being investigated.

He considers data science a new paradigm, different from the scientific method [2] in terms of (i) data, (ii) methods, (iii) models and (iv) outcomes:

(i) The data are often of observational nature, rather than being collected as in a controlled experiment as in the natural sciences.
(ii) Methods are typically domain- and data-specific, even though based on general (e.g. machine learning) approaches.
(iii) Models are created on demand and ad-hoc, and changed or updated frequently, rather than being proposed and refined over many years.
(iv) Regarding outcomes, "the scientific method is used to discover causal relationships between a small number of variables that represent the essential characteristics of the natural phenomena being analyzed" while data science is mainly used to discover correlations.

2.3 Current debates

In Brodie's view, empirical science and data science have another fundamental difference: the scientific method uses deductive reasoning, is hypothesis- or theory-driven, and works top-down, while data science is typically data-driven, uses inductive reasoning and works bottom-up (in contrast to the top-down views of data science highlighted by Ley and Bordas (2018) and Stadelmann et al (2019a), see section 2.2). Since a scientific discipline would require "fundamental principles and techniques applicable to all relevant domains", rather than being domain-, model- and method specific, Brodie does not consider data science a science yet, but rather "an immature, emerging domain that will take a decade to mature".

Regarding the development of data science as a discipline, Brodie (2019a) suggests that this process will be driven by the virtuous cycle of research, development and delivery (RD&D) underlying applied science, as will be the development of data science applications and education.

[2] In the context of this paper, with the term "scientific method" we refer to the common underlying iterative process across natural sciences, which employs empirical methods as fundamental mean to validate predictions, derived from newly-formulated hypotheses about a specific research question (Galilei, 1638; Newton, 1687; Popper, 1959).

F. Jack Smith (2006) compares computer science and data science with respect to their recognition as an academic discipline, stating that for both, they are often perceived as being merely within the realm of tools used by technicians. Smith remarks that an important indicator for the establishment of an academic discipline is the dissemination of scientific articles through peer-reviewed journals, which so far had been lacking in the area of data science, but which has started to change in the 2000s, as it did for computer science already from the 1950s.

Making the connection to the field of Artificial Intelligence (AI) and machine learning which is at the core of many data science problems, Michael I. Jordan notes that we are witnessing the creation of a new branch of engineering which should be developed in a human-centric way (Jordan, 2019a,b).

A similar point is made by Blei and Smyth (2017) who present a holistic view of data science. It includes not only the statistical and computational perspectives, but also a human perspective, where the latter involves domain knowledge and data understanding, the ability to fuse methods from both the statistical and computational domains, as well as the task to interpret and visualize the results in their context.

Taking again a more sceptical view, Irizarry (2020) claims that the definitions of data science as given above generally lack consensus on the fundamental principles and the author proposes, in agreement with Jeannette Wing (2019), that "data science is an umbrella term to describe the entire complex and multi-step processes used to extract value from data".

Provost and Fawcett (2013) discuss data science from the perspective of its application to the business world, and state that the "the ultimate goal of data science is improving decision making", supported by data engineering and processing including big data technologies which they however do not consider to be part of data science. Provost and Fawcett advise not to confuse the description of the day-to-day tasks of a data scientist (at the technical level, which often involves a large amount of data processing) with a formal definition of data science as an academic discipline.

2.4 State of the art

Looking forward, Jeannette Wing (2020) formulates three meta-questions about data science as a discipline:

(i) What is/are the driving deep question(s) of data science, similar to the questions about the origin of life in biology or the origin of the universe in astrophysics?[3]
(ii) What is the role of the domain in the field of data science, i.e. is the inclusion of the domain specific to data science?
(iii) What makes data science a science, i.e. what makes it more than the sum of its constituent disciplines computer science and statistics?

She then discusses 10 research challenge areas in data science, among others scientific understanding of learning algorithms, causal reasoning, trustworthy AI, privacy and ethics.

In summary, three diverging main themes emerge from the historic and current discussion regarding data science as a scientific discipline:

(i) Data science is often considered an extension/update of statistics ("statistics 2.0"), which is upscaled to meet the new challenges of the big data era, and it is shifting its focus from inference to prediction.
(ii) Data science is an interdisciplinary field, built upon varying selections of fields but mostly upon statistics and computer science, while adding data understanding and domain knowledge as a new perspective.
(iii) Data science can be approached from both top-down (model-driven, generating value, building data products) and bottom-up (data-driven) perspectives. It may be viewed as a new paradigm, which is different from the traditional scientific method which employs controlled experiments.

Thus, at the time of writing, no consensus seems to have formed yet on the question whether data science can be considered an independent academic discipline.

3 Aspects of data science

In the following, we present a summary of the main ideas presented in the individual workshop contributions which were received. More details can be found in the individual articles contained in the same volume of this journal.

[3] The authors would add the unsolved P versus NP problem in theoretical computer science here.

3.1 A new scientific paradigm?

Four workshop contributions (Doemer and Kempf, 2022; Heitz and Schumann, 2022; Ott et al, 2022; Stadelmann et al, 2022) tried to answer the question about the scientific nature of data science as a discipline.

Doemer and Kempf (2022) argue that Data Science can be viewed as a new paradigm in scientific practice, in addition to experimental, theoretical and computational science. As discussed in Section 2, data science can be viewed as either a data-driven (inductive), or hypothesis-driven (deductive) approach. However, the paradigm-shifting nature of data science comes with a problem still awaiting a practical solution: The increasingly complex setups producing huge amounts of data and information at various levels (e.g., meta data in addition to observational data). These are typically of merely observational nature for the data scientist, in contrast with those classically obtained through experiments (in line with the scientific method's principles), i.e. generated under controlled conditions and setup. As a potential solution to this problem, which would provide a basis that allows data science to be consistent with the requirements of transparency, traceability and reproducibility demanded by the scientific method, the authors suggest the adoption of tools, frameworks and platforms provided by *"XOps"* ("X for IT Operations") approaches, where X can be e.g. ML (Machine Learning), Data or AI. For example, MLOps is a set of best practices that aims to deploy and maintain ML models in production reliably and efficiently.

Heitz and Schumann (2022) state that data science consists of *two elements, one based on engineering and the other based on science*. The science element is concerned with creating insights based on phenomena/data measured in the real world, while the engineering element is concerned with creating value ("data products") by making use of derived insights. On the scientific side, it is argued that the way in which the insights are derived from the data must follow scientific principles such as empirical evidence, validity (e.g. in terms of statistical significance)[4]. and reproducibility. Such a scientific process is then able to make predictions that however need not necessarily be accompanied by a causal model, in contrast to (Brodie, 2019b). On the other hand, the engineering element involves anything that changes the course of the world, as opposed to plain knowledge creation. It includes not only the data product itself, but

[4] Regarding aspects of model validation besides accuracy and significance, see also the discussion in section 2.2.3 of Oberkampf and Roy (2010).

also decision making, which often depends on external factors not related to data analytics as well. Finally, the engineering side should also include ethical considerations, such as algorithmic fairness.

Ott et al (2022) make the case for a *systemic view* of the data science workflow, which extends the "classical" workflow (comprising data collection, cleaning, visualization, model building, evaluation and impact/value creation) with various stakeholders such as data scientists, business owners, domain experts and users giving feedback, as well as with societal influences and impacts, that all influence the outcome of a data science project. Four hypotheses towards this perspective are developed:

(i) There is a need for more abstraction and automation in the data workflow and pipeline engineering process.
(ii) Humans play an active role in the data science workflow (e.g. in active learning, or in identifying bias).
(iii) Data science will diversify at the intersection of domains (e.g. life science, health, economics and business etc.).
(iv) With increasing complexity, data science workflows evolve into complex networks, which can be studied and organized with the help of complex systems science.

Stadelmann et al (2022) propose their answer to the question of the scientific core of data science, which distinguishes it from its contributing disciplines and is not already part of one of them. For the authors, this overarching, unique principle is *data centrism*, i.e. putting data at the center and subject of study, something which is not the case for the contributing disciplines, neither for statistics and machine learning, nor for computer science or service engineering. The unique principle in data science is to create value out of *actual* data (but not ignoring tools and methods to improve data acquisition), and it is argued that recent trends such as explainability (e.g., explainable AI or XAI, see also (Melchior, 2022)) and trustworthiness, but also learning from less supervision, are grounded in the data centrism of data science. Finally, the authors argue that, besides its core of data centrism, data science includes several new areas of research which are not dealt with in the contributing disciplines per se, such as MLOps (see also (Doemer and Kempf, 2022)), or whose current surge can be attributed to a mindset shift originating in the use cases and culture shaped by data science, like applied semi- and weakly-supervised learning (Simmler et al, 2021), or explainable AI. One example of the latter are explanations of deep neural networks whose necessity arises out of data-driven applications in

safety-critical sectors like healthcare (Jin et al, 2022), while other aspects of explainability have been dealt with since longer (Keil and Wilson, 2000).

3.2 Explainability, rationality and trust

Furthermore, Melchior (2022) and Füchslin and Flumini (2022) discuss special topics in data science, namely the important issues of transparency and explainability as well as various ways of defining and automating the decision making process.

Melchior (2022) focuses on the notions of transparency, explainability and interpretability in data science in the context of machine learning models, which is an issue in particular for deep learning architectures where hand-crafted features are replaced with many layers of deep neural networks. Recently, explainable AI (XAI) has become a subject of research, in particular for applications with strong safety or ethical requirements, but also in the case of fundamental/natural sciences, where ML is used for knowledge discovery. It is argued that in order to achieve explainability and interpretability, domain experts and data scientists have to work together. Several concrete technical examples are given for the inclusion of domain knowledge in a deep learning model in order to facilitate learning of interpretable features, such as autoencoders or generative models, invertible flow networks or graph neural networks. The latter are particular promising in view of unifying symbolic and connectionist AI approaches.

Füchslin and Flumini (2022) give a definition of rational decision making (structured, inductive, verifiable, grounded) and it is claimed that the former is more appropriately complemented by arationality rather than irrationality. In summary, rational decision making exhibits two main features. Firstly, it is based on some sort of generally accepted scheme of reasoning (in mathematics expressed in an axiomatic manner) and some data/variables. Secondly, the process of reasoning can be expressed in a language that enables to make the reasoning transparent and comprehensible to a (sufficiently well-educated) other individual and uses terms/variables with a meaning that relates them to the objects one reasons about. Whereas irrationality lacks both of these features, arationality captures the concept of decision making that leads to sensible results but uses processes that one may or may not be able to describe mechanistically, but without the possibility to attribute a meaning to the data

representing the process steps in between input and output. Arational decision making includes for instance the notion of intuition. Discovering new proofs of mathematical theorems and generally generating insights and mathematical theories is typically based on a conjecture-proof workflow that includes arationality. According to the authors, an artificial mathematician must therefore also include arational decision making, which is provided for instance by deep neural networks, providing an implementation of AI, as the authors state in their title, as *Arational Intelligence*.

3.3 Education vs technical skill

Finally, the contribution of Helmer (2022) is concerned with teaching data science and the corresponding curriculum. As particular challenges, the author mentions the very diverse background of students with different levels of technical skills, the difficulty in providing a suitable computing environment for labs and exercises (local vs cloud based) given the short life cycles of relevant tools and frameworks, and the selection of appropriate use cases and datasets. It is argued that the curriculum should be structured according to the elements and layers of a typical data lifecycle model, in order to provide a structure and frame for the theoretical foundations. Regarding practical approaches, the author suggests, largely in agreement with Irizarry (2020), to structure the curriculum into backend (data engineering) and frontend (data analysis, machine learning) parts, to build the knowledge for developing and maintaining data processing pipelines, to teach data science at the graduate (rather than undergraduate) level, and to consider theoretical foundations at least as important as practical examples. Focusing too much on ever-changing tools and frameworks would shift the curriculum too much towards training, as opposed to education. Finally, it is argued that care should be taken not to standardize this still very young and fluid field too quickly.

In summary, the individual workshop contributions highlight different aspects of data science and address the research question of its scientific core from various complementary angles. They formed the basis for the common discussion, which is summarized in the following section.

4 Discussion

One of the potential controversial aspects when reasoning about the science in data science is the fundamental difference between experimental data, observational (field) data and citizen-based data, i.e. data collected in the context of citizen science projects. Unlike classic quantitative science, data science relies strongly on the latter two categories and not only on controlled experimental data. Observational and citizen-based data are both affected by the big problem of being potentially biased by humans in their selection or generation process. For example, people take most pictures with daylight.

Therefore, one of the goals of data science that distinguishes it from traditional science is to provide a rigorous methodology to handle data from the *real world* by accounting for the inevitable complexity (e.g., bias) or by modeling concepts which can not be directly observed and, therefore, need experiments or simulations (e.g., risk assessment). One proposal formed at the symposium then is the idea that an exemplary common trait in data science is the way insights are derived from (not controlled) data sources. While the insights derived belong primarily to the respective scientific domains, *how* these were derived pertains to data science and ultimately this constitutes one scientific aspect of the discipline.

Another common difficulty when arguing about the scientificity (i.e., referring to systematicity, logicality, certainty, and precision of knowledge (Xu, 2005)) of data science is solving the issue of explainability and trust. Basic founding principles of science are the quest for explainable models and reproducible experiments: Both of these elements are the basis to trust the ensuing results.

In data science, however, it is not always straightforward to rely on fully explainable models and, as already argued before, on controlled experiments, with the obvious result of doubts being cast on the amount of science present in this discipline. An interesting point of discussion in this context is how the concept of *trust* is associated to the explainability of the model, therefore often to its simplicity, rather than to its correctness. However, a simple model delivering wrong results can not be trustworthy, hence this association is not always reasonable. This reveals the need for clarity around the concept of trust within and towards data science. Specifically, alternative ways for building trust—in the outcomes of data science, and by extension into the discipline

itself—need to be found, which do not necessary rely on experiments or full explainability of models.

The following observation might serve as a starter in this direction: while science is intuition-inspired-then-fact-driven, data science is fully fact-driven-then-intuition-enabled. This stems from the observation that in science, theories are sparked by creativity (the intuition of an apple falling from a tree) and later confirmed by a fully rational (i.e., systematic and logical) process. In data science, theories are derived from data by rational models (e.g., number-crunching neural networks) but encoded in such incomprehensible ways (the network's weight matrix) that methods need to be built to bring human intuition back ex post to leverage the findings (XAI).

Whatever viewpoint individual participants took in the discussion, a small set of key words emerged as central elements to their statements on data science: *data*; *"the wild"* (i.e., real-world applications and use-cases); *pipeline*; and *data products*. While debaters couldn't unify behind a coherent picture of how these central issues are related, there was a consensus that

(i) Data is central to data science (data evokes theories and not just confirms them).
(ii) Data science is about the real world, specifically its *messiness* (for which it provides methods and tools to deal with).
(iii) As data products are the natural results of data science (its "claims"), the *process* of creating them (the pipeline) plays an important part in constituting the field.

5 Conclusions

Picking up from where Wing (2020) asked her three meta-questions, considering and eventually deliberating more profoundly on what "a" data science actually is through the lens of philosophy of science (Boyd et al, 1999; Losee, 2001) should be a fruitful endeavor, further shaping the ongoing debate. More specifically, reflecting on and addressing some of the following questions should improve clarity for practitioners and philosophers alike: What is the *purpose* of data science? For example, is it foremost about producing predictions, as many of today's real world applications suggest? Or is it also about creating explanations, too?

If it is also about creating explanations, then what is the *explanatory power* of data science? For example, is it really a black box, as some applications of neural networks would suggest? Is it simply about correlations, as some applications of statistical learning would suggest? Or can we actually learn something from data science about data generation and social mechanisms (Hedstrom, 2005), mid-range theories (Merton, 1949), causality (Pearl and Mackenzie, 2018) and so forth? If so, then this would imply that, at least theoretically, there is *"truth"* through data science. This again would imply that there is a role for rationality, intelligence and intuition in data science and underlying models. And if truth, then also a "measure" of what good science is (Moss and Edmonds, 2005).

And if there is truth, then what is the merit of data science? For example, does it help us to solve practical challenges pertaining to daily decision support tasks better, because it creates more precise *and* accurate predictions? Does data science contribute to conducting science better because it makes better use of an ever expanding repertoire of computational techniques and data repositories (e.g., data lakes)? Does it improve trust in science, because it increases explainability grounded in data?

Similar to the situation in Goethe's "sourcerer's apprentice", it may be that the spirits we summoned, we now cannot rid ourselves of again. No harm done. But at least we should know why we summoned a new scientific discipline. The questions raised above should help creating some clarity. Some attempts to answer them are found in the remaining contributions to this special issue on the 1^{st} Symposium on the Science of Data Science. Others are left for future work.

Acknowledgements The authors are grateful for the support of the ZHAW Datalab and the data innovation alliance in organizing ISSDS'21 as a satellite event of the 8^{th} Swiss Conference on Data Science, and for the participants of the symposium to share and co-shape each other's thoughts. Michael L. Brodie's input to an earlier draft of the symposium concept is very much appreciated.

References

Blei DM, Smyth P (2017) Science and data science. Proceedings of the National Academy of Sciences 114(33):8689–8692, DOI 10.1073/pnas.1702076114, URL https://www.pnas.org/content/114/33/8689

Boyd R, Gasper P, Trout JD (eds) (1999) The Philosophy of Science. MIT Press, Cambridge

Breiman L (2001) Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statistical Science 16(3):199 – 231, DOI 10.1214/ss/1009213726

Brodie ML (2019a) On developing data science. In: Applied Data Science, Springer, pp 131–160, DOI 10.1007/978-3-030-11821-1

Brodie ML (2019b) What is data science? In: Applied Data Science, Springer, pp 101–130, DOI 10.1007/978-3-030-11821-1

Cao L (2017) Data science: A comprehensive overview. ACM Comput Surv 50(3), DOI 10.1145/3076253

Cleveland WS (2001) Data science: an action plan for expanding the technical areas of the field of statistics. International Statistical Review 69(1):21–26, DOI https://doi.org/10.1111/j.1751-5823.2001.tb00477.x

Denning PJ (2005) Is computer science science? Communications of the ACM 48(4):27–31

Doemer M, Kempf D (2022) Is it Ops that make Data Science Scientific? Archives of Data Science, Series A 8(1):XX–YY

Donoho D (2017) 50 years of data science. Journal of Computational and Graphical Statistics 26(4):745–766, DOI 10.1080/10618600.2017.1384734

Escoufier Y, et al (1995) Preface. In: Escoufier Y, et al (eds) Data Science and its Applications, Academic Press, Tokyo, pp (vii)–(viii)

Füchslin RM, Flumini D (2022) AI as Arational Intelligence? Archives of Data Science, Series A 8(1):XX–YY

Galilei G (1638) Discorsi e dimostrazioni matematiche intorno a due nuove scienze attinenti la meccanica e i movimenti locali. Elzeviri, Leiden (NL)

Gould, IH (ed) (1971) IFIP Guide to Concepts and Terms in Data Processing. North-Holland Publishing Company

Gunzenhäuser R (1988) Entwicklung und Bedeutung der Informatik in den Hochschulen der Bundesrepublik Deutschland. In: Baur F (ed) Nutzungsbilanz moderner Informations-und Kommunikationssysteme aus Anwendersicht/User Experience in the Application of Modern Information and Communication Systems, Springer, pp 25–36, DOI 10.1007/978-3-642-83515-5

Hand DJ (2016) Editorial. 'Big data' and data sharing. Journal of the Royal Statistical Society: Series A (Statistics in Society) 179(3):629 – 631, DOI 10.1111/rssa.12185

Hayashi C (1998) What is Data Science? Fundamental Concepts and a Heuristic Example. In: Hayashi C, Yajima K, Bock HH, Ohsumi N, Tanaka Y, Baba

Y (eds) Data Science, Classification, and Related Methods, Springer Japan, Tokyo, pp 40–51
Hedstrom P (2005) Dissecting the Social. On the Principles of Analytical Sociology. Cambridge University Press, DOI 10.1017/CBO9780511488801
Heitz C, Schumann R (2022) Data Science: What is science, and what is engineering? Archives of Data Science, Series A 8(1):XX–YY
Helmer S (2022) Teaching Data Science: Constructing Pillars in a Fluid Field. Archives of Data Science, Series A 8(1):XX–YY
Irizarry RA (2020) The role of academia in data science education. Harvard Data Science Review 2(1), DOI 10.1162/99608f92.dd363929, URL https://hdsr.mitpress.mit.edu/pub/gg6swfqh
Jin D, Sergeeva E, Weng WH, Chauhan G, Szolovits P (2022) Explainable deep learning in healthcare: A methodological survey from an attribution view. WIREs Mechanisms of Disease n/a(n/a):e1548, DOI 10.1002/wsbm.1548
Jordan MI (2019a) Artificial intelligence—the revolution hasn't happened yet. Harvard Data Science Review 1(1), DOI 10.1162/99608f92.f06c6e61, URL https://hdsr.mitpress.mit.edu/pub/wot7mkc1
Jordan MI (2019b) Dr. AI or: How I Learned to Stop Worrying and Love Economics. Harvard Data Science Review 1(1), DOI 10.1162/99608f92.b9006d09, URL https://hdsr.mitpress.mit.edu/pub/2imtstfu
Keil FC, Wilson RA (eds) (2000) Explanation and Cognition. The MIT Press, Cambridge
Ley C, Bordas SPA (2018) What makes Data Science different? A discussion involving Statistics2.0 and Computational Sciences. International Journal of Data Science and Analytics 6(3):167–175, DOI 10.1007/s41060-017-0090-x
Losee J (2001) A Historical Introduction to the Philosophy of Science, 4th edn. Oxford University Press, Oxford
McCorduck P (1979) Machines who think: A personal inquiry into the history and prospects of artificial intelligence. CRC Press
Melchior M (2022) Incorporating Domain Knowledge for Learning Interpretable Features. Archives of Data Science, Series A 8(1):XX–YY
Merton RK (1949) On sociological theories of the middle range. In: Social Theory and Social Structure, Simon & Schuster, The Free Press, New York, pp 39–53
Moss S, Edmonds B (2005) Towards good social science. Journal of Artificial Societies and Social Simulation 8(4):13

Naur P (1974) Concise Survey of Computer Methods. Studentlitteratur, Lund, Sweden

Newton I (1687) Philosophiae naturalis principia mathematica. J. Societatis Regiae ac Typis J. Streater

Nilsson NJ (2009) The quest for artificial intelligence. Cambridge University Press, DOI 10.1017/CBO9780511819346

Oberkampf WL, Roy CJ (2010) Verification and Validation in Scientific Computing, 1st edn. Cambridge University Press

Ott T, Horn C, Garcia V (2022) The Sciences of Data – Moving Towards a Comprehensive Systems Perspective. Archives of Data Science, Series A 8(2):XX–YY

Pearl J, Mackenzie D (2018) The Book of Why. Basic Books, DOI 10.5555/3238230

Popper KR (1959) The logic of scientific discovery. The logic of scientific discovery., Basic Books, Oxford, England, pages: 480

Provost F, Fawcett T (2013) Data science and its relationship to big data and data-driven decision making. Big Data 1:51–59, DOI 10.1089/big.2013.1508

Schilling FP, Flumini D, Füchslin RM, Stadelmann T (2021) ISSDS 2021: 1st Intl. Symposium on the Science of Data Science. URL https://sds.data-innovation.org/sds2021-1st-international-symposium-on-the-science-of-data accessed: 2022-07-02

Simmler N, Sager P, Andermatt P, Chavarriaga R, Schilling FP, Rosenthal M, Stadelmann T (2021) A Survey of Un-, Weakly-, and Semi-Supervised Learning Methods for Noisy, Missing and Partial Labels in Industrial Vision Applications. In: 8th Swiss Conference on Data Science, IEEE, DOI 10.1109/SDS51136.2021.00012

Smith FJ (2006) Data science as an academic discipline. Data Science Journal 5:163–164, DOI 10.2481/dsj.5.163, URL https://datascience.codata.org/articles/abstract/10.2481/dsj.5.163/

Stadelmann T, Stockinger K, Braschler M, Cieliebak M, Baudinot G, Dürr O, Ruckstuhl A (2013) Applied data science in Europe: Challenges for academia in keeping up with a highly demanded topic. In: 9th European Computer Science Summit, Amsterdam, Niederlande, 8-9 October 2013

Stadelmann T, Braschler M, Stockinger K (2019a) Data science. In: Applied Data Science, Springer, pp 17–29, DOI 10.1007/978-3-030-11821-1

Stadelmann T, Braschler M, Stockinger K (2019b) Introduction to applied data science. In: Applied Data Science, Springer, pp 3–16, DOI 10.1007/978-3-030-11821-1

Stadelmann T, Klamt T, Merkt PH (2022) Data Centrism and the Core of Data Science as a Scientific Discipline. Archives of Data Science, Series A 8(1):1–16, DOI 10.5445/IR/1000143637

Sveinsdottir E, Frøkjær E (1988) Datalogy — The Copenhagen tradition of computer science. BIT Numerical Mathematics 28(3):450–472

Tukey JW (1962) The Future of Data Analysis. The Annals of Mathematical Statistics 33(1):1 – 67, DOI 10.1214/aoms/1177704711

Weihs C, Ickstadt K (2018) Data science: the impact of statistics. Int J Data Sci Anal 6(3):189–194, DOI 10.1007/s41060-018-0102-5

Wing JM (2019) The data life cycle. Harvard Data Science Review 1(1), DOI 10.1162/99608f92.e26845b4, URL https://hdsr.mitpress.mit.edu/pub/577rq08d

Wing JM (2020) Ten research challenge areas in data science. CoRR abs/2002.05658, URL https://arxiv.org/abs/2002.05658

Wu CFJ (1986) Future directions of statistical research in China: A historical perspective. Application of Statistics and Management 1:1 – 7

Wu CFJ (1997) Statistics = Data Science? URL http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf

Xu ZL (2005) Science and scientificity. Genomics Proteomics Bioinformatics 3(4):197 – 200, DOI 10.1016/s1672-0229(05)03026-3

AAPM 2022
JULY 10–14 | WASHINGTON, DC
64TH ANNUAL MEETING & EXHIBITION

Deep Learning-Based Simultaneous Multi-Phase Deformable Image Registration of Sparse 4D-CBCT

I. HERZIG[1], P. PAYSAN[2], S. SCHEIB[2], A. ZUEST[1], F-P. SCHILLING[3], J. MONTOYA[3], M. AMIRIAN[3], T. STADELMANN[3], P. EGGENBERGER[1], R. FÜCHSLIN[1], L. LICHTENSTEIGER[1]
1 Zurich University of Applied Sciences ZHAW, Institute for Applied Mathematics and Physics IAMP, Switzerland
2 Varian Medical Systems Imaging Laboratory, Daettwil AG, CH
3 Zurich University of Applied Sciences ZHAW, Centre for Artificial Intelligence CAI, Switzerland

INTRODUCTION

Respiratory gated 4D-CBCT suffers from sparseness artefacts caused by the limited number of projections available for each respiratory phase. These artefacts severely impact traditional deformable image registration methods used to extract motion information. We use a supervised deep learning method that is able to predict displacement vector-fields (DVF) from sparse 4D-CBCT despite the presence of artefacts.

MATERIALS & METHODS

Motion Simulation for Data Generation

Training data is generated by a motion simulation framework based on Ref. [1] but extended to use multiple phases. The simulation uses respiratory phase gated 4D-CT scans and a collection of recorded breathing curves.

Preprocessing 4D-CT: Deformable image registration (Deeds [3]) between the 10 consecutive breathing phases creates a set of 10 patient-specific cyclic DVFs.

Simulated 4D-CBCT Scan:

1. Select a combination of breathing curve and 4D-CT scan as inputs to the simulation. This enables data-augmentation by producing multiple simulated scans based on the same 4D-CT using different breathing curves.
2. For each x-ray acquisition, deform the max-exhale 4D-CT volume according to the curves' amplitude at this point using the previously generated DVFs.
3. Forward project the deformed volumes to simulate the x-ray acquisition.

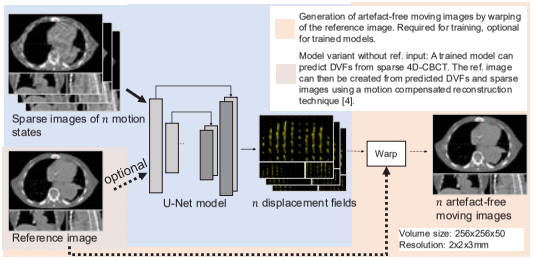
This yields a full set of 4D-CBCT projections where each projection corresponds to a different respiratory state and a set of ground truth volumes for all phases. An iterative reconstruction algorithm is then used to create the sparse (phase gated) training volumes from these .

Datasets

The training dataset consists of 560 simulated 4D-CBCT of 56 different patients, each scan using a different breathing curve; the generated data include fully sampled ground-truth images that are used to train the network. For validation and calculation of the reported metrics we use a new set of 120 scans of 12 unseen patients.

Models & Implementation

We trained U-Net-type convolutional neural network models to predict multiple (10) DVFs in a single forward pass given multiple sparse, gated CBCT and an optional artefact-free reference image as inputs. The predicted DVFs are used to warp the reference image to the different motion states, resulting in an artefact-free image for each state. The models are trained in a supervised way by comparing the resulting images to the ground truth. The trained models are four layers deep, the image size is halved at each downward step and nearest-neighbour upsampling is used in the upward branch. All activations are leaky ReLu (slope=0.2).



CONCLUSIONS

To the best of our knowledge, this is the first time CNNs are used to predict multi-phase DVFs in a single forward pass. This enables novel applications such as 4D-auto-segmentation, motion compensated image reconstruction, motion analyses, and patient motion modeling.

RESULTS

Our method clearly outperforms pairwise registration using the Deeds algorithm alone. PSNR improved from 28.6 to 38.6. In addition, the runtime of our learning-based method is orders of magnitude shorter (2 seconds instead of over 10 minutes). Our results also indicate slightly improved performance compared to pairwise registration with VoxelMorph (delta-PSNR=1.6). We also trained a model that does not require the artefact-free reference image during inference demonstrating only marginally compromised results (delta-PSNR=-0.7).

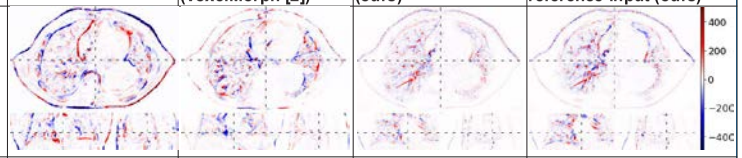
	Pairwise Deeds [3]	Pairwise U-Net (VoxelMorph [2])	Multiphase U-Net, (ours)	Multiphase U-Net, no reference input (ours)
Diff. to ground truth at max. amplitude				
PSNR	28.6	37.0	38.6	37.9
Runtime* (mm:ss)	10:40	00:04**	00:02	00:02

Table 1: Summary of results for our proposed registration methods compared to deeds and pairwise VoxelMorph registration algorithms. All metrics are averaged over a validation set of 12 unseen patients. All plots use level/window 0/1000 HU.
* Average runtime to register a single scan of 10 phases on a machine with 2xAMD EPYC 7513 32-Core CPUs (Deeds) and using a single Nvidia RTX A6000 (other methods)
** Pairwise registration of all phases requires 10 forward passes through the network

REFERENCES

[1] Paysan P et al. CT Based Simulation Framework for Motion Artifact and Ground Truth Generation of Cone-Beam CT. AAPM Annual Meeting 2019.
[2] Balakrishnan G et al. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. IEEE Transactions on Medical Imaging. 2019;38(8):1788-1800. doi:10.1109/TMI.2019.2897538
[3] Heinrich MP et al. MRF-based deformable registration and ventilation estimation of lung CT. IEEE Trans Med Imaging. 2013;32(7):1239-1248. doi:10.1109/TMI.2013.2246577
[4] Brehm M et al. Self-adapting cyclic registration for motion-compensated cone-beam CT in image-guided radiation therapy. Med Phys. 2012;39(12):7603-7618. doi:10.1118/1.4766435

ACKNOWLEDGEMENTS

Co-financed by Innosuisse, grant no. 35244.1 IP-LS.
S.S. and P.P. are full-time employees of Varian Medical Systems Imaging Laboratory GmbH

CONTACT INFORMATION

Lukas Lichtensteiger, ZHAW: licn@zhaw.ch

6 Research Output of the Natural Language Processing Group

6 Research Output of the Natural Language Processing Group

The NLP research group is led by Prof. Cieliebak and develops technologies for the analysis, understanding and generation of speech and text. We combine methods from linguistics, natural language processing (NLP) and artificial intelligence to enable natural language communication between humans and machines.

In our research, we work on topics such as text classification (e.g., sentiment analysis, hate speech detection), chatbots/dialogue systems, text summarization, speech-to-text, speaker recognition and natural language generation. At the end of 2022, the group consists of 9 researchers: 1 full professor, 3 senior scientists, 4 research assistants, and 1 IAESTE intern. Two of the research assistants pursue their master's degree, and there are four additional non-staff master's students.

The start of the *war in Ukraine* in February 2022 has deeply impressed our research team, and we thought about ways how we as NLP experts can help. This resulted in two projects: The first short-term project was a snapshot analysis of how news about the war spread on Twitter in the first months, and how this correlates to the attacks in real life. The second is a long-term project, where we are developing an information platform for Ukrainian refugees in Switzerland. The platform focusses on housing and working environments of refugees, and we are implementing a chatbot where they can find information about potential misuse and fraud patterns in these areas. This is a joint project with experts from Social Work ZHAW and funded by the "DIZH Rapid Action Call".

Swiss German has been a major topic within the research team for several years. In the past years, we have collected more than 500 hours of transcribed Swiss German audio, and manually annotated several thousand Swiss German texts. Using these, we have now built a speech-to-text system which can handle the seven major Swiss German dialects separately. In addition, we have implemented a translation system between standard German and Swiss German texts, and we organized a Shared Task on Swiss German text normalization. This is a joint research effort with FHNW, the University of Zurich, and the Swiss Association for Natural Language Processing (SwissNLP)

A breakthrough in NLP was the release of *Large Language Models*, in particular the *GPT* model in its various variants. We have built several showcases using these models, such as summarizing business meetings, generating images from a few keywords, and easily accessing information in databases.

In addition, we are working on several other topics such as automatic interview transcription, detection of malicious behavior on social media, a simulator for children in police interviews and organizing the SwissText conference, to name a few. One topic that rapidly gained momentum in 2022 was "*Theory of Evaluation*", where we are working on a fundamental framework for the evaluation of text generation systems (e.g., machine translation, summarization, chatbots etc.). We are proud that our work has been accepted and published at the ACL and the EMNLP conference, which are the two most prestigious conferences in NLP.

We would like to thank all project partners, collaborators, funding agencies and students for their support!

The NLP 2022 Team
Mark Cieliebak, Don Tuggener, Manuela Hürlimann, Jan Milan Deriu, Pius von Däniken, Katsiaryna Mlynchyk, Nicola Good, Daniel Neururer, and Simona Hovančíková

Probing the Robustness of Trained Metrics for Conversational Dialogue Systems

Jan Deriu, Don Tuggener, Pius von Däniken, Mark Cieliebak
Zurich University of Applied Sciences (ZHAW), Winterthur, Switzerland
deri@zhaw.ch

Abstract

This paper introduces an adversarial method to stress-test trained metrics to evaluate conversational dialogue systems. The method leverages Reinforcement Learning to find response strategies that elicit optimal scores from the trained metrics. We apply our method to test recently proposed trained metrics. We find that they all are susceptible to giving high scores to responses generated by relatively simple and obviously flawed strategies that our method converges on. For instance, simply copying parts of the conversation context to form a response yields competitive scores or even outperforms responses written by humans.

1 Introduction

One major issue in developing conversational dialogue systems is the significant efforts required for evaluation. This hinders rapid developments in this field because frequent evaluations are not possible or very expensive. The goal is to create automated methods for evaluating to increase efficiency. Unfortunately, methods such as BLEU (Papineni et al., 2002) have been shown to not be applicable to conversational dialogue systems (Liu et al., 2016). Following this observation, in recent years, the trend towards training methods for evaluating dialogue systems emerged (Lowe et al., 2017; Deriu and Cieliebak, 2019; Mehri and Eskenazi, 2020; Deriu et al., 2020). The models are trained to take as input a pair of context and candidate response, and output a numerical score that rates the candidate for the given context. These systems achieve high correlations to human judgments, which is very promising. Unfortunately, these systems have been shown to suffer from instabilities. (Sai et al., 2019) showed that small perturbations to the candidate response already confuse the trained metric. This work goes one step further: we propose a method that automatically finds strategies that elicit very high scores from the trained metric while being of obvious low quality. Our method can be applied to automatically test the robustness of trained metrics against adversarial strategies that exploit certain weaknesses of the trained metric.

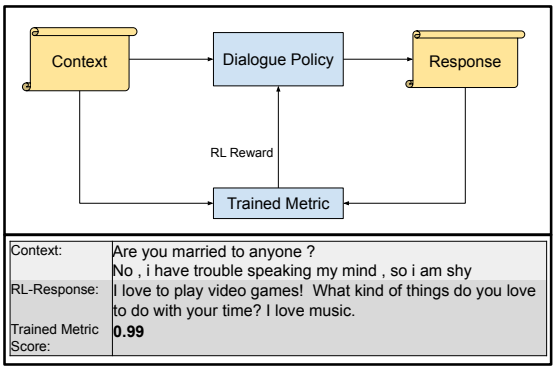


Figure 1: Overview of the process. It takes a context and an response generated by a dialogue policy and computes a score based on the trained metric. The score is then used as a reward to update the policy. In this example, the policy converges to a fixed response, which achieves an almost perfect score, although it is clearly a low-quality response. The policy always returns this response, regardless of the context, and the trained metric always scores it perfectly.

Our method uses a trained metric as a reward in a Reinforcement Learning setting, where we fine-tune a dialogue system to maximize the reward. Using this approach, the dialogue system converges towards a degenerate strategy that gets high rewards from the trained metric. It converges to three different degenerate types of strategies to which the policy converges in our experiments: the *Parrot*, the *Fixed Response*, and the *Pattern*. For each dataset and metric, an adversarial response is found, which belongs to one of the three strategy types. The responses generated from these strategies then achieve high scores on the metric. Even more, in most cases, the scores are higher than the scores achieved by human written responses. Figure 1 shows the pipeline. The dialogue policy receives a reward signal from the trained metric.

Over time, the policy converges to a fixed response, which objectively does not match the context but gets a near-perfect score on the trained metric. We release the code [1].

2 Related Work

Trained Metrics. In recent years the field of trained metrics gained traction after word-overlap methods have been shown to be unreliable (Liu et al., 2016). The first of these metrics is ADEM (Lowe et al., 2017), which takes as input a context, a reference, and the candidate response and returns a score. The main issue with ADEM is the reliance on references and annotated data (i.e., human ratings of responses), which are costly to obtain, and need to be redone for each domain. RUBER (Tao et al., 2018) extended ADEM by removing the reliance on annotated data for training. However, it still relies on a reference during inference. AutoJudge (Deriu and Cieliebak, 2019) removed the reliance on references, which allows the evaluation of multi-turn behavior of the dialogue system. However, AutoJudge still leverages annotated data for training. USR (Mehri and Eskenazi, 2020) is a trained metric that does not rely on either annotated data or any reference. It is trained in a completely unsupervised manner while still highly correlated to human judgment (0.4 Spearman Correlation). Similarly, MAUDE (Sinha et al., 2020) is trained as an unreferenced metric built to handle the online evaluation of dialogue systems.

Robustness of Trained Metrics. There is not yet much research on the robustness of trained metrics. Sai et al. (2019) evaluated the robustness of ADEM by corrupting the context in different ways. They show that by just removing punctuation, the scores of ADEM change, and in 64% of cases are superior to the scores given for the same response without removed punctuation. Other corruption mechanisms yielded similar results. Yeh et al. (2021) compared a large variety of automated metrics for dialogue system evaluation by comparing, e.g., turn- and dialogue-level correlation with human judgemnts and studying the impact of the dialogue length. They find that no single metric is robust against all alternations but see potential in ensembling different metrics. Novikova et al. (2017) investigate automated metrics in the task-oriented NLG domain and find that the metrics do not sufficiently reflect human ratings.

[1] https://github.com/jderiu/metric-robustness

Algorithm 1: Advantage Actor-Critic Algorithm, where π_θ denotes the policy, c denotes the context, r the response generated by the policy, and s denotes the score by the automated metric, i.e., the reward.

while *training* **do**
  sample c from pool of contexts;
  $r = \pi_\theta(c)$ generate response;
  $s = R(c, r)$ compute reward;
  fit action-value function Q_σ i.e., $\mathcal{L}(\sigma) = \frac{1}{2}\sum_i \left\| R(c,r) + Q_(c', r') - Q_\sigma(c,r) \right\|$; compute the advantage $A(r,c) = R(r,c) - Q(c,r) + Q(c',r')$;
  $\theta = \theta + \alpha \bigtriangledown J_{RL}(\theta)$ fit policy;
end

3 Method

Our method applies a trained metric as a reward signal $R(c, r)$ to update a dialogue system $\pi(c)$ in a reinforcement learning setting, where c denotes the context and r the response. The dialogue system is trained by generating a response for a context, which is then scored by the automated metric. The dialogue system is then updated using the score as the reward. This process is repeated for different contexts. We use the Actor-Critic framework to optimize the policy (Sutton et al., 1999). See Algorithm 1 for an overview. The policy gradient is defined as $\bigtriangledown J_{RL}(\theta) = \bigtriangledown_\theta log\ \pi_\theta(r|c) * A(r,c)$, where $\pi_\theta(r|c)$ defines the probability of the generated response for the given context, and $A(c, r)$ the advantage function.

The learned policy depends on the reward function, i.e., the automated metric. If the reward function is susceptible to adversarial attacks, the policy will likely generate an objectively suboptimal solution, which is rated highly by the automated metric. Conversely, we expect the policy to improve the dialogue systems' responses if the automated metric is robust against adversarial examples.

4 Experimental Setup

4.1 Datasets

We perform the evaluation on three widely-used datasets in the dialogue modelling domain. Namely, Dailydialog (Li et al., 2017), Empathetic Dialogues (Rashkin et al., 2019), and PersonaChat (Zhang et al., 2018).

Metric	Strategy	Response
		PersonaChat
ATT	Fixed	yea!!! 1!! 2!! 3!! * * * fucking fucking fucking * * [[[fucking * fucking *
BLM	Fixed	that sounds like a lot of fun. what do you like to do in your spare time?
MAUDE	Fixed	What kind of work do you have? What do you like to do in your free time?
USR FULL	Parrot	-
USR MLM	Fixed	i am a stay at home mom and i am trying to figure out what i want to do with my life
USR RET	Fixed	I love to be a musician. I love music. What kind of music do you listen to as a music lover
		Dailydialog
ATT	Fixed	! freaking out! one of these days! * * one * * freaking * * out! * even * * damn * * even damn
BLM	Fixed	that would be great! what do you do for a living, if you don't mind me asking?
MAUDE	Fixed	I hope it works out for you. What kind of car did you get?
USR FULL	Pattern	i'm not sure if i'd like to [copy context tokens]. i'll let you know if i do.
USR MLM	Fixed	i am not sure if i am going to be able to go out of my way to get to know each other or not.
USR RET	Parrot	-
		Empathetic Dialogues
ATT	Fixed	I know right? I felt SO SO ASHAmed of myself. I felt so embar assed.
BLM	Fixed	I'm so sorry to hear that. What happened, if you don't mind me asking?
MAUDE	Fixed	I wish I could go back in time and be a kid again. I miss those days.
USR FULL	Pattern	i don't think it's [random context noun]. i'm sorry to hear that. what do you mean by that?
USR MLM	Fixed	I don't know what I'm going to do if it doesn't work out. I'm not sure what to do.
USR RET	Parrot	-

Table 1: The strategies achieved for each metric and domain.

4.2 Metrics

We use various state-of-the-art automated metrics developed for evaluating conversational dialogue systems without reference, i.e., so-called unreferenced metrics.. These are metrics where no reference is needed, i.e. they only use the context and response to determine the score. They can be represented as a function $s = R(c, r)$, which rate the response r for a given context c.

We selected state-of-the-art trained metrics which achieve good correlations to human judgments to evaluate our approach—namely, USR (Mehri and Eskenazi, 2020), ATT (Gao et al., 2021), and MAUDE (Sinha et al., 2020). Additionally, we added the Blender language model score (BlenderLM) (Roller et al., 2020). For the ATT [2], MAUDE [3], and BlenderLM metrics [4], we use the out-of-the-box models provided by the respective authors. For the USR metric, we perform custom training on each dataset. Furthermore, we report the USR-retrieval (*USR Ret*), USR-masked-language-model *USR MLM*, and the USR-regression *USR Full* scores. Note that the *USR Full* is a combination of the *USR Ret* and *USR MLM* metric. More details can be found in Appendix A.

[2] https://github.com/golsun/AdversarialTuringTest

[3] https://github.com/facebookresearch/online_dialog_eval

[4] https://huggingface.co/facebook/blenderbot-400M-distill

4.3 Strategies

For our approach, we use Blenderbot as our policy (Roller et al., 2020) since it is currently a state-of-the-art conversational dialogue system [5]. We use the validation set for each domain to perform reinforcement learning. This is to avoid the dialogue systems being fine-tuned on already seen data. We use the test set to evaluate the reward over the number of episodes. We perform the reinforcement learning for 15 epochs, where each epoch is composed of 500 updates. We noted from pre-experiments that this is enough for a dialogue system to converge to a degenerate strategy. We track the average reward achieved on the test set after each epoch. Each experiment is repeated 10 times since we expect the policy to converge to slightly different strategies in different runs. We select the repetition which achieved the highest score (i.e., reward) and use it to determine the strategy. We also experimented with automated strategy detection, see Appendix B.

5 Results

The policies typically converge towards one of the following three degenerate strategies.

Parrot. Here, the policy simply copies parts of the context into the response. Sometimes, it applies slight changes. For instance, it changes the pronouns from "you" to "I".

Fixed Response. Here, the policy converges on a fixed response which it returns regardless of the

[5] Note that here we are referring to Blenderbot as a dialogue system. BLM is using the Blenderbot LM as a metric.

Dailydialog						
	USR RET	USR MLM	USR FULL	ATT	MAUDE	BLM
BL	0.440	0.426	4.951	0.0002	0.664	0.096
HU	0.928	0.409	7.904	0.0006	0.898	0.183
COPY	0.998	0.811	9.429	0.0002	0.921	0.233
FIXED	-	**0.505**	-	**0.435**	**0.985**	**0.239**
PARROT	**0.998**	-	-	-	-	-
PATTERN	-	-	**7.091**	-	-	-
Empathetic Dialogues						
	USR RET	USR MLM	USR FULL	ATT	MAUDE	BLM
BL	0.935	0.298	7.645	0.001	0.820	0.087
HU	0.891	0.384	7.611	0.120	0.942	0.264
COPY	0.996	0.885	9.617	0.054	0.935	0.358
FIXED	-	**0.912**	-	**0.731**	**0.976**	**0.333**
PARROT	**0.994**	-	-	-	-	-
PATTERN	-	-	**7.240**	-	-	-
PersonaChat						
	USR RET	USR MLM	USR FULL	ATT	MAUDE	BLM
BL	0.847	0.185	6.797	0.0006	0.844	0.070
HU	0.927	0.267	7.512	0.0024	0.951	0.153
COPY	0.925	0.794	8.933	0.0001	0.898	0.223
FIXED	**0.977**	**0.852**	-	**0.813**	**0.933**	**0.250**
PARROT	-	-	**7.542**	-	-	-
PATTERN	-	-	-	-	-	-

Table 2: Scores achieved by humans (HU), Blenderbot (BL) and the degenerate strategies with regard to the different metrics for each domain.

context.

Pattern. This is a mix between the *Parrot* and the *Fixed Response*. It creates a fixed template filled with parts of the context.

Table 1 shows the selected responses for each pair of domain and metric. For all metrics except *ATT*, the fixed response is composed of a grammatically correct sentence. Note that these responses are always returned by the fine-tuned dialogue system, regardless of the context.

5.1 Scores

Table 2 shows the main results. In almost all cases, the degenerated strategy outperforms the vanilla Blenderbot and humans with respect to the automated metric. The most striking example is the *ATT* metric, where the fixed response achieves scores by orders of magnitude better than the ones achieved by humans. For both *USR Ret* and *MAUDE*, the scores achieved by the fixed response are almost perfect, i.e., they are close to 1.0, which is the upper bound. Also, for *USR MLM*, the scores are significantly higher than the ones achieved by Blenderbot. Interestingly, the *USR FULL* seems to be more immune to the pattern that were found. However, even for *USR FULL*, the parrot strategy beats the humans by a significant margin in the *PersonaChat* domain.

Copy. We also display the scores achieved by simply copying the context on each metric, which is inspired by the *Parrot* strategy. The only metric which is immune to the *Copy* strategy is *ATT*. Under all the other metrics, the *Copy* achieves very high scores. In some cases, it achieves even better scores than the converged policy. For instance, for the *Dailydialog* domain, it achieves 0.811 points under the *USR MLM* metric, which is 0.3 point higher than the converged policy and twice as good as the human score.

6 Conclusion

Trained metrics for automatic evaluation of conversational dialogue systems are an attractive remedy for the costly and time-consuming manual evaluation. While high correlation with human judgments seems to validate the metrics regarding their ability to mimic human judging behavior, our analysis shows that they are susceptible to rather simple adversarial strategies that humans easily identify. In fact, all metrics that we used failed to recognize degenerate responses. Our approach is easily adaptable to any newly developed trained metric that takes as input a pair of context and response. There are no known remedies for this problem. Thus, the next open challenge is to find methods that improve the robustness.

References

Jan Deriu and Mark Cieliebak. 2019. Towards a Metric for Automated Conversational Dialogue System Evaluation and Improvement. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 432–437, Tokyo, Japan. Association for Computational Linguistics.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on Evaluation Methods for Dialogue Systems. *Artificial Intelligence Review*, pages 1–56.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS '18 Competition*, pages 187–208, Cham. Springer International Publishing.

Xiang Gao, Yizhe Zhang, Michel Galley, and Bill Dolan. 2021. An adversarially-learned turing test for dialog generation models. *arXiv preprint arXiv:2104.08231*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language Models are Unsupervised Multitask Learners.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Ananya B Sai, Mithun Das Gupta, Mitesh M Khapra, and Mukundhan Srinivasan. 2019. Re-Evaluating ADEM: A Deeper Look at Scoring Dialogue Responses. In *Proceedings of the thirty-third AAAI Conference on Artificial Intelligence*, volume 33 of *AAAI'19*, pages 6220–6227, Honolulu, Hawaii, USA.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. *arXiv preprint arXiv:2005.00583*.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 1057–1063, Cambridge, MA, USA. MIT Press.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. In *Proceedings of the thirty-second AAAI Conference on Artificial Intelligence*, AAAI'18, New Orleans, Louisiana USA.

Yi-Ting Yeh, Maxine Eskénazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. *ArXiv*, abs/2106.03706.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Correlation between Human Judgements and Trained Metrics

In this section, we evaluate the metrics with regards to their correlation to human judgments to show that these metrics have reasonable performance. For this, we sample 100 contexts for each domain. For each domain, we use a set of bots to create a response for each context. Furthermore, we add the human response to the pool of responses for each context. Then, we let crowdworkers annotate the responses. We correlate the scores of each metric on the same set of contexts and responses to the human annotations.

A.1 Domains and Bots

We perform the evaluation on the three datasets from the main paper.

Dailydialog. We prepared 5 bots using ParlAI (Miller et al., 2017). We fine-tune a GPT-2 (GPT) model (Radford et al., 2018), a BERT-Rank (BR) model, a sequence-to-sequence model (S2) with attention, and a weakly trained sequence-to-sequence model (DR). We also use the Blender model (Roller et al., 2020), although it was not specifically tuned on Dailydialog.

Empathetic Dialogues. We prepared the same pool of models as in Dailydialog.

PersonaChat. We mostly reuse the openly available systems of the ConvAI2 challenge (Dinan et al., 2020), namely, Lost in Conversation[6] (LC) and Huggingface (HF) [7] , and KVMemNN (KV). We also add the Blender model, which is also trained in this domain, a custom-trained BERT-Rank model (BR), and a sequence-to-sequence model (S2). Together with the DR model, the pool consists of 7 different dialogue systems.

A.2 Annotation Process

Since we perform the evaluation on a static-context setting, we also add the human response (i.e., the gold response) to the pool of systems. For evaluation, we use 600 samples for Dailydialog and Empathetic Dialogues each, and 800 samples for the PersonaChat domain. Each sample is composed of a context (sampled from the test set), and a generated response. We annotated the overall quality of each sample on a Likert scale from 0 (bad) to

[6] https://github.com/atselousov/transformer_chatbot

[7] https://github.com/huggingface/transfer-learning-conv-ai

	DD	ED	PC
USR RET	0.561	0.524	0.605
USR MLM	0.138	0.452	0.303
USR REG	0.559	0.573	0.585
ATT	0.154	0.385	-0.099
MAUDE	0.211	0.086	0.357
BLENDERLM	0.201	0.287	0.266

Table 3: Correlations of the automated metrics to human judgments. For all runs $p < 0.05$.

2 (good) using Mechanical Turk[8]. Each sample is annotated by three different humans. As the final score, we use the average score of the three annotations. For each metric, we apply the metric to all samples, and then compute the Spearman correlation between the human scores and the scores predicted by the metric.

A.3 Correlation to Human Judgements

Table 3 shows the correlations of the human judgments to each of the metrics for each domain. For all domains, the *USR* metric performs best, achieving strikingly high correlations to humans. *MAUDE* also achieves good correlation scores on the PersonaChat domain, and *ATT* performs well on the Empathetic Dialogues domain. *BlenderLM* has mediocre performance on all domains equally.

A.4 Original USR

Note that the *USR Ret* scores are significantly higher than in the original paper (Mehri and Eskenazi, 2020), which is due to the fact that we use more turns to represent the context, whereas the original implementation uses only the previous turn for the context. In the original implementation, *USR Ret* achieves a Spearman correlation of 48.67 on our annotated data. If we train our implementation of *USR Ret* using only one turn to represent the context, we also achieve a Spearman correlation of 40.34, which is comparable to the original. We did not experience a discrepancy on the *USR MLM* model, where the original model achieves the same correlation as ours.

B Strategy Selection

We observed in our experiments that the dialogue system almost always converges to one of three degenerate strategies. In order to atomize their detection in the experiments, we used a set of heuristics for their identification.

[8] https://www.mturk.com/

B.1 Heuristics

Since the strategies are very simple, we propose heuristics to detect the policy automatically. This avoids the need for manual inspection of a potentially large amount of log files. For this, we introduce the following measures.

- *Response Frequency.* The percentage of times that the same response is generated for all samples in the test set.
- *Lexical Variety.* The ratio between number of different tokens and the total number of tokens over all responses in the test set.
- *BLEU score.* The BLEU score between the context and the response. This is computed for each pair of context and responses and then averaged over all samples in the test set.
- *Jaccard score.* The Jaccard overlap between the context and response tokens. Analogous to the BLEU score, the Jaccard overlap is computed between each context-and response-pair, and then averaged over all samples in the test set.

These measures can be used to detect the various strategies the policy converges to. For instance, a high *Response Frequency* indicates that the policy converges to a fixed response. A high *BLEU* score and *Jaccard score* indicate that the policy converges to the parrot strategy. A low *Response Frequency*, a low *Lexical Variety* and a moderate *Jaccard score* indicate that the policy converges to a pattern. A pattern is composed of a fixed template where parts are filled with tokens from the context.

B.2 Application of the Heuristics

For each run, we use these metrics to determine which strategy the policy has converged on. The final strategy is extracted by selecting the best epoch across all 10 runs for each domain. If the *Response Frequency* is larger than 0.7, we extract the most common sentence and use this as our fixed response. If the *BLEU* score is larger than 0.2, we assign the parrot strategy. If the *Response Frequency* is smaller than 0.1, the *Lexical Variety* is smaller than 0.15, and the *Jaccard score* is larger than 0.05, it indicates a pattern emerged. In this case, we manually extract the pattern.

B.3 Overview

Table 4 shows the measures used to perform the automated strategy selection. The automated strategy

domain	metric	Avg Reward	Resp Freq	Lex Var	BELU	Jacccard	Strategy Inferred	Strategy Manual	Strategy Final
Persona Chat	ATT	0.77	0.14	0	0	0	Not Conclusive	Fixed Response	Fixed Response
Persona Chat	BLM	0.41	0.01	0.11	0.03	0.06	Not Conclusive	Fixed Response	Fixed Response
Persona Chat	MAUDE	0.98	0.7	0.01	0	0.07	Fixed Response		Fixed Response
Persona Chat	USR Full	7.7	0	0.09	0.42	0.48	Parrot		Parrot
Persona Chat	USR MLM	0.84	0.94	0.01	0.01	0.1	Fixed Response		Fixed Response
Persona Chat	USR Ret	1	0.8	0	0	0.07	Fixed Response		Fixed Response
Dailydialog	ATT	0.42	0.55	0.01	0	0.01	Not Conclusive	Fixed Response	Fixed Response
Dailydialog	BLM	0.26	0.32	0.01	0	0.05	Not Conclusive	Fixed Response	Fixed Response
Dailydialog	MAUDE	0.99	0.99	0	0	0.06	Fixed Response		Fixed Response
Dailydialog	USR Full	7.65	0	0.11	0.08	0.15	Pattern		Pattern
Dailydialog	USR MLM	0.52	1	0	0	0.04	Fixed Response		Fixed Response
Dailydialog	USR Ret	0.99	0	0.19	0.21	0.31	Parrot		Parrot
Empathetic Dialogues	ATT	0.78	0.98	0	0	0.04	Fixed Response		Fixed Response
Empathetic Dialogues	BLM	0.33	0.47	0.03	0	0.05	Not Conclusive	Fixed Response	Fixed Response
Empathetic Dialogues	MAUDE	0.98	0.96	0	0	0.06	Fixed Response		Fixed Response
Empathetic Dialogues	USR Full	8.67	0.01	0.07	0.04	0.1	Pattern		Pattern
Empathetic Dialogues	USR MLM	0.77	0.98	0	0	0.06	Fixed Response		Fixed Response
Empathetic Dialogues	USR Ret	1	0	0.17	0.33	0.44	Parrot		Parrot

Table 4: Scores achieved on the test set during the evaluation.

selection worked in 72% of cases. There are two main cases in which it was not conclusive. First, for the *ATT* metric, where for both the *Dailydialog* and *PersonaChat* domains no clear fixed response arose. However, after manual inspection, we noted that for the *PersonaChat* the policy generated the same tokens in various frequencies and orders. For the *Dailydialog* the most frequent response arose in 55% of cases. Thus, we used this fixed response. The second case is the *BLM* metric. For all the domains we selected the most frequent response, although it appeared in less than 70% of cases.

C Full Results

Table 5 shows all scores achieved by the dialogue systems on the respective metrics. Furthermore, we also added the average score of the Amazon Mechanical Turk judges, which ranges from (0-2).

D Technical Explanation

One potential reason why our approach is able to find a degenerate strategy lies in the exploration problem in reinforcement learning. Blender's language model can be interpreted as a policy which performs a sequence of actions, i.e., sampling a sequence of tokens. Thus, the language model loss during standard Blender training can be interpreted as an indicator for how sure the policy is of its actions. A high language model loss indicates that the policy assigns low probability scores to its actions. Conversely, a low language model loss indicates that the policy is sure of it's actions. This could be further investigated by measuring the entropy of the language model. Indeed, in all our experiments, we notice that the language model loss collapses toward a very small value. This indicates that the language model collapsed to a single simple strategy. Figure 2 shows the language model loss over the number of steps. The loss quickly collapses from an average of 4 points to around 0.5 points. At the same time the average reward (orange) rises from 0.78 to 0.92. Similarly, the response frequency rises from 0 to 0.94. In the middle, the loss rises again, which indicates the search for a new strategy. This coincides with a lower response frequency.

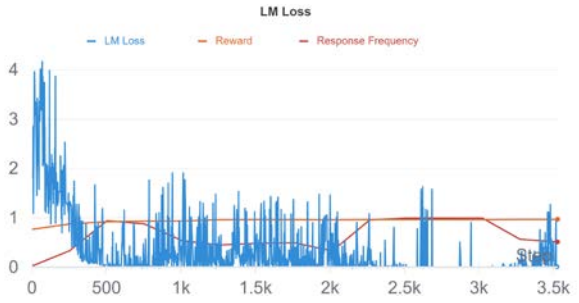


Figure 2: The language model loss (blue), the Average Reward (orange), and the Response Frequency (red) over time.

E Examples

In Tables 6, 7, and 8, we show examples of the outputs from the fine-tuned Blenderbot model. For each of the five metrics, we show the output to which Blenderbot converged to when using the metric as a reward. Furthermore, we show the score which the respective metric assigns to the generated response. Note that the *Parrot* strategies simply copy the text form the context. For the *Empathetic Dialogues* dataset, the degenerate strategy prepends a "I'm not sure" to the context. For the *PersonaChat*, the degenerate strategy prepends a "i've always wanted to". The *Copy* strategy (see Table 2 in main Paper), ignores these prefaces, and simply copies the context.

Dailydialog

	AMT	USR RET	USR MLM	USR FULL	ATT	MAUDE	BLM
BR	1.836	0.928	0.409	7.904	0.0006	0.898	0.177
BL	1.386	0.440	0.426	4.951	0.0002	0.664	0.096
HF	1.656	0.925	0.080	6.989	0.0026	0.866	0.371
HU	1.782	0.928	0.409	7.904	0.0006	0.898	0.183
S2	1.024	0.512	0.300	5.050	0.0003	0.895	0.183
DR	0.729	0.308	0.338	3.900	0.0001	0.891	0.204
PARROT	-	**0.998**	*0.811*	*9.429*	*0.0002*	*0.921*	*0.233*
FIXED	-	-	**0.505**	-	**0.435**	**0.985**	**0.239**
PATTERN	-	-	-	**7.091**	-	-	-

Empathetic Dialogues

	AMT	USR RET	USR MLM	USR FULL	ATT	MAUDE	BLM
BR	1.808	0.891	0.384	7.611	0.120	0.942	0.260
BL	1.640	0.935	0.298	7.645	0.001	0.820	0.087
HF	1.610	0.887	0.644	8.292	0.044	0.948	0.462
HU	1.816	0.891	0.384	7.611	0.120	0.942	0.264
S2	0.702	0.493	0.145	4.510	0.010	0.932	0.159
DR	0.822	0.354	0.182	3.759	0.001	0.936	0.199
PARROT	-	**0.996**	*0.8848*	*9.617*	*0.054*	*0.935*	*0.358*
FIXED	-	-	**0.912**	-	**0.731**	**0.976**	**0.333**
PATTERN	-	-	-	**7.240**	-	-	-

PersonaChat

	AMT	USR RET	USR MLM	USR FULL	ATT	MAUDE	BLM
BR	1.350	0.725	0.211	6.120	0.0020	0.946	0.138
BL	1.507	0.847	0.185	6.797	0.0006	0.844	0.070
HF	1.480	0.794	0.272	6.707	0.0023	0.925	0.152
HU	1.623	0.927	0.267	7.512	0.0024	0.951	0.153
KV	1.147	0.538	0.217	4.982	0.0023	0.852	0.122
LC	1.572	0.879	0.103	6.769	0.0011	0.918	0.195
S2	0.681	0.390	0.121	3.814	0.0013	0.845	0.111
DR	0.906	0.482	0.268	4.779	0.0002	0.537	0.220
PARROT	-	*0.925*	*0.794*	**8.933**	*0.0001*	*0.898*	*0.223*
FIXED	-	**0.977**	**0.852**	-	**0.813**	**0.933**	**0.250**
PATTERN	-	-	-	-	-	-	-

Table 5: Scores achieved by the various dialogue policies with regard to the different metrics. For each metric, we show the average scores achieved by each dialogue system and humans (HU) for each domain. We then show the average score achieved by applying the strategy to which the policy converged to. The AMT metric denotes Amazon Mechanical Turk ratings.

Turn	Utterance	Score	Strategy
usr_ret			
1	Did this just come on suddenly ?		
2	I felt OK until an hour or so ago , and then I just doubled over .		
3	Have you exercised strenuously or played sports recently ?		
RL-Response	i feel ok until an hour or so ago, and then i just doubled over. have you exercised strenuously or played sports recently	0.9986	Parrot
usr_mlm			
1	Yes . Who is speaking ?		
2	This is Kelly speaking . I am calling from London . How are you , Mr . Chang ?		
3	Oh , Kelly , I couldn't recognize your voice . Well , I'm fine .		
RL-Response	i am not sure if i am going to be able to go out of my way to get to know each other or not.	0.4820	Fixed
usr_full			
1	You really shouldn ' t write it down . You should memorise it . You can change it to a more convenient number if you like .		
2	Ok , I ' Ve entered my PIN . There are several options on the screen . I guess I just choose which service I want , don ' t I ?		
3	Yes . Choose the service you need to use . If you choose to withdraw money , the machine will ask you how much you want .		
RL-Response	i'm not sure if i'd like to change my pin to a more convenient number if i like. i'll let you	7.466	Pattern
att			
1	Hello , hot water overflowed from the bathtub . Would you help me ?		
2	Would you turn off the tap ?		
RL-Response	hahaha!!! thats so funny!! im laughing so hard!! its so true!! ive been laughing so	0.1176	Fixed
maude			
1	Good morning , Mr Zhang , I wonder whether you have got an answer from your superior .		
2	Yes , we had a meting about it yesterday afternoon .		
RL-Response	I hope it works out for you. What kind of car did you get?	0.9901	Fixed
blender			
1	There is a Matisse exhibition at the Museum of Modern Art . Larry and I will go .		
2	Did you buy tickets already ?		
3	No . Do we need tickets ?		
RL-Response	what do you like to do in your spare time? do you have any hobbies?	0.2078	Fixed

Table 6: Randomly sampled responses for the Dailydialog dataset. We randomly chose a context, and let the fine-tuned Blenderbot create a response, which at this point is degenerate. We then let the metric which was used as a reward score rate the generated response.

Turn	Utterance	Score	Strategy
usr_ret			
1	My boyfriend's brothers got into a fist fight one night and he jumped in the middle.		
2	That sucks. Did anyone get hurt?		
3	Some things got broke and there were some bruises, but it was really scary to watch that.		
RL-Response	i'm sure your boyfriend's brothers got into a fist fight one night and he jumped in the middle. did anyone get hurt?	0.9982	Parrot
usr_mlm			
1	Financially, things will turn around for my family soon!		
2	I hope and wish so. In this tough economic climate many are finding it difficult to cope with financial responsibilities		
3	Yes, and we had a layoff of five months. I believe in a brighter day, though, so I know it will improve.		
RL-Response	I don't know what I'm going to do if it doesn't work out. I'm not sure what to do.	0.7689	Fixed
usr_full			
1	I had a major project and felt very ready to present it to everyone at the office		
2	How'd it go		
3	It went well, I think I gained confidence by being so prepared		
RL-Response	i don't think i gained confidence by being so prepared. i'm sorry to hear that. what do you mean by that?	9.068	Pattern
att			
1	I once got in a really bad accident on my way home from work and it really upset me		
2	Wow! Were you hurt? Was anyone else hurt?		
3	luckily the only problem I had was my neck was sore for a few weeks but nobody else was hurt		
RL-Response	I know right? I felt SO SO ASHAmed of myself. I felt so embar assed.	0.8138	Fixed
maude			
1	like who on earth writes var int value = 1, this is so ridiculous i am so annoyed at my brother, after everything i tought him, he doesn't even know the basics		
RL-Response	I wish I could go back in time and be a kid again. I miss those days.	0.9946	Fixed
blender			
1	My friend left my front door open when she was house-sitting for me. My cat snuck out and was killed.		
RL-Response	i feel like i know what i'm going to do, but i don't know what to do with myself.	0.1299	Fixed

Table 7: Randomly sampled responses for the Empathetic Dialogues dataset. We randomly chose a context, and let the fine-tuned Blenderbot create a response, which at this point is degenerate. We then let the metric which was used as a reward score rate the generated response.

Turn	Utterance	Score	Strategy
usr_ret			
1	nemo , i love him already . its a tea mix , green and black .		
2	the tea sounds cool . is there honey in it ?		
3	i add honey to it , yes . i drink it while running museum tours .		
RL-Response	I love to be a musician. I love music. What kind of music do you listen to as a music lover	0.9997	Fixed
usr_mlm			
1	hi i am a musician with big dream		
RL-Response	i am a stay at home mom and i am trying to figure out what i want to do with my life	0.8854	Fixed
usr_full			
1	it is tough and smelly . 25 people and 3 toilets .		
2	that would be awful . you really should move .		
3	i cant yet . need to save money , spend it too quickly .		
RL-Response	i've always wanted to move 25 people and 3 toilets that would be awful you really should move i cant yet need to save money	8.498	Parrot
att			
1	me too ! i aspire to become a physical therapist and run a 5k !		
2	awesome . i just started a new job as a life coach		
3	that is great ! i get straight as so maybe you could coach me in social skills .		
RL-Response	yea!!! 1!! 2!! 3!! * * * [* * fucking * fucking fucking * [fucking * *	0.7417	Fixed
maude			
1	how are you doing today		
2	hello , how are you today ?		
3	i am doing really well		
RL-Response	What are you up to today? What do you like to do in your free time?	0.9923	Fixed
blender			
1	haha agreed ! or a lot . either way . i try to eat well though		
2	me too . should i trade my mini van in for a sports car .		
3	yes . who does not love a sports car .		
RL-Response	what kind of mini van do you have? i have a corvette and a camaro	0.1970	Fixed

Table 8: Randomly sampled responses for the PersonaChat dataset. We randomly chose a context, and let the fine-tuned Blenderbot create a response, which at this point is degenerate. We then let the metric which was used as a reward score rate the generated response.

Detection of Typical Sentence Errors in Speech Recognition Output

Bohan Wang* EPFL bohan.wang@epfl.ch
Ke Wang* EPFL k.wang@epfl.ch
Siran Li* EPFL siran.li@epfl.ch
Mark Cieliebak ZHAW ciel@zhaw.ch

Abstract

This paper presents a deep learning based model to detect the completeness and correctness of a sentence. It's designed specifically for detecting errors in speech recognition systems and takes several typical recognition errors into account, including false sentence boundary, missing words, repeating words and false word recognition. The model can be applied to evaluate the quality of the recognized transcripts, and the optimal model reports over 90.5% accuracy on detecting whether the system completely and correctly recognizes a sentence.

1 Introduction

Automatic Speech Recognition (ASR) systems develop technologies to recognize and translate spoken language into text by machines (Yu and Deng, 2016). Sentence error detection on ASR systems is important for the two reasons: a) This can help to set proper punctuation marks; b) For multiple speakers, speaker recognition often fails at the change between two speakers, which results in single words at beginning or end of an utterance being assigned to the wrong person. A practical application domain of our work is to detect complete and correct sentences in ASR systems to mitigate the aforementioned problems.

In prior works, research focused mainly on grammatical error detection (Agarwal et al., 2020; He, 2021). In this paper, we focus on dealing with the specific errors emerging in speech recognition, such as missing words or incorrect sentence boundaries (detailed in Sec. 3.3). In addition, previous works on enriching speech recognition emphasize on finding correct sentence boundaries in whole transcripts (Liu et al., 2006, 2005). However, in real-time speech recognition, we have access to only individual sentences instead of full transcripts, and they don't take other typical speech recognition errors (apart from incorrect sentence boundaries) into account (Tuggener and Aghaebrahimian, 2021).

Recently, transformer models have shown state-of-art performance in generating word embeddings and extracting intrinsic features of word sequences. In specific, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), Generative Pre-trained Transformer (GPT) (Radford et al., 2019) and BIG-BIRD (Zaheer et al., 2020) have achieved promising performance to learn high quality language representations from large amounts of raw text. The token representations produced by these transformers pre-trained on unsupervised tasks also help improve the performance of a supervised downstream task.

In this paper, we fine-tune the pre-trained transformers (BERT, GPT2 and BIG-BIRD) on the speech recognition error detection task, to build a binary classification model detecting speech recognition errors. The performance of sequentially linking BERT embedding and a down-stream text classification network is also studied. We compare and analyze the performances of several classification models. The models are ensembled through a Random Forest to further improve the performance. Finally, we analyse the performance of BERT-based classifier on a multi-label dataset.

The paper is structured as follows: In Sec. 2, we explain the models and experimental design. In Sec. 3, we describe how the dataset is generated. We discuss the experimental results in Sec. 4.

*These authors contributed equally to this work.

2 Methods

2.1 Models

In this section, we use three state-of-art transformer models BERT (Devlin et al., 2018), GPT2 (Radford et al., 2019), BIG-BIRD (Zaheer et al., 2020) are considered.

Besides, we also test the performance of using BERT embedding plus a downstream text classification network. For the classification networks, we use either a bi-direction LSTM and a TextCNN. We use a one-layer TextCNN with kernels sizes to be $2, 3$ and 4. For LSTM, we use a one-layer bi-directional LSTM network (Gers et al., 2000), followed by an attention layer and a fully connected layer. The number of hidden states is 256. Specifically, the attention layer is found to be essential.

2.2 Ensemble learning

We ensemble the five trained classifiers with random forest. Configuration and the final classification performance are shown in Sec. 4.2.

3 Data preparation

3.1 Dataset sources

For the model to have better generalizing capacity, a training set from diverse sources covering diverse topics and occasions is necessary. The following corpora are included in our proposed dataset:

News reports (Thompson, 2017): 143, 000 articles from 15 American publications

Ted 2020 Parallel Sentences Corpus (Reimers and Gurevych, 2020): around 4000 TED Talk transcripts from July 2020

Wikipedia corpus (Foundation): over 10 million topics

Topical-Chat (Gopalakrishnan et al., 2019): nearly 10 thousand human dialog conversations spanning 8 broad topics

3.2 Dataset Creation

To make the selected datasets suit our speech recognition model, we remove some non-English tokens, sentence ending symbols ('.', '!', '?'), duplicated sentences and also short sentences (less or equal to 5 words) to avoid some recognition errors. After pre-processing on the data from the sources, we create the following two datasets:

Standard Dataset: contains 0.3 million sentences from News reports, 0.3 million sentences from Ted corpus, 0.3 million sentences from Wikipedia corpus, 0.2 million sentences from Topical-Chat, in total 1.1 million sentences. We split the Standard Dataset randomly over all data sources into train set, ablation set and test set, with a proportion of 8:1:1.

Large Dataset: contains 2.3 million sentences from News reports, 0.4 million sentences from Ted corpus, 2 million sentences from Wikipedia corpus, 0.2 million sentences from Topical-Chat; in total 5 million sentences. We split it into train and test set, with a proportion of 19:1.

We train and compare performances of various models on the Standard Dataset. As a comparison, we evaluate the performance of BERT trained on the large dataset to see how an enlarged training set affects generalization ability for this task.

3.3 Generate positive and negative samples

For creating positive samples, punctuation is removed (except abbreviations such as it's, Mr., I've, etc.) and words are converted to lower case.

For creating negative samples, we mimic typical errors of the speak recognition system, which are detailed in the following, and we propose corresponding methods to create negative samples with respect to typical errors.

False sentence boundary: When a speech recognition system fails to correctly separate two sentences, the first sentence would be cut off in the middle and part of the sentence would be assigned to the next sentence (illustrated in Fig. 1 (a)). For such negative samples, we group the sentences by three, and randomly separate the three sentences into 2-4 sentences (so that on average negative samples created in this way would have equal length with positive samples). While choosing random separating points, the genuine sentence separations points, punctuation and typical words for starting subsentences (e.g. that, which, because, etc.) are avoided, and thus reduce the probability that a generated sample is still a complete sentence by chance (e.g. 'I like you because you are beautiful' to 'I like you'.)

Missing words: A speech recognition system can fail to recognize one or several words from a sentence, and as a result some words may be missing in the produced transcripts (Fig. 1 (b)). For such negative samples, we randomly remove 1 word for sentences up to 3 words, and 2-4 words from longer sentences.

Repeating words: The system can record speak-

ers' unintended repeated words (Fig. 1 (c)). For such negative samples, we randomly repeat 1 word for sentences within 3 words, and 1-3 words from longer sentences.

False word recognition: The system can mistakenly recognize one word as another word (Fig. 1 (d)). For such negative samples, we randomly replace 1 word for sentences within 3 words, and 1-3 words from a longer sentences, by random words from another sentence.

Finally, the punctuation is removed and words are converted to lower case.

Figure 1: Typical errors in speech recognition system

After creating the positive and negative samples, the sentences longer than 100 words are removed, for they are too long to appear in speech recognition. We create the same number of negative samples as that of positive samples, so that we have a balanced dataset. The ratio between different types of negative samples is 2:1:1:1. The type *False Sentence Boundary* corresponds to two times the number of other negative sample types since *False Sentence Boundary* contains two types of false sentences, those which are cut off and those which are assigned with extra words.

4 Experiments and Discussion

In this section, we report the results of our experiments. We describe below the setup, and then evaluate the different models in Sec. 4.1. In Sec. 4.2, based on the models, we train a Random Forest classifier to further aggregate the models and improve the performance. In Sec. 4.3, we compare the performance of BERT trained on Standard and Large Dataset. Finally, we show the result of BERT trained on a Multi-Labeled Dataset in Sec. 4.4.

Training details: We train each model for 5 epochs with batch size 64 using Adam optimizer. The initial learning rate is set as $3e-5$ for fine-tuning transformer models and $1e-3$ for downstream classification networks. To prevent overfitting, we only save the model with optimal performance on test set after each epoch.

4.1 Results on Standard Dataset

As explained in Sec. 2, we train five models on the Standard Dataset containing 1 million proper sentences and 1 million non-proper sentences to evaluate their performances.

The results of this experiment are presented in Table 1.

Model	Test Accuracy
BERT	89.27%
GPT-2	88.67%
BIG-BIRD	**90.26%**
BERT embedding + Bi-LSTM	86.33%
BERT embedding + TextCNN	81.40%

Table 1: Test accuracy of five models on Standard Dataset

From the results, we can see that the transformers provide much better results than the models sequentially linking BERT embedding and either a BiLSTM or TextCNN. Specifically, BIG-BIRD provides the optimal performance, with 90.26% test accuracy. BERT and GPT2 provide similar test accuracy, 89.27% and 88.67% respectively.

4.2 Ensemble learning with Random Forest

In this section, we combine the five trained models (in Table 1) with random forest in order to produce one optimal predictive model. The idea of the ensemble learning is to train a random forest classifier with the combination of the predicted classes from the models. The random forest classifier can generate a final classification through a majority vote mechanism.

To prevent random forest from overfitting the train set, we use a separate ablation set, instead of the train set which the models are trained on. The best parameters after 10-fold cross-validation are 100 decision trees, and a maximum depth of 3. The test accuracy of the random forest reaches 90.51%, higher than the optimal accuracy among the individual models (90.26%), but not to a large extent. This is probably since the transformers (along with their embedding) share similar structures and do not diverge much on decisions.

4.3 Results on Large Dataset

In this section, we train BERT on the large dataset (5 times the size of the Standard Dataset) with less epochs (1 epoch in contrast to 5 epochs). Overall, the model is trained with the same iterations as with Standard Dataset. With the same training details described before (but only for one epoch), results show that training with Large Dataset provides a

higher test accuracy (90.36%), compared with the accuracy trained with Standard Dataset (89.27%).

The results suggest that, provided with enough computational capacity, we can further improve our model's generalization ability by training on a larger dataset.

4.4 Result on multi-label dataset

In this section, we further create a Multi-Label Dataset, which contains the same samples as the Standard Dataset, whereas the negative samples are distinctively labeled (including *false sentence boundary*, *false word recognition*, *missing words*, and *repeating words*) instead of uniformly labeled as *negative*.

We train a BERT model on this dataset, and it reached 85.01% classification test accuracy. The precision, recall and F1-score of each class is given in Table 2.

Sample Class	Precision	Recall	F1 Score	Support
Complete Sentence	0.87	0.94	0.90	109857
False Sentence Boundary	0.83	0.81	0.82	42677
False Word Recognition	0.84	0.70	0.77	21897
Missing Words	0.64	0.50	0.56	21711
Repeating Words	0.96	0.99	0.98	21781

Table 2: Precision, Recall and F1-Score of each sample class

From the result, we can see that the simplest task is to identify repeated words in the sentences (F1-score near 0.98). Identifying complete sentences is also a relatively easy task, with a F1-score of 0.90. The hardest task for the model is detecting whether there are missing words in the sentence. It achieves only 64% precision and 50% recall on this task.

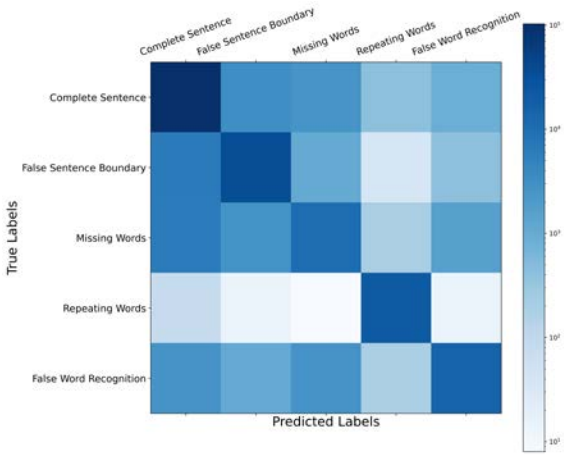


Figure 2: Confusion matrix for BERT trained on Multi-Label Dataset

The confusion matrix is drawn in Fig. 2. From this figure, we can further see that the classifier finds it difficult to classify between complete sentences and sentences with missing words, even though in most of the cases more than one word is missing in the erroneous sentences. This is understandable because in most cases, not every word is indispensable, even we lose some words, and maybe the meaning is not exactly the same but the sentence still makes sense grammatically.

4.5 Result on real-world ASR outputs

Finally we test our trained multi-modal BERT model on the real-world ASR outputs from CEASR corpus (Ulasik et al., 2020). The predictions are shown in Fig. 3, where we can see the model is able to capture real-world ASR errors correctly, while we also provide an example where the model fails.

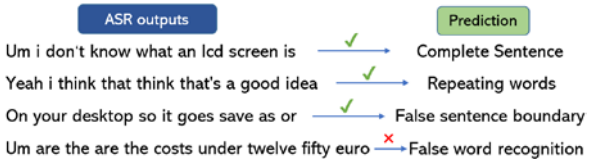


Figure 3: Prediction on real-world ASR outputs

5 Conclusion

In this paper, a dataset for detecting speech recognition errors was created, where four different types of typical speech recognition errors were taken into account. Experimental results show that transformer models are capable of providing good performance on classification of the constructed dataset for speech recognition error, reporting approximately 90% accuracy for BERT, GPT2 and BIG-BIRD. A Random Forest was trained based on the five models, and further improved the test accuracy to over 90.51%. Overall, the results suggest that using state-of-art transformer models can provide good quality for detecting the errors in speech recognition systems, and provide feedback on further improvements of speech recognition systems. In our future works, special adjustments might be needed to better cope with identifying missing words in recognized sentences.

References

Nancy Agarwal, Mudasir Ahmad Wani, and Patrick Bours. 2020. Lex-pos feature-based grammar error detection system for the English language. *Electronics*, 9(10):1686.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep

bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Wikimedia Foundation. Wikimedia downloads.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Zhenhui He. 2021. English grammar error detection using recurrent neural networks. *Scientific Programming*, 2021.

Yang Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1526–1540.

Yang Liu, Andreas Stolcke, Elizabeth Shriberg, and Mary Harper. 2005. Using conditional random fields for sentence boundary detection in speech. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 451–458.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Andrew Thompson. 2017. All the news: 143,000 articles from 15 American publications. =https://www.kaggle.com/snapcrack/all-the-news.

Don Tuggener and Ahmad Aghaebrahimian. 2021. The Sentence End and Punctuation Prediction in NLG text (SEPP-NLG) shared task 2021. In *Swiss Text Analytics Conference–SwissText 2021, Online, 14-16 June 2021*. CEUR Workshop Proceedings.

Malgorzata Anna Ulasik, Manuela Hürlimann, Fabian Germann, Esin Gedik, Fernando Benites de Azevedo e Souza, and Mark Cieliebak. 2020. Ceasr: a corpus for evaluating automatic speech recognition. In *12th Language Resources and Evaluation Conference (LREC) 2020*, pages 6477–6485. European Language Resources Association.

Dong Yu and Li Deng. 2016. *Automatic speech recognition*, volume 1. Springer.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big Bird: Transformers for Longer Sequences. In *NeurIPS*.

SDS-200: A Swiss German Speech to Standard German Text Corpus

Michel Plüss[a], Manuela Hürlimann[b], Marc Cuny[c], Alla Stöckli[c], Nikolaos Kapotis[b], Julia Hartmann[a], Malgorzata Anna Ulasik[b], Christian Scheller[a], Yanick Schraner[a], Amit Jain, Jan Deriu[b], Mark Cieliebak[bc], Manfred Vogel[a]

[a]University of Applied Sciences and Arts Northwestern Switzerland, Windisch
[b]Zurich University of Applied Sciences, Winterthur
[c]SpinningBytes AG, Winterthur
michel.pluess@fhnw.ch

Abstract

We present SDS-200, a corpus of Swiss German dialectal speech with Standard German text translations, annotated with dialect, age, and gender information of the speakers. The dataset allows for training speech translation, dialect recognition, and speech synthesis systems, among others. The data was collected using a web recording tool that is open to the public. Each participant was given a text in Standard German and asked to translate it to their Swiss German dialect before recording it. To increase the corpus quality, recordings were validated by other participants. The data consists of 200 hours of speech by around 4000 different speakers and covers a large part of the Swiss German dialect landscape. We release SDS-200 alongside a baseline speech translation model, which achieves a word error rate (WER) of 30.3 and a BLEU score of 53.1 on the SDS-200 test set. Furthermore, we use SDS-200 to fine-tune a pre-trained XLS-R model, achieving 21.6 WER and 64.0 BLEU.

Keywords: Corpus, Less-Resourced/Endangered Languages, Speech Recognition/Understanding, Speech Resource/Database, Statistical and Machine Learning Methods

1. Introduction

We present Schweizer Dialektsammlung (SDS-200), a corpus of Swiss German dialectal speech with the corresponding Standard German text. The data consists of 200 hours of speech. We make the corpus publicly available [1].

Swiss German is a family of German dialects spoken by around five million people in Switzerland. It differs from Standard German regarding phonetics, vocabulary, morphology, and syntax and is primarily a spoken language. While it is also used in writing, particularly in informal text messages, it lacks a standardized orthography. This leads to difficulties for automated text processing due to spelling ambiguities and huge vocabulary size. Therefore, it is often preferable to work with Standard German text, for which automated processing tools exist in abundance. The main challenge is that Swiss German is not a unified language but a collection of dialects, which sometimes differ significantly in phonetics, grammar, and vocabulary. The immense vocabulary makes it hard to create a Swiss German Automatic Speech Recognition (ASR) system. Due to these reasons, Swiss German is a low-resource language. One way to tackle Swiss German ASR is an end-to-end Swiss German speech to Standard German text approach. This can be viewed as a speech translation (ST) task with similar source and target languages. Training a model for this task requires a substantial amount of data. Unfortunately, not enough public data is available for Swiss German. The largest available corpus, the Swiss Parliaments Corpus (SPC) (Plüss et al., 2021), is limited to the Bernese dialect. However, there are many different dialects in Switzerland, some of which differ substantially from Bernese because the difference between dialects can be significant, especially regarding vocabulary and pronunciation; as many dialects as possible should be part of the training data.

For SDS-200, we created a web recording tool[2] which is open to the public. The idea is that the public can record Standard German sentences in their Swiss German dialect. Other participants then validate the recordings. Almost 4000 different participants from all over Switzerland helped create a high-quality corpus covering a large part of the Swiss German dialect landscape. To cover a wide range of topics and increase vocabulary diversity, we used texts from Swiss newspapers and the German Common Voice corpus. The code of the tool is open source[3].

The remainder of this paper is structured as follows: Related work is discussed in section 2. The data collection process is described in section 3. Corpus preparation and statistics can be found in section 4. In section 5, we describe a baseline model trained on the corpus. Section 6 wraps up the paper and gives directions for future work.

2. Related Work

End-to-end approaches are widely used in deep learning, especially natural language processing (NLP). In the domain of speech translation, suitable corpora are

[1] https://swissnlp.org/datasets/

[2] https://dialektsammlung.ch/de

[3] https://github.com/stt4sg/dialektsammlung-public

scarce. The MuST-C dataset (Di Gangi et al., 2019) provides 400 h of English speech data with sentence-aligned text for eight different languages (German, French, Spanish, Italian, Dutch, Portuguese, Romanian, and Russian). The MuST-C data is collected from TED talks, providing a variety of topics and speakers (male/female, native/non-native speakers). TED talks are manually transcribed and translated, providing a high-quality data source.

Europarl (Iranzo-Sánchez et al., 2020) is another ST corpus with speech and sentence-aligned text for 6 European languages (English, German, French, Spanish, Italian, and Portuguese) containing between 20 and 89 hours of audio for 30 pairs. The sentence alignment is done automatically. Due to the automatic alignment, audio data with low alignment confidence is discarded, and the data quality is lower than manual text alignment. Europarl contains speeches held in the European Parliament.

Four public datasets contain Swiss German audio with transcripts. SPC (Plüss et al., 2021) is the largest corpus with 293 hours of data in the Bernese dialect recorded in the Bernese cantonal parliament. The text and audio are automatically aligned by using commercial Standard German ASR systems, followed by a forced sentence alignment using the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). The ArchiMob dataset (Scherrer et al., 2019) includes 69 hours of Swiss German speech and Swiss German transcript. There are no Standard German transcripts available. The Radio Rottu Oberwallis dataset (Garner et al., 2014) includes 8 hours of speech, 2 of which are provided with Standard German transcripts. Swiss-Dial (Dogan-Schönberger et al., 2021) is a high-quality dataset including eight different Swiss German dialects with roughly 3 hours of audio data per dialect. The sentences are crawled from newspapers and Wikipedia and then manually translated into the selected eight Swiss German dialects. The translated sentences are then recorded sentence by sentence in a studio setting.

SDS-200 combines the strengths of the existing corpora in Swiss German ASR with a large size of 200 hours, Standard German transcripts, and perfect alignment. What makes it unique is the coverage of a large part of the Swiss German dialect landscape and that almost 4000 different speakers made the recordings. We now describe the components in more detail.

3. Data Collection

Our data collection tool is based on the Common Voice platform (Ardila et al., 2020). We adapted the annotation guidelines to the special case of Swiss German. We use the two-step annotation process of the original platform consisting of a recording step and a validation step (see Figures 1 and 2). For the recording step, we presented Standard German sentences from Swiss newspapers, covering diverse topics and Switzerland-specific named entities, and texts from the German Common Voice corpus to the participants. They were then asked to translate each sentence into their Swiss German dialect and record it. For the validation step, the participants were presented with a sentence-recording pair and asked if the recording contained an accurate Swiss German translation of the Standard German sentence.

The goal was to create a corpus with as many hours and as much dialect and topic diversity as possible. We worked extensively with the Swiss media to reach as many people as possible. To enhance the engagement, we organized two contests on our platform. The leaderboard contest awarded prices to the participants with the most recordings, factoring in the quality of their translations. The *Clash of Cantons* contest was a competition between the 26 Swiss cantons.

3.1. Sentence Selection

The sentences used for the recordings were derived from Swiss newspapers and the German dataset of Common Voice. We used newspaper articles from all categories from the past five years. As the speakers' task consisted of translating the sentences from Standard German to Swiss German, not just reading them, we expected the speakers' cognitive effort to be larger, hence the error probability to be higher. Keeping this in mind, we carefully selected sentences to ensure lexical diversity and reduce sentence complexity. To this end, we selected only sentences between 5 and 12 tokens long. We applied the following filtering criteria:

- Exclude sentences containing tokens that occur less than 1000 times per billion words. We use the Exquisite Corpus[4] to compute the word frequencies.
- Exclude sentences with a large number of rare words having an average word frequency below 10'000 per billion words.
- We removed sentences with dates and numbers with more than three digits. This is to reduce inconsistencies in how speakers read or translate the prompts.
- Sentences containing citations, e-mail addresses, hashtags, and phrases in brackets are also removed.
- We kept only complete sentences. We used simple heuristics to remove incomplete sentences. For instance, each sentence begins with an uppercase letter or a digit, and a sentence should contain at least one noun, pronoun, or proper noun and one verb.

The final set of prompts contains 1'267'195 sentences. Our tool samples newspaper sentences in 80% of cases, and in 20% of cases, it samples from the German Common Voice pool.

[4] https://github.com/LuminosoInsight/exquisite-corpus

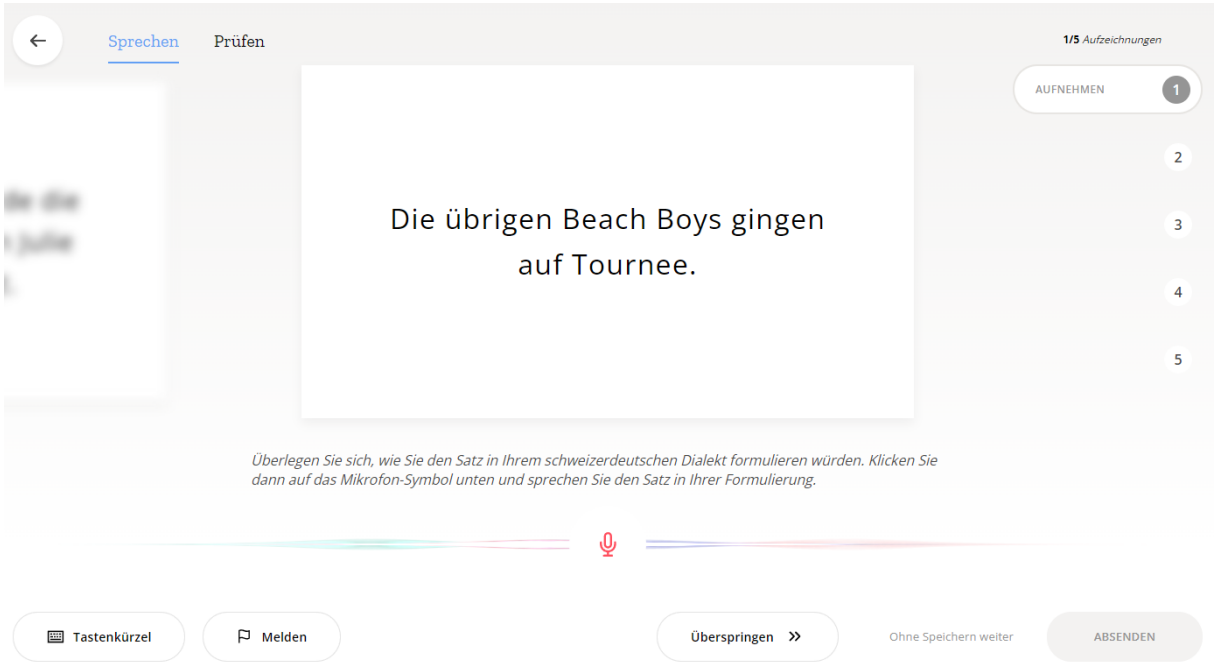


Figure 1: Recording step in our tool. "*Die übrigen Beach Boys gingen auf Tournee.*" is the sentence to be recorded.

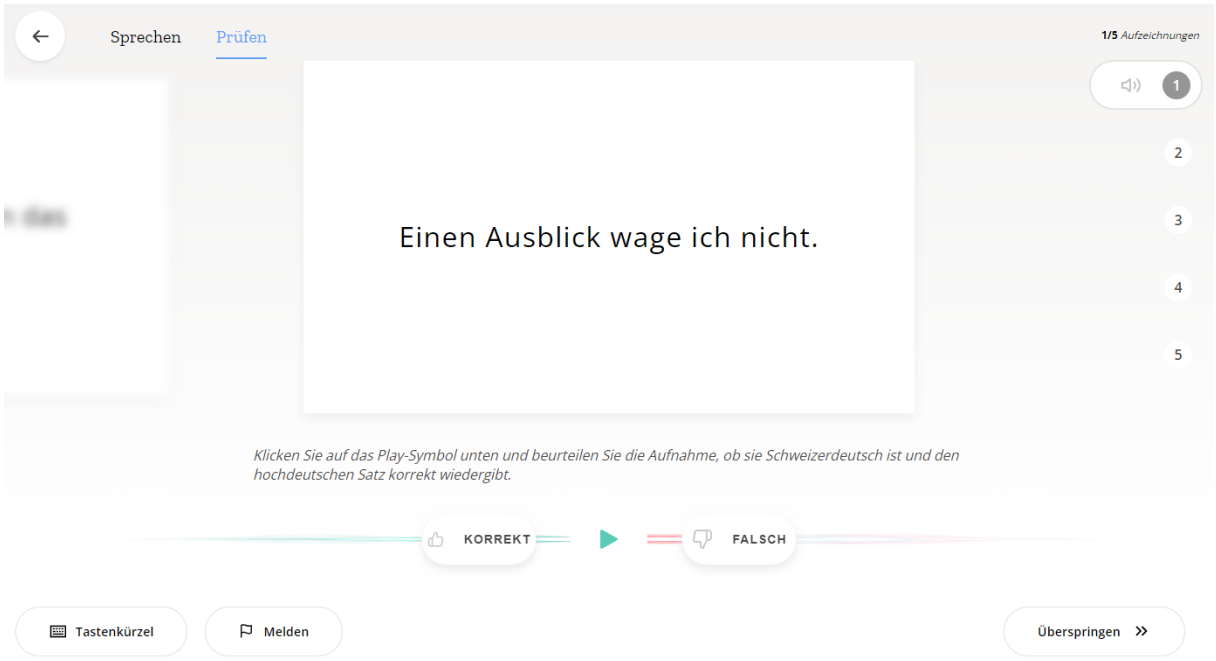


Figure 2: Validation step in our tool. "*Einen Ausblick wage ich nicht.*" is the Standard German sentence. The recording must be played and then judged as correct ("*Korrekt*") or wrong / inaccurate ("*Falsch*").

3.2. Recording Tool

We made two adaptions to the original Common Voice (Ardila et al., 2020) platform. First, we added the possibility for the participants to specify the zip code of origin of their dialect[5]. This allows us to investigate dialects in different granularity levels: coarse dialect regions, cantons, fine-grained dialect regions, and even individual municipalities. Additional demographic information such as age and gender selection is already

[5]The origin of a participant's dialect could for example be the place where he or she grew up and / or went to school. The specified zip code is not to be confused with the current place of residence, which would not allow reliable inference of a participant's dialect.

Split	Hours	Sentences	Speakers
train (raw)	188.9	144′468	3428
train (filtered)	178.3	135′271	3247
validation	5.2	3638	288
test	5.4	3636	281

Table 1: Data splits of the Dialektsammlung corpus.

available in Common Voice. Second, we adapt the annotation guidelines to cover the special case of Swiss German. The annotation is performed in two steps: a recording step and a validation step.

Step 1: Recording. During the recording step, depicted in Figure 1, the participant is shown a Standard German sentence and asked to translate it to Swiss German speech. Sentences are recorded in packages of 5 and can be skipped or reported if necessary. One crucial point for our Swiss German speech to Standard German text use case is the inherent translation step the participant has to do before recording. As an example, the participant is presented with the following Standard German sentence: "*Robben verstand dies wie viele andere Spieler nicht.*". The participant should then think about how he or she would formulate this sentence in his or her Swiss German dialect, e.g. "*De Robben het das wie vieli anderi Spieler nid verstande.*", before actually recording the Swiss German version. This can include vocabulary as well as grammar changes, such as changing the past tense from Standard German "*verstand*" to Swiss German "*het (...) verstande*", which is necessary because the imperfect tense does not exist in Swiss German, where the perfect tense is used instead. We display an explanation popup with examples before the first recording to make this clear to participants. We also display a short explanation below the sentence to be recorded (see Figure 1).

Step 2: Validation. Figure 2 depicts the validation function. Participants are asked to listen to other recordings and judge whether the recording contains an accurate Swiss German translation of the Standard German sentence. Recordings are again validated in packages of 5 and can be reported or skipped if necessary. Similar to the recording function, we display a detailed explanation with examples of wrong (e.g. recording is in Standard German rather than Swiss German) or inaccurate (e.g. wrong tense) translations when a participant visits the validation page for the first time.

3.3. Collection Process

To reach as many people as possible, we collaborated with a range of national and local newspapers, television networks, and radio stations. In addition, four well-known Swiss comedians agreed to record a short video supporting the project and share it on their social media accounts, some of them reaching more than 100'000 followers.

To keep the participants motivated, we organized two contests, the leaderboard contest and the *Clash of Cantons*.

Leaderboard. The leaderboard contest was a competition between all registered participants. For each participant, we computed a score based on the number of recordings, the number of validations given, and the number of positive validations received. The top ten of the leaderboard were awarded attractive Switzerland-themed prizes. Furthermore, the participant with the highest recording quality (lowest rejection rate) was awarded a special prize.

Clash of Cantons. The *Clash of Cantons* was a competition between the 26 Swiss cantons. The idea was to spark a competition between the cantons and for participants to "fight" for their respective canton. The winning canton was picked according to its number of recordings, weighted by their average quality, normalized by the population of the canton.

The data of the corpus described here was collected over seven months, with 58 % of recordings made during the 38 days where the two contests were held. The current version contains 200 hours of raw speech data in MP3 format with a sampling rate of 32 kHz.

4. Corpus Preparation and Data Statistics

4.1. Data filtering

Crowd-sourced data needs filtering to ensure high data quality. We used the public validation process to filter bad samples such as empty, truncated, or silent recordings and wrong translations.

Of all recorded data, 33% have been validated, and of these samples, 88% have been accepted. To also use a large amount of unvalidated samples, we allow unvalidated samples as well under the following conditions:

- The speaker *has some* validated recordings and more than 80% of the validated clips are accepted.
- The speaker *has no* validated recordings and the duration is within 2 to 12 seconds.

We found that we were able to filter out many clips with recording problems (e.g., empty recordings) with the second rule. Since the added unvalidated data likely contains some invalid samples, they will need to be filtered further as more clips are validated. We also provide the unfiltered train data so that corpus users can compile their own filter rules.

4.2. Corpus Structure

We provide randomly generated train, validation, and test splits, ensuring that each speaker is part of only one split. The target size of the validation and test splits is 5.3 hours each. Table 1 shows the number of hours, sentences, and speakers of each split. To ensure optimal quality, validation and test splits only contain validated samples. Furthermore, to obtain balanced sets

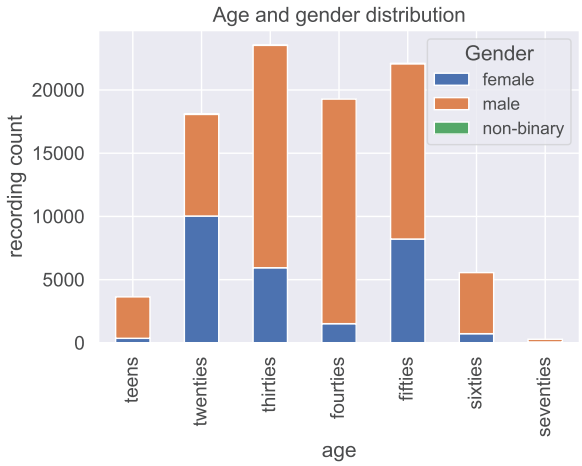


Figure 3: Number of utterances per speaker's age group and gender.

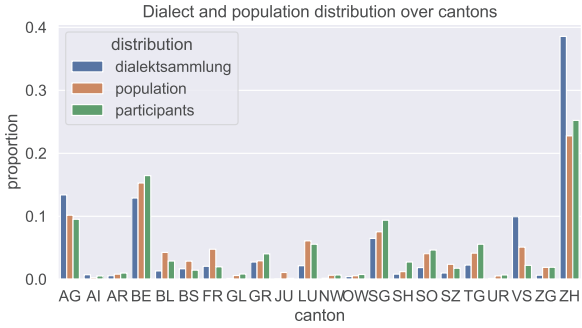


Figure 4: Canton distribution in the dataset compared with the relative population and the relative number of unique speakers per respective canton. Only cantons where Swiss German is spoken are shown.

and a larger variety of speakers, we only allow speakers with 5 to 200 recorded sentences to be part of either validation or test splits.

4.3. Data Statistics

On average, an utterance is 4.8 seconds long with a standard deviation of 1.3 seconds. The shortest and longest utterances are 2 and 11.2 seconds long, respectively. In Figure 5 we display the utterance length distribution.

By crowdsourcing the data, we obtain a diverse set of speakers regarding age, gender, and dialect. In total, the filtered SDS-200 contains 142'545 utterances with 138'553 unique sentences. The vocabulary consists of 41'289 German words. Out of 3816 speakers, 8% are male, 6% are female, 86% did not reveal their gender, and 4 participants are non-binary. In terms of utterances, 19% of utterances are voiced by females, 46% by males, and 35% of unknown gender. On average, each participant recorded 37 utterances with a standard deviation of 364 utterances. The participant with the most speech donations recorded 13'333 utterances. In

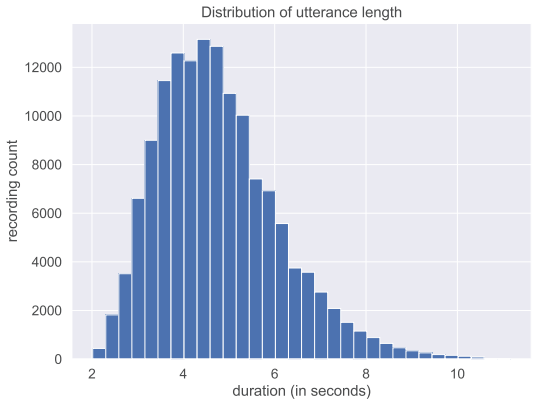


Figure 5: Distribution of utterance lengths in the SDS-200.

Figure 3 we display the age and gender distribution over the recorded utterances. In Figure 4 we show the distribution over the number of recordings for each canton and compare them with the population of the respective cantons[6] and the proportion of unique speakers. The collected dialects follow the dialect distribution in Switzerland closely, with some exceptions. For Appenzell Innerrhoden, we have four times more utterances than the relative population. Wallis and Zürich have almost twice as many utterances. In the canton Wallis, one speaker recorded 10'368 out of 11'739 samples. The cantons Baselland, Glarus, Jura, Luzern, Nidwalden, Uri, and Zug are underrepresented in the SDS-200.

5. Baseline

We conducted experiments to demonstrate the use of the SDS-200 corpus for speech translation. We further evaluated how the corpus can be combined with the SPC (Plüss et al., 2021). Finally, we assessed how large-scale pre-training on unlabeled speech data can improve the performance by finetuning XLS-R Wav2vec models (Babu et al., 2021) on the SDS-200 train set.

Transformer Baseline. We employed Transformer (Vaswani et al., 2017) based models implemented in the FAIRSEQ S2T library (Ott et al., 2019; Wang et al., 2020) as our baselines. These models consist of a two-layer convolutional subsampler followed by a Transformer network with 12 encoder layers and six decoder layers. For the Transformer network, we employed eight attention heads, an embedding dimension size of 512, and a dropout rate of 0.15. We used the default model hyper-parameters and learning rate schedules provided by the library without any task-specific tuning. We evaluated the model performance when training on SDS-200 alone as well as the combination of SDS-200 and the SPC. After training, we

[6]We use the canton information as an indicator for the dialect.

Model	Train data	Model parameters	WER		BLEU	
			valid	test	valid	test
Transformer	SDS-200	72M	31.3	30.3	52.1	53.1
Transformer	SDS-200+SPC	72M	24.9	24.7	60.9	61.0
XLS-R (0.3B)	SDS-200	317M	27.2	26.9	54.9	54.6
XLS-R (1B)	SDS-200	965M	21.7	21.6	63.9	64.0

Table 2: Performance of the Transformer Baseline and XLS-R Wav2Vec models finetuned on the SDS-200 train set. We report Word Error Rate (WER) and BLEU scores obtained from evaluating on the SDS-200 valid and test splits.

averaged the weights of the ten checkpoints with the lowest validation loss to obtain the final model.

XLS-R fine-tuning. For the Wav2vec experiments, we employed XLS-R models (Babu et al., 2021) that were pre-trained on 436K hours of unlabeled speech data covering more than 128 languages and are publicly available[7]. Importantly, Swiss German was not part of the training data. Of the available pre-trained models, we evaluated XLS-R (0.3B) and XLS-R (1B), whereas the number in braces denotes the number of model parameters. XLS-R Wav2vec models consist of a convolutional feature encoder, followed by a stack of transformer blocks. Details of the architecture configurations can be found in (Babu et al., 2021). For the fine-tuning on the SDS-200 corpus, we followed the procedure and hyper-parameters described by the authors.

Results. The results of our experiments are shown in Table 2. Both additional labeled training data and large-scale self-supervised pre-training on unlabeled speech data lead to performance improvements. The strong performance of XLS-R (0.3B) highlights the benefits of latter in low-resource settings, even if the target language was not available during pre-training. Notably, for all our experiments, we did not use any external language model.

6. Conclusion

In this work, we presented SDS-200, a speech translation dataset for Swiss German speech to Standard German text. The main characteristics of this corpus are the large variety of Swiss German dialects that are covered and the large number of speakers that contributed to the data collection. The baseline achieved 30.3 WER score, and 53.1 BLEU score on the SDS-200 test set. The current version contains around 200 hours of speech.

Our goal is to increase the size of the corpus in the future, which will allow for even better performance. We plan to find new ways to engage the public, for instance, by adding gamification components to keep the engagement high. The current version is publicly available.

[7] https://github.com/pytorch/fairseq/tree/main/examples/wav2vec/xlsr

7. Acknowledgements

First and foremost, we would like to thank all participants for their contribution to this corpus. Furthermore, we thank Tamedia for providing the newspaper texts. We also thank Claudio Zuccolini, Frölein da Capo, Mike Müller, Renato Kaiser, Simon Enzler, and Sina for their public relations efforts.

8. Bibliographical References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2021). Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv*, abs/2111.09296.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, C., Tang, Y., Ma, X., Wu, A., Okhonko, D., and Pino, J. (2020). fairseq S2T: Fast Speech-to-Text Modeling with fairseq. In *Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*.

9. Language Resource References

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., and Turchi, M. (2019). MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the*

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2012–2017, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dogan-Schönberger, P., Mäder, J., and Hofmann, T. (2021). SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German.

Garner, P. N., Imseng, D., and Meyer, T. (2014). Automatic Speech Recognition and Translation of a Swiss German Dialect: Walliserdeutsch. In *Proceedings of Interspeech*, Singapore, September.

Iranzo-Sánchez, J., Silvestre-Cerdà, J. A., Jorge, J., Roselló, N., Giménez, A., Sanchis, A., Civera, J., and Juan, A. (2020). Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Plüss, M., Neukom, L., Scheller, C., and Vogel, M. (2021). Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German Text Corpus. In *Swiss Text Analytics Conference 2021*, Proceedings of the Swiss Text Analytics Conference 2021.

Scherrer, Y., Samardžić, T., and Glaser, E. (2019). ArchiMob: Ein multidialektales Korpus schweizerdeutscher Spontansprache. *Linguistik Online*, 98(5):425–454, November.

SwissText 2022
Swiss Text Analytics Conference 2022

Proceedings of the Swiss Text Analytics Conference 2022

Lugano, Switzerland, June 8-10, 2022.

Edited by

Manuela Hürlimann *
Roberto Mastropietro **
Daniele Puccinelli **
Fabio Rinaldi **
Mark Cieliebak *

* Zurich University of Applied Sciences (ZHAW), Centre for Artificial Intelligence, 8401 Winterthur, Switzerland
** University of Applied Sciences and Arts of Southern Switzerland (SUPSI), Department of Innovative Technologies, 6962 Lugano-Viganello, Switzerland

Table of Contents

Welcome to the 7th Swiss Text Analytics Conference (SwissText 2022)

The SwissText conference series was started by the Zurich University of Applied Sciences (ZHAW) in 2016. The first edition, which attracted more than 170 people, was already a big success. The conference was turned over to the Swiss Association for Natural Language Processing (SwissNLP) in 2020, and it is now organized by Swiss universities in cooperation with SwissNLP.

This year, SwissText was hosted at the University of Applied Science and Arts of Southern Switzerland (SUPSI), and co-organized by the ZHAW.
It was the first Swisstext on premise edition after two years of pandemic that forced the organizers to hold the conferences online. It was great to see attendees talking face to face again and enjoying the coffee breaks and the networking opportunities outdoors.

The conference hosted four pre-conference workshops:

- 2nd Swiss German Speech to Standard German Text shared task
- Keyword extraction in scientific documents
- GSWNorm, Shared Task & Workshop on the normalization of written Swiss German
- NLP and Insurance

We received 23 submissions for the applied track, 6 for the junior track and 4 for the demo track. You will find the full junior papers and the abstracts of the applied talks and posters in separate sections in these proceedings. The workshops papers and abstracts are included as well.

A great addition to the conference program was the second edition of the “Battle of NLP ideas”, where groups of researchers presented outstanding ideas for potential new scientific projects.

Beside the presentations hosted in conference rooms, a foyer was dedicated to 13 sponsors and a few academic partners booths. In the same room the poster and demo session took place.

We would like to thank our keynote speakers Marco Passarotti (Università Cattolica, Milano), Raul Rodriguez-Esteban (Roche) and Enrique Alfonseca (Google)
Their perspectives and contributions were much appreciated. A big thank you also to our workshop organizers and to all members of our programme committees for their excellent work.

We are grateful to our sponsors and partners, who supported us in spite of the fact that the economy was still depressed due to the pandemic. In particular, we would like to acknowledge the generous support by Innosuisse (the Swiss Innovation Agency) and our co-organizer, the Data Innovation Alliance.

Last but foremost, we would like to thank Manuela Hürlimann, Daniele Puccinelli and Fabio Rinaldi, who were the main organizers. They were assisted by Oscar, Mariangela, Erica, Antonio and many more for various tasks. Without them this conference would not have been possible. Additionally, we are thankful for the support of the programme committees for the Applied and Junior tracks.

It was a great pleasure for us to organize and chair this conference. We hope that all participants enjoyed the conference and the location.

Mark Cieliebak and Roberto Mastropietro
Conference Chairs

Speech-to-Text Technology for Hard-of-Hearing People

by Manuela Hürlimann (Centre for Artificial Intelligence, Zurich University of Applied Sciences), Jolanda Galbier (Pro Audito Schweiz) and Mark Cieliebak (Centre for Artificial Intelligence, Zurich University of Applied Sciences)

Hard-of-hearing people face challenges in daily interactions that involve spoken language, such as meetings or doctor's visits. Automatic speech recognition technology can support them by providing a written transcript of the conversation. Pro Audito Schweiz, the Swiss federation of hard-of-hearing people, and the Centre for Artificial Intelligence (CAI) at the Zurich University of Applied Sciences (ZHAW) conducted a preliminary study into the use of Speech-to-Text (STT) for this target group. Our survey among the members of Pro Audito found that there is large interest in using automated solutions for better understanding in everyday situations. We now propose to take the next step and develop an application which uses ZHAW's high-quality STT models.

Figure 1: A group discussion – this is a situation in which the proposed application could support hard-of-hearing people (Photo: colourbox.de).

The average person holds more than 25 conversations per day, which can be very challenging for people with hearing loss, as their auditory perception of spoken language is limited. Pro Audito provides an interpreting service ("Schriftdolmetschen"), where a trained human interpreter accompanies the hard-of-hearing person and creates a written transcript of the interaction on the fly. While this is highly appreciated with 1,800 hours of speech transcribed each year, the financial compensation by the Swiss disability insurance is currently limited to professional and educational settings and the cost is capped [L1]. We received an Innovation Cheque from Innosuisse to run a preliminary study consisting of a needs analysis and market research. Our goal was to find out how STT could be used to create an offer for people with hearing loss that provides more flexibility and independence.

Needs analysis

The needs analysis was conducted via a detailed survey among the members of Pro Audito, which was answered by 166 respondents, of which 87% have moderate or severe hearing loss. We found that 28% already use technical support to facilitate understanding, which consists mostly of external microphones, headsets or rerouting sound to their hearing aid via Bluetooth (e.g., when watching TV). Some people already use STT apps, where the most frequently named use cases are appointments at the doctor or optometrist, meetings (both online and on-site, see Figure 1) and conversations in crowded spaces with background noise (such as restaurants). 57% of our respondents can imagine using STT technology to facilitate their understanding – the most frequently named languages are Standard German, Swiss German, French and English. They were also asked what would be important features of an STT application: it should be as easy as possible to use and provide high-quality recognition (e.g., accuracy, robustness to noise, specialised vocabulary) with minimum latency. Many of our respondents would be willing to pay for a STT solution, either as a one-off purchase or on a monthly subscription basis. Most people would be willing to pay between 50 and 150 CHF one-off or 10 CHF per month.

Market Research

We reviewed existing STT solutions for people with hearing loss and found that currently no single solution ticks all the boxes – some have good recognition accuracy but a poor user interface, others are very easy to use but quickly become unstable when tested in real-life conditions. We are currently developing STT models for various languages at ZHAW. We believe that the best way forward is to develop a dedicated application for hard-of-hearing people and integrate our models for the following reasons:

- Latency: For real-time STT, latency needs to be minimised as much as possible. This means that ideally the model runs on-device, since using external cloud providers introduces an additional time-lag. Creating STT models which are small enough to run on a device such as a smartphone yet have high prediction accuracy is an important challenge.
- Privacy: Users will in some cases want to transcribe sensitive information, such as a conversation with a doctor. With a local model, privacy can be guaranteed.
- Customisation: The use cases from our survey offer significant challenges such as a large number of speakers, spontaneous speech, and background noise. If we use our own STT models, we have full control over their customisation.

Furthermore, it is important that this application can run on an inexpensive device to be accessible to as many users as possible; this is a further argument in favour of a smartphone app.

Future Activities

We propose to develop an application for hard-of-hearing people based on our STT models, which will use a high-precision microphone to record audio – either from the hearing aid itself, a partner microphone, or a wireless lapel microphone. The audio is then transmitted via Bluetooth to the user's smartphone. For minimum latency as well as maximum privacy and customisation, the transcription will be carried out on-device and will be displayed in an easy-to-use interface.

Pro Audito and ZHAW are now looking for partners interested in jointly developing and operating this application - if you are interested, please refer to the contact information below.

Link: [L1] https://kwz.me/hjf

Please contact:
Mark Cieliebak, ZHAW School of Engineering, Switzerland
ciel@zhaw.ch

On the Effectiveness of Automated Metrics for Text Generation Systems

Pius von Däniken and **Jan Deriu** and **Don Tuggener** and **Mark Cieliebak**
Centre for Artificial Intelligence
ZHAW School of Engineering
{vode,deri,tuge,ciel}@zhaw.ch

Abstract

A major challenge in the field of Text Generation is evaluation, because we lack a sound theory that can be leveraged to extract guidelines for evaluation campaigns. In this work, we propose a first step towards such a theory that incorporates different sources of uncertainty, such as imperfect automated metrics and insufficiently sized test sets. The theory has practical applications, such as determining the number of samples needed to reliably distinguish the performance of a set of Text Generation systems in a given setting. We showcase the application of the theory on the WMT 21 and Spot-The-Bot evaluation data and outline how it can be leveraged to improve the evaluation protocol regarding the reliability, robustness, and significance of the evaluation outcome.

1 Introduction

The field of Text Generation is a subfield of Natural Language Processing (Celikyilmaz et al., 2020). We define text generation tasks as those where many different texts may constitute an optimal solution to a given problem. Examples are automated summarization, machine translation, dialogue systems, paraphrasing, caption generation, or natural language generation.

One unsolved issue in the field of Text Generation is the evaluation, be it human or automated evaluation. Human evaluation is more reliable but more cost and time intensive, and automated evaluation is erroneous but performed in a fraction of time and cost (Amidei et al., 2019; Hashimoto et al., 2019; Celikyilmaz et al., 2020; Deriu et al., 2021). One of the main issues is the lack of theoretically founded guidelines when running an evaluation. For instance, how many samples are needed to be able to significantly distinguish the performance of two systems? Or how do we handle the errors made by automated metrics? Under which circumstances is it still possible to run an evaluation campaign that yields significant results? In this work, we make a first step towards developing such a theoretical foundation, which can be used as a guideline to answer the above questions. For this, we consider what we call *binary metrics*. These are metrics that classify the output of a text generation system as being either adequate or inadequate. This allows us to measure the performance of a text generation system as the ratio of adequate responses it generates. Furthermore, it allows us to reason about the performance of the metric in terms of true positives and true negatives.

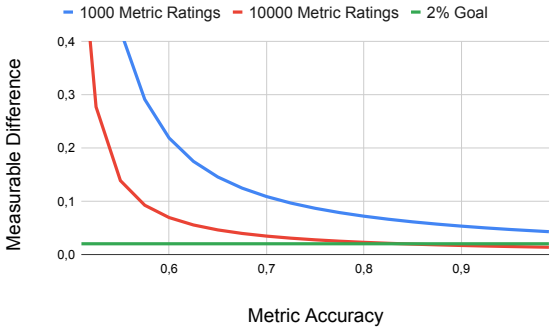


Figure 1: Measurable difference of the performance of two text generation systems depending on the accuracy of a binary metric. We add the 2% line as discussed in the text.

		#Automated Ratings				
		0	$1k$	$5k$	$10k$	$50k$
#Human Ratings	0	1.000	0.109	0.049	0.035	0.015
	10	0.379	0.106	0.049	0.034	0.015
	100	0.134	0.085	0.046	0.033	0.015
	$1k$	0.043	0.040	0.032	0.027	0.015
	$2.5k$	0.027	0.026	0.024	**0.020**	0.013
	$5k$	**0.019**	0.019	0.018	0.017	0.012

Table 1: Mixed Case: Measurable difference for a metric with accuracy of 70% depending on the number of human rating mixed with the number of automated ratings. The values discussed in the text are highlighted in bold.

For this setting, we derive various theoretically

founded guarantees and guidelines that can be used to run an evaluation campaign. For instance, consider Figure 1 (derived by our theory). If we assume a binary metric that has an accuracy of 70%, and if we have access to 1000 automatically rated samples (blue line), then we can reliably distinguish between two text generation systems that have a difference in performance of 10 percentage points. To distinguish two systems with a smaller difference, for instance of 2%, we would need a better metric and many more samples. That is, we need for instance a metric with an accuracy of at least 85% and 10000 automatically rated samples by this metric.

Our theory provides analogous assessments of how many human evaluations are required to reliably distinguish text generation systems. When we say that the performance of two systems can be reliably distinguished, we mean that the difference in their performance is statistically significant. Similarly, a measurable difference in performance is one that leads to statistical significance given the experiment parameters.

In addition, our theory allows for the mix of human and automated evaluation. For this, consider Table 1 where we depict the number of human and automatic ratings required by a metric with 70% accuracy. For instance, to distinguish two text generators with 2 percentage points difference, we need either at least 5000 human ratings, or 2500 human ratings mixed with 10'000 automated ratings.

Our theoretical framework allows us to design our evaluation with theoretical guarantees regarding the significance of the resulting measurements. Given a monetary budget and our theory, one can decide whether to invest in more human annotations, in developing better automated metrics, or in sampling more automated ratings. Our approach can also be used to showcase the limits of a given setting: for instance in Figure 1, we see that using only 1000 automated ratings leads to a minimally measurable difference of 4% even with a perfect metric.

In the remainder of the paper, we derive the theoretical framework for binary metrics and apply it to two showcases: the WMT-21 shared task (Freitag et al., 2021b) and the Spot-The-Bot evaluation (Deriu et al., 2020). We analyse how well these evaluations adhere to the constraints imposed by our theory and demonstrate how the quality of the evaluations can be improved. To serve the community, we will release the formulas as code and as a web interface [1] that allows practitioners to enter their evaluation settings and receive an analysis of the measurable differences in their settings.

2 Definitions

In this section, we introduce the basic definitions that we need for the derivations. First, we define the general setting of Text Generation, then we cover binary metrics, and finally we describe text generation systems.

2.1 General Setting

Definition 1 (Text Generation Environment) *A* text generation environment *is composed of a triple* $\langle \mathcal{I}, \mathcal{O}, \Phi \rangle$*, where* $\mathcal{I}$ *denotes the set of inputs,* $\mathcal{O}$ *the output space, and* $\Phi : \mathcal{I} \times \mathcal{O} \rightarrow \{0, 1\}$ *an oracle that assess whether an output is adequate for a given input.*

For instance, for Machine Translation $\mathcal{I}$ denotes all sentences in the source language and $\mathcal{O}$ all sentences in the target language, while for a chatbot $\mathcal{I}$ contains all dialogue contexts and $\mathcal{O}$ all possible responses in a dialogue. Note that $\mathcal{I}$ and $\mathcal{O}$ can be of infinite size. We regard Φ as an oracle that segments the output space for a given input into adequate and inadequate outputs [2].

Definition 2 (Adequate Responses) $\forall i \in \mathcal{I}$*, we call* $\mathcal{R}^i_+ = \{o \in \mathcal{O} | \Phi(i, o) = 1\}$ *the set of* adequate responses *for input* i*, and* $\mathcal{R}^i_- = \{o \in \mathcal{O} | \Phi(i, o) = 0\}$ *the set of* inadequate responses.

2.2 Binary Metric

In this work, we set our focus to binary metrics, i.e., metrics that classify the output of a text generation system as being either adequate or inadequate. The choice of binary metrics allows us to reason about the performance of a text generation (TG) system as the ratio of adequate responses[3].

[1] https://github.com/vodezhaw/binary_metric_tool

[2] In most real-world setting Φ is approximated with human ratings.

[3] This lies in contrast with metrics that simply return a scalar value (e.g, BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), USR (Mehri and Eskenazi, 2020)) that is difficult to interpret. For instance, if BLEU returns a value of 0.34 for one system and 0.32 for the second system, can we really state that the first system is better than the second (Callison-Burch et al., 2006)? We can use these types of metrics to create binary metrics by selecting a threshold that defines the border between adequate and inadequate responses (e.g., all COMET

We first define the notion of a binary metric, then we show what it means for a binary metric to be error-free or error-prone with regards to Φ.

Definition 3 (Binary Metric) *A* binary metric M_b *is a function* $M_b : \mathcal{I} \times \mathcal{O} \to \{0, 1\}$ *which takes a pair of input and output, and returns either 0 or 1. We interpret the return of 1 as claiming that the output is an adequate output for the given input, and 0 claiming that the output is not adequate.*

Next, we define the notion of an error-free metric. That is, how we expect the metric to behave in the optimal case (i.e. its ability to replicate the oracle Φ).

Definition 4 (Error-Free Binary Metric) M_b^* *is an* error-free *binary metric* $\iff \forall i \in \mathcal{I}, o \in \mathcal{O} : (M_b^*(i, o) = 1 \iff o \in \mathcal{R}_+^i)$.

That is, an error-free binary metric always rates an adequate output as 1 and an inadequate output as 0. Since most metrics do not perform perfectly regarding Φ, we formulate the cases where a metric makes mistakes and the calculation of its performance as follows.

Definition 5 ((ρ, η)-optimal binary metric) *Let* $\rho, \eta \in [0, 1]$ *and* M_b *a binary metric. Then* M_b *is a* (ρ, η)-optimal binary metric *if* $Pr[M_b(i, o) = 1 | o \in \mathcal{R}_+^i] = \rho$ *and* $Pr[M_b(i, o) = 0 | o \notin \mathcal{R}_+^i] = \eta$.

That is, we define the performance of a binary metric as its probability to correctly classify an output as being adequate or not. Thus, the error of a binary metric can be assessed similar to the error of a binary classifier, i.e., ρ is equivalent to the true positive ratio and η to the true negative ratio. Note that $\rho = \eta = 1$ defines an error-free binary metric, whereas all other cases are error-prone. In the case where ρ and η have the same value, $\rho = \eta$, this value is the accuracy of $M_b^{\rho,\eta}$. Note that in practise, ρ and η must be estimated from data.

2.3 Text Generation

We define a text generation system as a function that takes an input from the input-space and generates an output.

Definition 6 ((Optimal) Text Generator) *A* Text-Generator (TG) *is a mapping* $\pi : \mathcal{I} \to \mathcal{O}$ *which generates for each input* i *an output* o*. A TG is* optimal $\iff \forall i \in \mathcal{I} : \pi(i) \in R_+^i$

values above 0.78 are regarded as adequate). This introduces errors, which can be measured.

Next, we introduce the notion of an imperfect text-generator. There are many different ways the errors of a TG can be modeled. We model it as its capability of generating adequate responses.

Definition 7 (α-optimal TG) *Let* π *be a TG and* $\alpha \in [0, 1]$*. Then* π *is an* α-optimal *TG if* $Pr[\pi(i) \in \mathcal{R}_+^i] = \alpha$ *for all* $i \in \mathcal{I}$.

That is, the probability of the text generation system to generate an adequate output is denoted as α. The task of a binary metric is to estimate the α value of a TG system, which has a concrete meaning: Assume that we compare two systems, where $\alpha^{\pi_1} = 0.5$, and $\alpha^{\pi_2} = 0.49$, then these numbers have a clear semantic: π_1 outputs an adequate output in 50% of cases and π_2 in 49% of cases. Thus, one system generates adequate outputs more often than the other. We denote the difference in performance as ϵ. In the following, we will use α^π to denote the rate at which a system π generates adequate responses, and π^α to refer to a system which is α-optimal.

3 Theory: Estimating α with Binary Metrics

In this section, we show how binary metrics can be used to estimate the performance α of text generation systems. For the remainder of the text, assume that $\mathcal{T}_\Phi = \{(i_j, o_j, r_j^*) | 1 \leq j \leq n_\phi\}$ is a set of input-output rating triples of size n_ϕ, where i_j are inputs, $o_j = \pi^\alpha(i_j)$ denotes the output generated by an α-optimal TG system for input i_j, and $r_j^* = M_b^*(i_j, o_j)$ denotes the error-free rating of the j^{th} input-output-pair. Analogously, let $\mathcal{T}_M = \{(i_j, o_j, r_j) | 1 \leq j \leq n_M\}$ be a set of input output rating triples of size n_M, where $r_j = M_b^{\rho,\eta}(i_j, o_j)$ denotes the rating of an error-prone (ρ, η)-optimal binary metric.

We consider three different cases: 1) the error-free case, 2) the error-prone metric case, and 3) the mixed case. The error-free case is where we have access to r_j^*. For instance, we can interpret human evaluation as an example of the error-free case. In the error-prone metric case, we have access only to an (ρ, η)-optimal binary metric. Finally, the mixed case is a novel approach that leverages error-free ratings, which are usually costly to obtain, with error-prone ratings, which are cheaper but are needed en-masse for automated metrics with low ρ and η values, as we will see. Usually, in evaluation campaigns, either the first or second setting is applied.

We apply a Bayesian approach to estimate α by treating it as a random variable, which allows us to model various sources of uncertainty stemming from α, ρ and η, which all need to be estimated from data. The full derivations are given in Appendix A.

3.1 Error-Free Case

Here, we start with the most simple case and introduce the formula to estimate α given error-free ratings r_j^*. Given n_ϕ error-free ratings, α is estimated by $\tilde{\alpha} = \frac{n_+}{n_\phi}$, where $n_+ = \sum_{i=1}^{n_\phi} r_j^*$. This formula can be derived via the frequentist approach or the Bayesian. For the Bayesian approach, we assume a uniform prior over α (i.e. $\alpha \sim Beta(1,1)$). The resulting posterior distribution for α given n_+ is:

$$\begin{aligned} &P(\alpha|N_+ = n_+) \propto P(N_+ = n_+|\alpha) * P(\alpha) \\ &\propto Beta(n_+ + 1, n_\phi - n_+ + 1) \end{aligned} \tag{1}$$

and the value of α is estimated using the mode of $Beta(n_+ + 1, n_\phi - n_+ + 1)$, which corresponds to $\frac{n_+}{n_\phi}$.

3.2 Error-Prone Metric Case

In the error-prone metric case, the probability that $r_j = 1$ depends on ρ and η. Hence, if $r_j = 1$, we cannot assume that $r_j^* = 1$ as well, since the binary metric can be error-prone. For the error-prone setting, we consider two cases, one where ρ and η are provided (e.g. from an earlier evaluation campaign), and one where ρ and η must be estimated from data (i.e., from comparison to error-free ratings).

3.2.1 Provided ρ, η

Here, we assume that the exact values of ρ and η are known. The probability that the binary metric returns a positive label is thus given by:

$$P(r_j = 1) = \alpha\rho + (1-\alpha)(1-\eta) \tag{2}$$

From this, we derive the formula to estimate α using the Bayesian formulation.

Theorem 1 (Estimate α with error-prone metric) *Let $m_+ = \sum_{i=1}^{n_M} r_j \sim Binom(P(r_j = 1), n_M)$ be the number of pairs i_j, o_j rated as adequate $M_b^{\rho,\eta}(i_j, o_j) = 1$. Then we estimate α by computing the mode of the following distribution:*

$$\begin{aligned} &P(\alpha|M_+ = m_+, \rho, \eta) \\ &\propto P(M_+ = m_+|\alpha, \rho, \eta)P(\alpha) \end{aligned} \tag{3}$$

If we assume a uniform prior of α, i.e., $P(\alpha) \sim U(0,1)$, this reduces to: $\tilde{\alpha} = \frac{\frac{m_+}{n_M} + \eta - 1}{\rho + \eta - 1}$

Note that the above formulation does not allow for $\rho + \eta = 1$, in which case our estimator would be undefined. In the following we will assume that $\rho + \eta > 1$. This is a relatively safe assumption since in the case where $\rho + \eta < 1$, we can derive a new metric $M_b^{\rho',\eta'}$ by flipping the predictions of $M_b^{\rho,\eta}$: $M_b^{\rho',\eta'}(i,o) = 1 - M_b^{\rho,\eta}(i,o)$. In this case $\rho' + \eta' = (1-\rho) + (1-\eta) = 2 - (\rho + \eta) > 1$.

3.2.2 Estimated ρ, η

Here, we assume that ρ and η must be estimated from data, which introduces uncertainty. In our case, we estimate ρ and η from error-free ratings (i.e., how well the error-prone metric agrees with the error-free ratings). In practise, the error-free assessments stem from human annotations, which are regarded as the ground truth. To weave the estimation of ρ and η into the Bayesian framework, we treat them as random variables. For this, assume that we have access to a dataset $T_{\rho,\eta} = \{(i_j, o_j, r_j^*, r_j) | 1 \leq j \leq M\}$ of both error-free and error-prone ratings for pairs of inputs and outputs. Denote $\mathcal{T}_{\rho,\eta}^+ = \{(i_j, o_j) | r_j^* = 1\}$ as the set of true positive samples, and $\mathcal{T}_{\rho,\eta}^- = \{(i_j, o_j) | r_j^* = 0\}$ as the set of true negative samples. Thus, assuming a uniform prior over ρ, we apply the same reasoning as in Section 3.1 to compute the posterior distribution $\rho \sim Beta(m^{TP} + 1, |\mathcal{T}_{\rho,\eta}^+| - m^{TP} + 1)$, where m^{TP} denotes the number of true positive samples, rated as positive by $M_b^{\rho,\eta}$. Analogously, $\eta \sim Beta(m^{TN} + 1, |\mathcal{T}_{\rho,\eta}^-| - m^{TN} + 1)$, where m^{TN} denotes the number of true negative samples, rated as negative by $M_b^{\rho,\eta}$. Note that to estimate ρ and η, having a large sample size for both $\mathcal{T}_{\rho,\eta}^+$ and $\mathcal{T}_{\rho,\eta}^-$ is important, otherwise the estimation of ρ or η would have a higher uncertainty.

To incorporate the uncertainty of ρ and η into the estimation of α, we need to marginalize ρ and η from the joint likelihood $P(m_+, \rho, \eta|\alpha)$ to get $P(m_+|\alpha)$.

Theorem 2 (Est. α, ρ, η with error-prone metric) *Let $m_+ = \sum_{i=1}^{n} r_j \sim Binom(P(r_j = 1), n)$ be the number of samples rated positively by $M_b^{\rho,\eta}$. Then we estimate α by computing the mode of the following distribution:*

$$\begin{aligned} &P(\alpha|M_+ = m_+) \propto P(M_+ = m_+|\alpha)P(\alpha) \\ &\propto P(\alpha) \int_0^1 \int_0^1 p(m_+|\alpha, \rho, \eta)p(\rho)p(\eta)\mathrm{d}\rho\mathrm{d}\eta \end{aligned} \tag{4}$$

Note that we are not aware of a closed form solution for the above distribution and the computation of the mode. Thus, we approximate the solution using numerical methods in practise (See Appendix B).

3.3 Mixed Case

The mixed case combines the error-free and the error-prone cases. Here, we assume that we are given a small number of error-free samples (human annotations), which are costly to obtain, and a larger set of error-prone samples (ratings by an automated metric), which are easier to obtain [4].

Theorem 3 (Mixed α estimation) *Let* $n_+ = \sum_{i=1}^{|\mathcal{T}_\Phi|} r_j^* \sim Binom(\alpha, |\mathcal{T}_\Phi|)$ *the number of samples where* $M_b^* = 1$, *and* $m_+ = \sum_{i=1}^{n} r_j \sim Binom(P(r_j = 1), |\mathcal{T}_M|)$ *be the number of samples where* $M_b^{\rho,\eta} = 1$. *Then we estimate* α *by computing the mode of the following distribution:*

$$\begin{aligned} &P(\alpha|M_+ = m_+, N_+ = n_+) \\ &\propto P(M_+ = m_+, N_+ = n_+|\alpha)P(\alpha) \\ &\propto P(\alpha|N_+ = n_+) \\ &\times \int_0^1 \int_0^1 P(M_+ = m_+|\alpha, \rho, \eta)p(\rho)p(\eta)\mathrm{d}\rho\mathrm{d}\eta \end{aligned} \tag{5}$$

Note that the difference to the error-prone case is that $P(\alpha)$ is replaced by $P(\alpha|n_+)$, which can be expressed by a closed form beta distribution (see Section 3.1). Thus, we can compute the mixed case by first computing the error-free case to get an initial estimate of α, and then estimate the error-prone case. More generally, this approach lets us also combine ratings from multiple different error-prone metrics by applying Equation 5 iteratively. One would plug in the posterior from one metric as the prior for the next.

Having outlined the estimation of α for different scenarios, we now show how they can be used to determine the minimal number of samples needed to distinguish TGs in a significant manner.

4 Minimal Number of Samples Needed to Make Reliable Distinctions between TG Systems

We now come back to the main question of this paper: how many samples are needed to be able to significantly distinguish the performance of two text generation systems? The intuition is that the closer the performance of the two TG systems is, the more samples are needed. Thus, we investigate the setting where their difference in performance $|\alpha^{\pi_1} - \alpha^{\pi_2}| = \epsilon$ is small. Using the formulas from Section 3, we can compute the estimates shown in Table 1. There are seven variables involved in this computation:

- ρ and η denote the (unknown) performance of the automated binary metric. The better it is, the less samples are needed.
- α denotes the (unknown) performance of the TG system to be evaluated.
- γ as the significance level that is wished to be achieved.
- $|\mathcal{T}_\Phi|$ denotes the size of the set of rated input-output pairs that stem from a error-free binary metric.
- $|\mathcal{T}_\mathcal{M}|$ denotes the size of the set of rated input-output pairs that stem from an error-prone binary metric.
- $|\mathcal{T}_{\rho,\eta}|$ denotes the set of samples needed to estimate ρ and η.

To compute if one system is significantly better, the probability of one system being better than the other must be compared to the significance level (e.g., 0.05). We compute the probability that $\alpha_1 > \alpha_2$ as follows:

$$P(\alpha_1 > \alpha_2) = \int_0^1 \int_{\alpha_2}^1 p(\alpha_1)p(\alpha_2)\mathrm{d}\alpha_1\mathrm{d}\alpha_2 \tag{6}$$

The difference between π^{α_1} and π^{α_2} is significant at the γ-level if $P(\alpha_1 > \alpha_2) < 1 - \frac{\gamma}{2}$ or $P(\alpha_1 > \alpha_2) < \frac{\gamma}{2}$.

Equation 6 holds for any two random variables. In the particular case of normal distributions this is a reformulation of a two-sided z-test of the null hypothesis that both variables have the same mean. Equation 6 is therefore applicable to all the three cases of α estimation (i.e., error-free, error-prone, and mixed) by inserting the posterior distributions.

By applying normal approximations for $p(\alpha_1)$ and $p(\alpha_2)$, and using simulations we can compute the minimal distinguishable difference ϵ for a given set of fixed parameters. The details of the simulations are given in Appendix B.

5 Showcases: Application in Practise

In order to show that the theoretical findings translate to practical applications, we apply our theory to two real-world settings: the WMT21 metric shared task (Freitag et al., 2021b) and the Spot-The-Bot data (Deriu et al., 2020). Since the two tasks have significantly different settings (e.g., machine trans-

[4] Note that our setting also allows for $\mathcal{T}_\Phi \subseteq \mathcal{T}_M$.

lation and dialogue systems, different types of human annotations, and different types of metrics) this shows that our theory is applicable to a variety of text generation tasks. The showcases highlight the different dimensions that can be manipulated when designing an evaluation. In showcase 1, we highlight the number of ratings needed, whereas, in showcase 2, we focus on the influence of the metric performance.

5.1 Showcase 1: WMT Metrics Shared Task

For the WMT21 Metrics shared task, the authors evaluated the performance of 15 automated metrics by comparing their ratings to human ones on the output of several MT systems and several language pairs. In this work, we only focus on the English to German language pair and the news domain, where seven machine translation systems were evaluated. The data provided by the shared task can be expressed as follows using our notation: We regard the expert human multidimensional quality metrics (MQM) (Lommel et al., 2014) annotations as our error-free ratings. We binarize the scalar output of this metric by stating that only translations without any mistakes are regarded as adequate (i.e., $o \in \mathcal{R}^i_+ \iff MQM(i,o) = 0$). This means only responses that have been judged as being completely correct by all annotators are considered adequate. For this setting there are $|\mathcal{T}_\Phi| = 527$ error-free annotated samples for each machine translation system. We can reuse these annotations to estimate ρ and η, thus, $|\mathcal{T}_{\rho,\eta}| = 527$ [5]. For the error-prone metric outputs, WMT provides $|\mathcal{T}_M| = 1000$ samples for each machine translation system and each error-prone metric. For the error-prone metrics, we use *BleuRT* (Sellam et al., 2020) as the metric with the highest ρ and η estimates, and *SacreBLEU* (Post, 2018) as the most popular metric. We consider three machine translation systems: FacebookAI (FBAI) (Tran et al., 2021), VolcTrans-GLAT (VT) (Qian et al., 2021), and HuaweiTSC (HU) (Wei et al., 2021), which have the most interesting combinations of performance (the full Tables are in Appendix D).

5.1.1 WMT: Theoretical Bounds of ϵ

Here, we showcase the theoretical bounds of the ϵ values that can be distinguished significantly depending on the number of ratings and the performance of the metrics. We consider *BleuRT* with an estimated $\rho = \eta \approx 0.6$ (see Section 3.2.2 on how to compute these estimates), *SacreBLEU* with $\rho = \eta \approx 0.52$ and the performances of the machine translation systems are around $\alpha \approx 0.65$ (see section 3.3 on how to compute the estimate). Figure 2 shows the theoretical ϵ values that can be distinguished for various values of $|\mathcal{T}_M|$ and $|T_\Phi|$. For instance, with 527 error-free ($|\mathcal{T}_\Phi|$) and 1000 error-prone samples ($|\mathcal{T}_M|$), we can distinguish an ϵ of 5.6% for both *BleuRT* and *SacreBLEU*. Thus, the impact of the automated metrics is low for higher number of human ratings. However, for $\mathcal{T}_0 = 100$ the impact of the metric performance is larger: $\epsilon = 0.112$ vs. $\epsilon = 0.13$. The effect is even larger with access to more automated ratings. Thus, using 10000 *BleuRT* ratings with 100 human ratings allows to distinguish the same ϵ as with 527 human ratings and 1000 *SacreBLEU* ratings, which is much costlier.

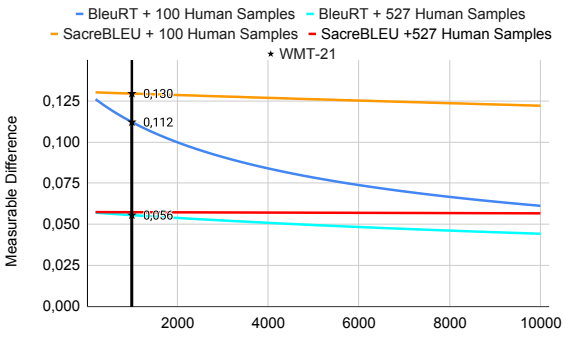


Figure 2: The measurable ϵ depending on $|\mathcal{T}_M|$ (x-Axis) for the BleuRT scenario ($\rho = \eta = 0.6$) and the SacreBLEU scenario ($\rho = \eta = 0.52$) and for $|\mathcal{T}_\Phi| = 100$ or 527. The vertical line shows the WMT setting with $|\mathcal{T}_M| = 1000$.

5.1.2 WMT: Practical Results

Here we analyse the results obtained when applying the theoretical framework to real data to estimate α, and assess whether the pairwise differences are significant or not. Table 2 shows the results for four scenarios: using all 527 error-free ratings, using only 100 error-free ratings (low-cost scenario), using 100 error-free ratings with an additional 1000 error-prone ratings from SacreBLEU, and using 100 error-free ratings with an additional 1000 error-prone ratings from BleuRT. The results include for each pair of systems the estimated ϵ values, and the probability that the first TG system is better than

[5] Note that we estimate ρ and η for each machine translation system separately since we noted that most trained metrics have different performances depending on the various machine translation systems. See Appendix C.

$\lvert\mathcal{T}_\Phi\rvert = 527, \lvert\mathcal{T}_\mathcal{M}\rvert = 0$		
$\pi(\alpha_1) - \pi(\alpha_2)$	ϵ	$P(\alpha_1 > \alpha_2)$
FBAI(0.67) - VT (0.64)	0.02	0.798
FBAI(0.67) - HU (0.58)	**0.09**	0.998
VT(0.64) - HU (0.58)	**0.06**	0.978
$\lvert\mathcal{T}_\Phi\rvert = 100, \lvert\mathcal{T}_\mathcal{M}\rvert = 0$		
$\pi(\alpha_1) - \pi(\alpha_2)$	ϵ	$P(\alpha_1 > \alpha_2)$
FBAI(0.67) - VT (0.65)	0.02	0.615
FBAI(0.67) - HU (0.58)	0.09	0.904
VT(0.65) - HU (0.58)	0.07	0.843
SacreBLEU: $\lvert\mathcal{T}_\Phi\rvert = 100, \lvert\mathcal{T}_\mathcal{M}\rvert = 1000$		
$\pi(\alpha_1) - \pi(\alpha_2)$	ϵ	$P(\alpha_1 > \alpha_2)$
FBAI(0.67) - VT (0.64)	0.02	0.631
FBAI(0.67) - HU (0.57)	0.09	0.918
VT(0.64) - HU (0.57)	0.07	0.854
BleuRT: $\lvert\mathcal{T}_\Phi\rvert = 100, \lvert\mathcal{T}_\mathcal{M}\rvert = 1000$		
$\pi(\alpha_1) - \pi(\alpha_2)$	ϵ	$P(\alpha_1 > \alpha_2)$
FBAI(0.66) - VT (0.62)	0.04	0.742
FBAI(0.66) - HU (0.56)	0.09	0.933
VT(0.62) - HU (0.56)	0.05	0.801

Table 2: Predicted WMT21 evaluation using BleuRT and SacreBLEU on three machine translation systems.

the second system. In the first scenario, we see that FBAI and VT cannot be significantly distinguished, which is consistent with the theory that states only $\epsilon > 0.057$ can be distinguished (see Figure 2), whereas the other system pairs can be distinguished. In the second scenario, we reduce the number of error-free samples to only $|\mathcal{T}_\Phi| = 100$, which makes all the TG systems non distinguishable from each other. Again, this is consistent with the theory that states only $\epsilon > 0.131$ can be distinguished using 100 consistent samples. When we add error-prone ratings, the probabilities of the first TG being better than the second increase, however not enough to be significantly distinguishable. This goes for both automated metrics, which is still consistent with the theory. The problem lies in the fact that the performance of the automated metrics is too low to have a strong impact on the evaluation. For instance, the theory predicts that using $10'000$ error-prone SacreBLEU samples will only lead to being able to distinguish $\epsilon > 0.120$. In this setting, adding even more error-prone samples will not help (even with $|\mathcal{T}_M| = 10^9$), since the uncertainty of ρ and η is too high due to $|\mathcal{T}_{\rho,\eta}| = 527$.

Thus, the practical application shows that the outcomes using real data is consistent with the theory. Unfortunately, the setting does not allow to distinguish FBAI and VT. For this more error-free ratings are needed, or better metrics.

5.2 Showcase 2: Spot The Bot (STB)

For the second show case, we use the Spot The Bot (STB) data, where dialogues between two dialogue systems are sampled and humans classified each interlocutor to be a human or a bot. STB contains pairwise ratings for six dialogue systems. In our setting, we use three of them: Blenderbot (BL) (Roller et al., 2021), Lost in Conversation [6] (LiC), and KVMemNN (KV) (Dinan et al., 2020). In this setting the error-free metric is the (aggregated) human judgment, which is already binary. We consider a response as adequate if all annotators labelled it as coming from a human. For the error-prone metric, we use the USR (Mehri and Eskenazi, 2020) metric, which is also a scalar metric that we binarize with a threshold[7]. The STB dataset yields $|\mathcal{T}_\Phi| = |\mathcal{T}_{\rho,\eta}| \approx 600$ error-free ratings per dialogue system. For creating $\mathcal{T}_M$, we sample new pairwise dialogues and let USR rate each turn of the dialogue. This yields $|\mathcal{T}_M| = 10'000$ samples per dialogue system.

5.2.1 STB: Theoretical Bounds

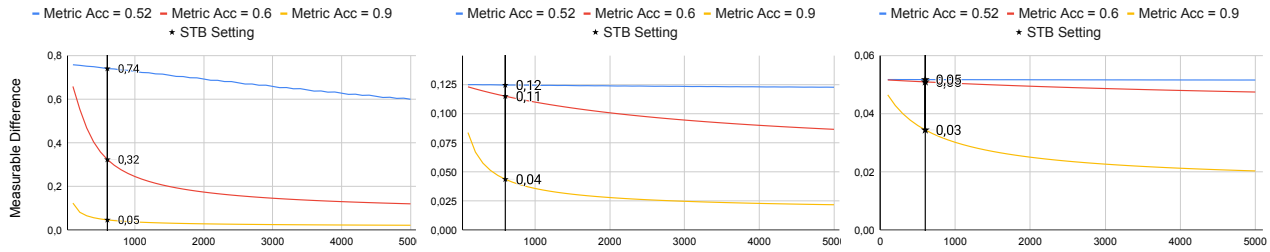
Figure 3 shows the theoretical ϵ values that can be achieved depending on $|\mathcal{T}_{\rho,\eta}|$. The values are depicted for three different settings of $|\mathcal{T}_\Phi|$ (i.e, human ratings). Each setting shows the measurable ϵ for three different $\rho = \eta$ combinations. The figure reveals the impact of $|\mathcal{T}_{\rho,\eta}|$ for $|\mathcal{T}_{\rho,\eta}| < 1000$. For instance, for $|\mathcal{T}_{\rho,\eta}| = 600$, a metric with $\rho = \eta = 0.6$ is only able to distinguish an $\epsilon = 0.11$, however, when increasing $|\mathcal{T}_{\rho,\eta}|$ to 5000 a difference of $\epsilon = 0.08$ can be measured. On the other hand, when the performance of the metric is too low (e.g., $\rho = \eta = 0.52$) the impact of higher $|\mathcal{T}_{\rho,\eta}|$ is negligible regardless of $|\mathcal{T}_\Phi|$.

5.2.2 STB: Practical Results

Table 3 shows the measured values for α and ϵ for three scenarios. The first two scenarios are analogous to the WMT setting, where we use $|\mathcal{T}_\Phi| = 600$ error-free ratings in the first scenario and $|\mathcal{T}_\Phi| = 100$ error-free ratings in the second scenario (assuming that we labeled only 100 samples due to cost reasons). For the third scenario we again use $|\mathcal{T}_\Phi| = 100$ error-free ratings, combined with $|\mathcal{T}_M| = 10'000$ error-prone ratings from the USR metric. The results show that for the first scenario all the pairs of systems are distinguishable,

[6] https://github.com/atselousov/transformer_chatbot

[7] See Appendix C.



(a) $|\mathcal{T}_\Phi| = 0$ (b) $|\mathcal{T}_\Phi| = 100$ (c) $|\mathcal{T}_\Phi| = 600$

Figure 3: Measurable difference (y-axis) depending on the number of samples available to estimate ρ and η (i.e, $|\mathcal{T}_{\rho,\eta}|$ on the x-axis). For fixed $|\mathcal{T}_M| = 10'000$. The vertical line denotes the STB setting with $|\mathcal{T}_{\rho,\eta}| = 600$

$\pi(\alpha_1) - \pi(\alpha_2)$	ϵ	$P(\alpha_1 > \alpha_2)$
$\|\mathcal{T}_\Phi\| = 600, \|\mathcal{T}_\mathcal{M}\| = 0$		
BL (0.38) - LiC (0.30)	0.08	0.999
BL (0.30) - KV (0.24)	0.13	1.000
LiC (0.30) - KV (0.24)	0.06	0.989
$\|\mathcal{T}_\Phi\| = 100, \|\mathcal{T}_\mathcal{M}\| = 0$		
BL (0.38) - LiC (0.30)	0.08	0.882
BL (0.38) - KV (0.25)	0.14	0.983
LiC (0.30) - KV (0.25)	0.06	0.827
USR: $\|\mathcal{T}_\Phi\| = 100, \|\mathcal{T}_\mathcal{M}\| = 10000$		
BL (0.36) - LiC (0.28)	0.08	0.889
BL (0.36) - KV (0.22)	0.13	0.989
LiC (0.28) - KV (0.22)	0.06	0.851

Table 3: Predicted STB evaluation using USR on three dialogue systems.

which is consistent with the theory and the original Spot The Bot results. When reducing the number of error-free samples to $|\mathcal{T}_\Phi| = 100$, only the pair BL-KV is distinguishable. This is consistent with the theory, which predicts that two systems with $\epsilon > 0.126$ are significantly distinguishable. However, adding $|\mathcal{T}_M| = 10'000$ error-prone ratings only increases the probability of the first TG system being better than the second by a small amount. The reason is that the performance of USR is too low to have a strong impact, which is consistent with the theory. Thus, to benefit from automated evaluation one needs a better metric and more samples to estimate ρ and η.

6 Related Work

Evaluation of Text Generation systems is a long-standing issue. Considerations about the proper evaluation of TG systems have been emerging rapidly in the last years. One line of inquiry is how to properly conduct human evaluations and what kind of guidelines and setups lead to consistent results (Novikova et al., 2018; Van Der Lee et al., 2019; Santhanam and Shaikh, 2019; Freitag et al., 2021a; Clark et al., 2021; Belz et al., 2021; Mohankumar and Khapra, 2022). Another line of research investigates the reliability of automated metrics for NLG evaluation. Novikova et al. (2017, inter alia) find that automated metrics poorly reflect human judgements in general. Sai et al. (2022, Sec. 6) provides an extensive overview of criticism on automated metrics in NLG.

There are few efforts to underlay (parts of the) TG evaluation paradigm with a theory-grounded base: To theoretically solidify human NLG evaluation and provide more statistically significant results in pairwise evaluations, a recent approach leverages utility theory in economics (Ethayarajh and Jurafsky, 2022) to showcase issues arising from the use of Likert scale ratings and averaging them. Chaganty et al. (2018) propose a method to combine automated metrics with human rankings to debiase a metric under a budget constraint. They provide a theory-grounded proof that their calculated mix of human and automated ratings is optimal and conclude that error-prone evaluation metrics are a bottleneck for reducing the cost of evaluations. Related to our Bayesian approach of modelling uncertainty in the evaluation of systems, a number of approaches aims to model uncertainty in the annotation process and the aggregation of annotations using a Bayesian approach (Paun et al., 2018, e.g.). Card et al. (2020) analyze the statistical power of different evaluation scenarios prevalent in NLP. In particular, they study the number of samples needed to detect a difference of 1 BLEU as significant. However, to the best of our knowledge, no efforts to model the uncertainties ingrained in TG evaluation in a holistic theory has been proposed so far.

7 Conclusion

We introduced a theoretical framework for binary metrics that can be used to extract guidelines for designing an evaluation of text generation systems. The framework estimates the performance of a text generation system from a mix of human and automated ratings giving guarantees of which level of significance can be achieved. Using the formulas, one can design the evaluation setup and compute estimates of how many human and automated samples are needed for a significant evaluation. We applied the theory to two very different real-world cases and exemplified how the theory can be leveraged to improve the significance of the results. We provide a tool that allows the computation of the formulas so that different settings can be tested.

The current theory is limited to binary metrics, but in future work, we will extend the theory to more types, such as comparative or scalar metrics. Furthermore, we will apply the theory to a wider range of tasks and domains. In general, we hope to have set in motion efforts to arrive at a sound formalization of the evaluation of text generation systems to increase the robustness, reliability, and significance of future evaluation campaigns.

Limitations

Human Ratings. We assume that human ratings are perfect, which is not the case (Clark et al., 2021). While it might be the case that the MQM ratings are close to error-free, there is no guarantee. To handle the fact that human ratings are not error-free we would need to measure this, which could be done via agreement scores.

Uniform Input and Outputs. We assume that each input and each output have the same difficulty of being evaluated. However, it is more likely that in practise, each metric has a different ρ and η value depending on the input. This is however very hard to include in the theory.

Uniform Text Generation Systems. Similarly to the above point, we assume that ρ and η are independent of the text generation system. However, preliminary experimental results (see Appendix C) showed that metrics tend to have different performances for different TG systems. Thus, ρ and η need to be estimated separately for each TG system.

Domain Dependence. The same argument can also be made about the domain. Metrics trained on one domain will perform differently when applied to another domain. Thus, the ρ and η values must be measured again for each domain.

Binary Metrics. The current theory is limited to binary metrics. However, in practise there are many different types of metrics and evaluation types. For instance, in a next step the theory should be extended to cover comparative metrics (i.e., metrics that state which of the two outputs is better).

Approximations. The estimations of the mixed case and the estimated ρ, η case must be approximated numerically since we did not find a closed form solution. This will inevitably lead to mistakes in the estimated values. This can be circumvented by making the numerical approximation more precise with the downside of needing more computational power (see Appendix B).

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. The use of rating and Likert scales in natural language generation human evaluation tasks: A review and some recommendations. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 397–402, Tokyo, Japan. Association for Computational Linguistics.

Anja Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021. The reprogen shared task on reproducibility of human evaluations in nlg: Overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 249–258.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in

natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653.

Pinzhen Chen, Jindřich Helcl, Ulrich Germann, Laurie Burchell, Nikolay Bogoychev, Antonio Valerio Miceli Barone, Jonas Waldendorf, Alexandra Birch, and Kenneth Heafield. 2021. The University of Edinburgh's English-German and English-Hausa submissions to the WMT21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 104–109, Online. Association for Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

Jan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Alvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. 2020. Spot the bot: A robust and efficient framework for the evaluation of conversational dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3971–3984, Online. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.

Kawin Ethayarajh and Dan Jurafsky. 2022. How human is human evaluation? Improving the gold standard for NLG with utility theory. *arXiv preprint arXiv:2205.11930*.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció*, (12):455–463.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Akash Kumar Mohankumar and Mitesh M Khapra. 2022. Active evaluation: Efficient nlg evaluation with few pairwise comparisons. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8761–8781.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. Rankme: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on*

Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Lihua Qian, Yi Zhou, Zaixiang Zheng, Yaoming Zhu, Zehui Lin, Jiangtao Feng, Shanbo Cheng, Lei Li, Mingxuan Wang, and Hao Zhou. 2021. The volctrans GLAT system: Non-autoregressive translation meets WMT21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 187–196, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215, Online. Association for Computational Linguistics.

Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, Hao Yang, and Ying Qin. 2021. HW-TSC's participation in the WMT 2021 news translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 225–231, Online. Association for Computational Linguistics.

A Derivations for α-estimation

In Section 3 we have introduced several ways to estimate the success rate α of a Text-Generator π. We will now elaborate some of these in more detail.

First, we want to estimate α based on consistent ratings from M_b^*. For this we need a set of inputs, the corresponding outputs from π, and the rating from M_b^*: $\mathcal{T}_\Phi = \{(i_j, o_j, r_j^*) | 1 \leq j \leq n_\phi\}$, where $o_j = \pi(i_j)$ and $r_j^* = M_b^*(i_j, o_j)$. We note that, in this case, the probability that a given pair is rated adequate is α, since:

$$\begin{aligned} P(r_j^* = 1) &= P(M_b^*(i_j, \pi(i_j)) = 1) \\ &= P(\pi(i_j) \in \mathcal{R}_+^{i_j}) \\ &= \alpha \end{aligned}$$

We can therefore treat r_j^* as outcomes of Bernoulli trials with success probability α. The number of successful trials N_+ is therefore a random variable with binomial distribution: $N_+ \sim Binom(\alpha, n_\phi)$. The concrete outcome for a given experiment is $n_+ = \sum_{j=1}^{n_\phi} r_j^*$. To estimate α we use the proportion of successful trials, meaning the fraction of adequate responses: $\tilde{\alpha} = \frac{n_+}{n_\phi}$. Due to the Law of Large Numbers this will converge to the expected value $\mathbb{E}[r_j] = \alpha$.

Bayesian Formulation We choose to work in a Bayesian framework as it provides a convenient way to unify the multiple sources of evidence and uncertainty we want to tackle. The first source of information comes from $\mathcal{T}_\Phi$. In particular we have seen that the number of input-output pairs rated as adequate, N_+, follows a binomial distribution. This means that $P(N_+ = n_+|\alpha) = \binom{n_\phi}{n_+}\alpha^{n_+}(1-\alpha)^{n_\phi - n_+}$. We want to derive a posterior distribution for α based on the evidence: $p(\alpha|N_+ = n_+)$. For this we can apply Bayes' Theorem: $p(\alpha|N_+ = n_+) \propto P(N_+ = n_+|\alpha)p(\alpha)$, where $p(\alpha)$ expresses our prior belief of the possible values for α. In this setting $p(\alpha)$ is called the prior, $P(N_+ = n_+|\alpha)$ likelihood, and $p(\alpha|N_+ = n_+)$ the posterior. Since we in general cannot assume anything about α we choose a uniform prior $\alpha \sim \mathcal{U}(0, 1)$. This means before seeing any evidence we consider any possible value of α to be equally likely. Of course there are other reasonable choices for priors, but in general uniform priors are

a good choice since the resulting estimators will closely match traditional frequentist approaches.

Another approach is to choose a so-called conjugate prior based on the type of likelihood we are confronted with. A conjugate prior for a given likelihood will result in a posterior from the same family (but different parameters) as the prior. In our case, the Beta distribution is a conjugate prior for a Binomial likelihood. Beta distributions have two shape parameters a and b and assuming $\alpha \sim Beta(a, b)$ then $p(\alpha) = \frac{\alpha^{a-1}(1-\alpha)^{b-1}}{B(a,b)}$. Here $B(a, b)$ is the beta function of a and b and serves as the normalizing constant, ensuring that $p(\alpha)$ integrates to 1. The beta function is defined in terms of the Gamma function Γ, an extension of factorials.

Luckily, we can show that $\mathcal{U}(0, 1)$ and $Beta(1, 1)$ are the same distribution. We first note that both distributions are defined on the same domain $(0, 1)$. In particular, the uniform distribution is constant 1 over the domain. By definition of the Beta distribution we have that if $\alpha \sim Beta(1, 1)$ then $p(\alpha) = \frac{\alpha^{1-1}(1-\alpha)^{1-1}}{B(1,1)} = \frac{1}{B(1,1)} = 1$.

Next we will show how to compute the posterior for the general case where $\alpha \sim Beta(a, b)$:

$$
\begin{aligned}
&p(\alpha|N_+ = n_+) \\
&\propto P(N_+ = n_+|\alpha)p(\alpha) \\
&\propto \binom{n_\phi}{n_+}\alpha^{n_+}(1-\alpha)^{n_\phi - n_+}\frac{\alpha^{a-1}(1-\alpha)^{b-1})}{B(a,b)} \\
&\propto \alpha^{n_+ + a - 1}(1-\alpha)^{n_\phi - n_+ + b - 1} \\
&\sim Beta(n_+ + a, n_\phi - n_+ + b)
\end{aligned}
$$

We see that the resulting posterior is indeed another Beta distribution. In particular if we choose $a = b = 1$, or a uniform prior, we get that $\alpha|N_+ = n_+ \sim Beta(n_+ + 1, n_\phi - n_+ + 1)$ as in Section 3.

Error-prone Metric Next, we want to estimate α given a set of inputs, outputs from π, and ratings from a error-prone metric $M_b^{\rho,\eta}$ with known ρ and η. We define $\mathcal{T}_M = \{(i_j, o_j, r_j)|1 \leq j < n_M\}$ where $o_j = \pi(i_j)$ and $r_j = M_b^{\rho,\eta}(i_j, o_j)$. The probability that any given r_j is 1 is:

$$
\begin{aligned}
&P(r_j = 1) \\
&= P(M_b^{\rho,\eta}(i_j, \pi(i_j) = 1) \\
&= P(M_b^{\rho,\eta}(i_j, \pi(i_j)) = 1|\pi(i_j) \in \mathcal{R}_+^{i_j}) \\
&P(\pi(i_j) \in \mathcal{R}_+^{i_j}) \\
&+ P(M_b^{\rho,\eta}(i_j, \pi(i_j)) = 1|\pi(i_j) \notin \mathcal{R}_+^{i_j}) \\
&P(\pi(i_j) \notin \mathcal{R}_+^{i_j}) \\
&= \rho\alpha \\
&+ (1 - P(M_b^{\rho,\eta}(i_j, \pi(i_j)) = 0|\pi(i_j) \notin R_+^{i_j})) \\
&(1 - P(\pi(i_j) \in R_+^{i_j})) \\
&= \rho\alpha + (1-\eta)(1-\alpha) \\
&= \alpha(\rho + \eta - 1) + (1 - \eta)
\end{aligned}
$$

What we can concretely measure (or count) on $\mathcal{T}_M$ is the number of times the error-prone metric gives an adequate rating. We define this as $m_+ = \sum_{j=1}^{n_M} r_j$. Since we sum n_M Bernoulli trials with success rate $\alpha(\rho + \eta - 1) + (1 - \eta)$, the sum has a Binomial distribution: $M_+ \sim Binom(\alpha(\rho + \eta - 1) + (1 - \eta), n_M)$. Therefore our likelihood is:

$$
\begin{aligned}
&P(M_+ = m_+|\alpha, \rho, \eta) \\
&= \binom{n_M}{m_+}(\alpha(\rho + \eta - 1) + (1 - \eta))^{m_+} \\
&(1 - (\alpha(\rho + \eta - 1) + (1 - \eta)))^{n_M - m_+}
\end{aligned}
$$

We notate the likelihood as $P(M_+ = m_+|\alpha, \rho, \eta)$ to indicate the dependence on ρ and η, even though they are assumed deterministic. Unfortunately, we are not aware of any conjugate prior for α that would allow us to derive a closed form posterior from this likelihood. Nevertheless, we can show that for $\alpha \sim \mathcal{U}(0, 1)$ the mode of the posterior is at $\frac{\frac{m_+}{n_M} - (1-\eta)}{\rho + \eta - 1}$. For this we will have to find the point where the derivative of the posterior with respect to α is 0. To simplify the notation we will write $f(\alpha) = \alpha(\rho + \eta - 1) + (1 - \eta)$ and $f'(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha} f(\alpha) = \rho + \eta - 1$.

We will first compute the derivative of the posterior with respect to α using a uniform prior (i.e.

$p(\alpha) = 1$):

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\alpha} p(\alpha | M_+ = m_+) \\
&\propto \frac{\mathrm{d}}{\mathrm{d}\alpha} P(M_+ = M_+ | \alpha, \rho, \eta) p(\alpha) \\
&\propto \frac{\mathrm{d}}{\mathrm{d}\alpha} P(M_+ = M_+ | \alpha, \rho, \eta) 1 \\
&\propto \frac{\mathrm{d}}{\mathrm{d}\alpha} (f(\alpha)^{m_+} (1 - f(\alpha))^{n_M - m_+}) \\
&\propto (\frac{\mathrm{d}}{\mathrm{d}\alpha} f(\alpha)^{m_+})(1 - f(\alpha))^{n_M - m_+} \\
&+ f(\alpha)^{m_+} (\frac{\mathrm{d}}{\mathrm{d}\alpha} (1 - f(\alpha))^{n_M - m_+}) \\
&\propto m_+ f(\alpha)^{m_+ - 1} f'(\alpha) (1 - f(\alpha))^{n_M - m_+} \\
&+ f(\alpha)^{m_+} (n_M - m_+) (1 - f(\alpha))^{n_M - m_+ - 1} (-f'(\alpha)) \\
&\propto m_+ f'(\alpha) f(\alpha)^{m_+ - 1} (1 - f(\alpha))^{n_M - m_+} \\
&- (n_M - m_+) f'(\alpha) f(\alpha)^{m_+} (1 - f(\alpha))^{n_M - m_+ - 1}
\end{aligned}$$

To find the mode we set the derivative to zero and solve for α. We will use the convenient fact that $f'(\alpha)$ is constant independent of α:

$$\begin{aligned}
&m_+ f'(\alpha) f(\alpha)^{m_+ - 1} (1 - f(\alpha))^{n_M - m_+} \\
&- (n_M - m_+) f'(\alpha) f(\alpha)^{m_+} (1 - f(\alpha))^{n_M - m_+ - 1} \\
&= 0 \iff \\
&m_+ f'(\alpha) f(\alpha)^{m_+ - 1} (1 - f(\alpha))^{n_M - m_+} \\
&= (n_M - m_+) f'(\alpha) f(\alpha)^{m_+} (1 - f(\alpha))^{n_M - m_+ - 1} \\
&m_+ f(\alpha)^{m_+ - 1} (1 - f(\alpha))^{n_M - m_+} \\
&= (n_M - m_+) f(\alpha)^{m_+} (1 - f(\alpha))^{n_M - m_+ - 1} \\
&m_+ (1 - f(\alpha))^{n_M - m_+} \\
&= (n_M - m_+) f(\alpha) (1 - f(\alpha))^{n_M - m_+ - 1} \\
&m_+ (1 - f(\alpha)) = (n_M - m_+) f(\alpha) \\
&m_+ = (n_M - m_+) f(\alpha) + m_+ f(\alpha) \\
&m_+ = (n_M - m_+ + m_+) f(\alpha) \\
&\frac{m_+}{n_M} = f(\alpha) = \alpha(\rho + \eta - 1) + (1 - \eta) \\
&\alpha = \frac{\frac{m_+}{n_M} - (1 - \eta)}{\rho + \eta - 1}
\end{aligned}$$

Uncertainty in ρ and η If we do not already know the specific ρ and η for a given error-prone metric, we will have to estimate them from data. For this we need ratings from a the error-prone metric as well as an error-free metric to compare to. Assume we are given the set $\mathcal{T}_{\rho,\eta} = \{(i_j, o_j, r_j, r_j^*) | 1 \leq j < n_{\rho,\eta}\}$, where $r_j = M_b^{\rho,\eta}(i_j, o_j)$ and $r_j^* = M_b^*(i_j, o_j)$. Note that unlike $\mathcal{T}_\Phi$ and $\mathcal{T}_M$ we do not make any assumptions about how o_j was generated.

By definition ρ is the true positive rate of the error-prone metric and η the true negative rate. We can therefore estimate them independently from each other by splitting $\mathcal{T}_{\rho,\eta}$ into two sets based on whether r_j^* is 1 or 0: $\mathcal{T}_{\rho,\eta}^+ = \{(i, o, r, r^*) \in \mathcal{T}_{\rho,\eta} | r^* = 1\}$ and $\mathcal{T}_{\rho,\eta}^- = \{(i, o, r, r^*) \in \mathcal{T}_{\rho,\eta} | r^* = 0\}$.

To estimate ρ we have to count the number of times $r_j = 1$ when $r_j^* = 1$ too, in other words we have to count the number of true adequate ratings: $n_{TP} = \sum_{i,o,r,r* \in \mathcal{T}_{\rho,\eta}^+} r$. By definition we know that $\rho = P(r = 1 | r* = 1)$ and therefore $N_{TP} \sim Binom(\rho, |\mathcal{T}_{\rho,\eta}^+|)$. We can apply the same Bayesian reasoning as at the start of this Appendix to derive a posterior distribution for ρ. Assuming a uniform prior over ρ, we have that $\rho | n_{TP} \sim Beta(n_{TP} + 1, |\mathcal{T}_{\rho,\eta}^+| - n_{TP} + 1)$. The estimation of η is exactly analogous.

At this point we could just use point estimates for ρ and η and treat them as deterministic like above. Unfortunately this has a high chance of throwing off the point estimate (mode) of α.

We will therefore consider the joint likelihood $P(M_+ = m_+, \rho, \eta | \alpha)$ and marginalize ρ and η. We will reuse results from above. Recall we were given the set $\mathcal{T}_M = \{(i_j, o_j, r_j) | 1 \leq j < n_M\}$ where $o_j = \pi(i_j)$ and $r_j = M_b^{\rho,\eta}(i_j, o_j)$. We counted the number of adequate ratings $m_+ = \sum_{j=1}^{n_M} r_j$ and we saw that $P(M_+ = M_+ | \alpha, \rho, \eta) = \alpha(\rho + \eta - 1) + (1 - \eta)$. Based on that we can compute the likelihood as follows:

$$\begin{aligned}
&P(M_+ = m_+ | \alpha) \\
&= \int_0^1 \int_0^1 P(M_+ = m_+, \rho, \eta | \alpha) \mathrm{d}\rho \mathrm{d}\eta \\
&= \int_0^1 \int_0^1 P(M_+ = m_+ | \rho, \eta, \alpha) p(\rho) p(\eta) \mathrm{d}\rho \mathrm{d}\eta
\end{aligned}$$

and the posterior as follows:

$$\begin{aligned}
&p(\alpha | M_+ = m_+) \\
&\propto p(\alpha) P(M_+ = m_+ | \alpha) \\
&\propto p(\alpha) \int_0^1 \int_0^1 P(M_+ = m_+ | \rho, \eta, \alpha) p(\rho) p(\eta) \mathrm{d}\rho \mathrm{d}\eta
\end{aligned}$$

We will show how approximate this numerically in Appendix B.

Combining error-free and error-prone ratings Finally, we show how we can combine both error-free and error-prone ratings into a single estimate for α. Here we assume that we have estimates for

ρ and η, for example in the form of Beta-posteriors, as derived previously: $\rho \sim Beta(a_\rho, b_\rho)$ and $\eta \sim Beta(a_\eta, b_\eta)$. Similarly, we build upon the previous setting where we counted the number of adequate ratings from the error-free metric, $N_+ \sim Binom(\alpha, n_\phi)$, and the number of adequate ratings from the error-prone metric, $M_+ \sim Binom(\alpha(\rho+\eta-1)+(1-\eta), n_M)$. Our observed n_+ and m_+ have the joint likelihood:

$$\begin{aligned} &P(M_+ = m_+, N_+ = n_+|\alpha) \\ &= P(M_+ = m_+|\alpha)P(N_+ = n_+|\alpha) \end{aligned}$$

We assume here that M_+ and N_+ are independent when conditioned on α.

We are now ready to compute the posterior for α. Using a Beta prior $\alpha \sim Beta(a_\alpha, b_\alpha)$ we get:

$$\begin{aligned} &p(\alpha|M_+ = m_+, N_+ = n_+) \\ &\propto p(\alpha)P(M_+ = m_+, N_+ = n_+|\alpha) \\ &\propto p(\alpha)P(N_+ = n_+|\alpha)P(M_+ = m_+|\alpha) \\ &\propto p(\alpha)P(N_+ = n_+|\alpha) \\ &\int_0^1 \int_0^1 P(M_+ = m_+|\alpha, \rho, \eta)p(\rho)p(\eta)\mathrm{d}\rho\mathrm{d}\eta \\ &\propto p(\alpha|N_+ = n_+) \\ &\int_0^1 \int_0^1 P(M_+ = m_+|\alpha, \rho, \eta)p(\rho)p(\eta)\mathrm{d}\rho\mathrm{d}\eta \end{aligned}$$

Looking at the last step, we see that we can combine the prior $p(\alpha)$ with the partial likelihood $P(N_+ = n_+|\alpha)$ to get a partial posterior $p(\alpha|N_+ = n_+)$ that gets multiplied with the likelihood of M_+. We have already seen that since α has a Beta prior and N_+ has a binomial likelihood, $\alpha|N_+$ is also a Beta distribution. This suggests a two-step procedure, where in the first step we derive a posterior from error-free ratings. In the second step we use that estimate as the new prior for deriving the posterior from error-prone ratings.

Notes on $\mathcal{T}_\Phi$, $\mathcal{T}_M$, and $\mathcal{T}_{\rho,\eta}$ Note that in practise there are some considerations to be made. Since we use human ratings, we can use them both for estimating ρ and η but also to estimate α. Thus, we use $T_\Phi = T_{\rho,\eta}$, which is also necessary since ρ and η are different for each TG system (see example in Appendix C). Thus, it is often not advisable to use the ratings for other systems to estimate ρ and η. However, this phenomenon needs to be explored in more detail.

For the estimation of ρ and η, we need to make sure that $T^+_{\rho,\eta}$ and $T^-_{\rho,\eta}$ are of large enough size. Since if we have only a few samples in $T_{\rho,\eta}$ where $r^*_j = 0$ then the estimate for η will be uncertain. This can be problematic when evaluating very strong or very poor systems (e.g., $\alpha > 0.9$ or $\alpha < 0.1$) as there will be only a few samples with $r^*_j = 0$ or $r^*_j = 1$ respectively.

In many cases we can reuse the samples in T_Φ for T_M, i.e., $T_\Phi \subseteq T_M$ since we can use the automated metric to rate the samples, which were annotated by humans. However, it is not clear what effect this will have on the final estimate of ϵ. Exploring this phenomenon is part of future work.

B Derivations for ϵ-simulation

In this section we will show how we derive the values for the minimally distinguishable difference between two systems. We do this by first simulating a concrete experiment based on theoretical parameters. We substitute the simulated experiment into Equation 5. We will also show how we numerically approximate Equation 5.

Simulation Until now we have considered the case where α , and possibly ρ and η, are unknown and need to be estimated from data. In that case we use Equation 5 to derive a posterior estimate for α. The whole estimation is based on counts from three sources $\mathcal{T}_\Phi$, $\mathcal{T}_M$, and $\mathcal{T}_{\rho,\eta}$. Assume we know the following properties: α, ρ, η, $n_\phi = |\mathcal{T}_\Phi|$, $n_M = |\mathcal{T}_M|$, $n_{\rho,\eta} = |\mathcal{T}_{\rho,\eta}|$, as well as the proportion ψ of truly adequate responses in $\mathcal{T}_{\rho,\eta}$.

To simulate the number of adequate ratings from the error-free metric n_+ we round its expected value, $\mathbb{E}[n_+] = \alpha n_\phi$, to the nearest integer: $n^{sim}_+ = \lfloor \alpha n_\phi + \frac{1}{2} \rfloor$. To simulate the number of adequate ratings from the error-prone metric m_+, we round its expected value, $\mathbb{E}[m_+] = (\alpha(\rho+\eta-1)+(1-\eta))n_M$, to the nearest integer: $m^{sim}_+ = \lfloor (\alpha(\rho+\eta-1)+(1-\eta))n_M + \frac{1}{2} \rfloor$. We have seen that to estimate ρ we need to know the number of true positive ratings n_{TP} of the error-prone metric as well as the total number of positive ratings in $\mathcal{T}_{\rho,\eta}$ which we notated as $|\mathcal{T}^+_{\rho,\eta}| = n^*_p$. We can simulate the latter by rounding its expected value, $\mathbb{E}[n^*_p] = \psi n_{\rho,\eta}$, to the nearest integer: $n^{sim}_p = \lfloor \psi n_{\rho,\eta} + \frac{1}{2} \rfloor$. To simulate n_{TP} we have to plug the simulated n^{sim}_p into the expected value: $n^{sim}_{TP} = \lfloor \rho n^{sim}_p + \frac{1}{2} \rfloor$. Finally, we follow the same process to simulate the data for η. Let $n^*_n = |\mathcal{T}^-_{\rho,\eta}|$, which we simulate as $n^{sim}_n = n_{\rho,\phi} - n^{sim}_p$. The number of true negatives of the error-prone metric is simulated as: $n^{sim}_{TN} = \lfloor \eta(n_{\rho,\phi} - n^{sim}_p) + \frac{1}{2} \rfloor$.

We can then use these simulated values to calculate the calculate the posterior $p^{sim}(\alpha)$ based on Equation 5. For this we first have to simulate our belief over ρ and η: $\rho^{sim} \sim Beta(n_{TP}^{sim} + 1, n_p^{sim} - n_{TP}^{sim} + 1)$ and $\eta^{sim} \sim Beta(n_{TN}^{sim} + 1, n_n^{sim} - n_{TN}^{sim} + 1)$. We again set a uniform prior, $\alpha \sim Beta(1,1)$ and compute the simulated posterior:

$$p^{sim}(\alpha|N_+ = n_+^{sim}, M_+ = n_+^{sim})$$
$$\propto p(\alpha)P(N_+ = n_+^{sim}|\alpha)$$
$$\int_0^1 \int_0^1 P(M_+ = m_+^{sim}|\alpha,\rho,\eta)p^{sim}(\rho)p^{sim}(\eta)\mathrm{d}\rho\mathrm{d}\eta$$

For the tables in Appendix E we make the following simplifying assumptions: we assume that the input-output pairs in $\mathcal{T}_\Phi$ and $\mathcal{T}_{\rho,\eta}$ are the same. This means that $n_{\rho,\eta} = n_\phi$ and $\psi = \alpha$.

Computing ϵ_γ We will now show how we use $p^{sim}(\alpha)$ to compute the minimal distinguishable difference between two systems π_1 with success rate α_1 and π_2 with success rate α_2.

Assume we know the distributions $p(\alpha_1)$ and $p(\alpha_2)$, we can then compute their means $\mu_i = \mathbb{E}[\alpha_i]$ and variances $\sigma_i^2 = \mathbb{V}[\alpha_i]$. These can be used to derive normal approximations for α_i: $\alpha_i^{\mathcal{N}} \sim \mathcal{N}(\mu_i, \sigma_i)$. In that case the difference $\epsilon = \alpha_1^{\mathcal{N}} - \alpha_2^{\mathcal{N}}$ also follows a normal distribution: $\mathcal{N}(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_1^2})$. We can now formulate a z-test to see whether there is a significant difference between $\alpha_1^{\mathcal{N}}$ and $\alpha_1^{\mathcal{N}}$. The null hypothesis H_0 is that both systems perform the same, meaning $\mu_1 = \mu_2$ or $\epsilon = 0$. Under H_0 we have that $\frac{\epsilon}{\sqrt{\sigma_1^2+\sigma_1^2}} \sim \mathcal{N}(0,1)$. To reject H_0 at a certain significance level γ, we have to show that $|\frac{\epsilon}{\sqrt{\sigma_1^2+\sigma_1^2}}| > \Phi^{-1}(1-\frac{\gamma}{2})$. Here Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution and we notate $Z_\gamma = \Phi^{-1}(1-\frac{\gamma}{2})$. In that case, all $|\epsilon| > \sqrt{\sigma_1^2 + \sigma_2^2}Z_\gamma$ will be significant under this test. The minimal significant difference at the γ level is then $\epsilon_\gamma = \sqrt{\sigma_1^2 + \sigma_2^2}Z_\gamma$.

Given our simulated posterior $p^{sim}(\alpha)$ we can compute its mean, $\mu^{sim} = \int_0^1 \alpha p^{sim}(\alpha)\mathrm{d}\alpha$ and variance $\sigma_{sim}^2 = \int_0^1 (\alpha - \mu^{sim})^2 p^{sim}(\alpha)\mathrm{d}\alpha$. We have to make one final assumption: if we estimate α_1 and α_2 under exactly the same conditions, meaning with the same n_ϕ, n_M, $n_{\rho,\eta}$ and the same error-prone metric $M_b^{\rho,\eta}$, and their difference ϵ is relatively small, then their variances should be the same. Using this assumption we compute: $\epsilon_\gamma^{sim} = \sqrt{2\sigma_{sim}^2}Z_\gamma$.

Caveats At this point we will reflect on the several layers of approximations we go through to arrive at an numerical estimate for ϵ_γ. We start out by simulating an experiment where we replace all key observables by their expected values under our experiment assumptions (i.e. the chosen fixed values of α, ρ, η and sample sizes). Of course in a real world setting those values could deviate from their expected values due to bad luck. This will influence both the mean and variance of the resulting estimate. We then compute the simulated posterior using numerical approximation (see next paragraph), which could be imprecise. We then further approximate the posterior by a normal distribution. In practice, we work with large enough sample sizes, that the normal approximation should be relatively accurate.

The overall implication is that the theoretical values of ϵ_γ we use throughout this work provide a useful guideline but it is unclear how exact they are.

Numerical Approximation of Posteriors A problem we face repeatedly is that we are interested in the expected values of a function of a continuous random variable, such as $\int_a^b f(x)p(x)\mathrm{d}x$, which might not have an easily computable closed form. This is for example the case for the integrals over ρ and η in Equation 5, but also when computing the mean and variance of the posterior.

We will now elaborate how we approximate expected values of a continuous variable by middle Riemann sums. Assume we are given a random variable x with domain $(0,1)$, its density function $p(x)$, and its cumulative density function $CDF_x(x') = P(x < x') = \sum_0^{x'} p(x)\mathrm{d}x$. The main idea is to partition the domain into a discrete number equally sized slices. Every partition gets identified by its midpoint and the total total density within that partition. Let N_x be the number of slices, the larger N_x the preciser our approximation

will be. We define:

$$\forall 0 \leq i < N_x$$
$$\boldsymbol{x}[i] = \frac{i + \frac{1}{2}}{N_x}$$
$$\boldsymbol{P}_x[i] = \int_{\frac{i}{N_x}}^{\frac{i+1}{N_x}} p(x)d(x)$$
$$= \int_0^{\frac{i+1}{N_x}} p(x)d(x) - \int_0^{\frac{i}{N_x}} p(x)d(x)$$
$$= CDF_x(\frac{i+1}{N_x}) - CDF_x(\frac{i}{N_x})$$

Here $\boldsymbol{x}[i]$ represents the midpoint of the interval $[\frac{i}{N_x}, \frac{i+1}{N_x}]$ and $\boldsymbol{P}_x[i]$ the total probability mass in that interval. To approximate expected values we can now replace integrals by sums: $\mathbb{E}[f(x)] = \int_0^1 f(x)p(x)\mathrm{d}x \approx \sum_{i=0}^{N_x-1} f(\boldsymbol{x}[i])\boldsymbol{P}_x[i]$.

If we want to apply this discretization to α, ρ, and η, we need access to their cumulative distribution functions. In our framework, these variables are either uniformly or more generally Beta distributed. The cumulative distributions for these are available in most numerical software libraries and therefore computing the discretization is relatively straight-forward.

Applying this discretization to α, ρ, and η we can restate Equation 5 in approximate form:

$$\boldsymbol{P}_{\alpha|N_+=n_+,M_+=m_+}[i] = \boldsymbol{P}_\alpha[i]P(N_+ = N_+|\boldsymbol{\alpha}[i]) \sum_{j=0}^{N_\rho-1}\sum_{k=0}^{N_\eta-1}\left(P(M_+ = m_+|\boldsymbol{\alpha}[i], \boldsymbol{\rho}[j], \boldsymbol{\eta}[k])\boldsymbol{P}_\rho[j]\boldsymbol{P}_\eta[k]\right) \quad (7)$$
$$\forall 0 \leq i < N_\alpha$$

This results in a discretized form of the posterior with the same granularity N_α as for the prior. We can then approximate the mean of the posterior as:

$$\mathbb{E}[\alpha|N_+ = n_+, M_+ = m_+] \approx \sum_{i=0}^{N_\alpha} \boldsymbol{P}_{\alpha|N_+=n_+,M_+=m_+}[i]\boldsymbol{\alpha}[i]$$

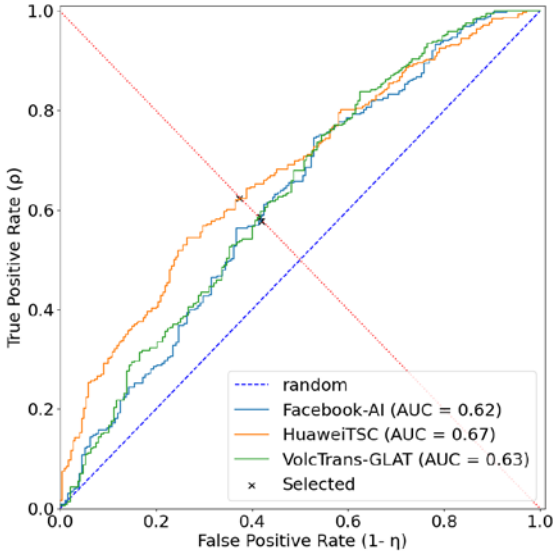


Figure 4: ROC curves for BleuRT predicting MQM annotations for 3 MT system. The markers show the threshold we select in our experiments. The blue diagonal corresponds to a random baseline. The red diagonal visualizes points where $\rho = \eta$.

and the variance as:

$$\mathbb{V}[\alpha|N_+ = n_+, M_+ = m_+] = \mathbb{E}[\alpha^2|N_+ = n_+, M_+ = m_+] - (\mathbb{E}[\alpha|N_+ = n_+, M_+ = m_+])^2 \approx \sum_{i=0}^{N_\alpha-1} \boldsymbol{P}_{\alpha|N_+=n_+,M_+=m_+}[i]\boldsymbol{\alpha}^2[i] - \left(\sum_{i=0}^{N_\alpha-1} \boldsymbol{P}_{\alpha|N_+=n_+,M_+=m_+}[i]\boldsymbol{\alpha}[i]\right)^2$$

We use $N_\alpha = 2000$ and $N_\rho = N_\eta = 1000$ in all our experiments.

C ROC Curves of Metrics

While our theory assumes binary metrics that will only produce 0 or 1 ratings, most real-world automated metrics produce scalar ratings $\in \mathbb{R}$. In our case all metrics under consideration produce scalar ratings. To apply our framework we have to transform scalar ratings into binary ratings. We can do this by selecting a threshold τ that partitions the ratings into binary classes (on either side of the threshold). We define a scalar metric as a function of input output pairs to the reals: $M_s : \mathcal{I} \times \mathcal{O} \to \mathbb{R}$. We interpret the rating as a preference, such that, if $M_s(i, o_1) > M_s(i, o_2)$, then we say that according to M_s, o_1 fits i better than o_2. Given a scalar metric M_s and a threshold $\tau \in \mathbb{R}$ we can derive the

associated binary metric:

$$M_b^\tau(i,o) = \begin{cases} 1 & \text{if } M_s(i,o) > \tau \\ 0 & \text{else} \end{cases} \quad (8)$$

The question is now, how to select τ. This is a well known problem in binary classification. Intuitively, every possible threshold τ is associated with a pair of corresponding ρ and η.

Figure 4 shows the Receiver Operator Characteristic (ROC) curves for BleuRT as a predictor of M_b^* for three machine translation systems. In an ROC plot, the true positive rate is plotted against the false positive rate at various thresholds τ. We note that in our framework the true positive rate is ρ and the false positive rate is $1 - \eta$.

Assume we are given a set of inputs and outputs, the ratings from M_s and ratings from an error-free binary metric M_b^*: $\mathcal{T}_S = \{(i_j, o_j, s_j, r_j^*) | 1 \leq j \leq n_S\}$, where $s_j = M_s(i_j, o_j)$ and $r_j^* = M_b^*(i_j, o_j)$. We can consider the values s_j as candidate thresholds, as these are exactly the cases where the predictions would switch in Equation 8. For each candidate threshold, we can binarize the predictions and compute the associated ρ and η. We select the threshold that minimizes $|\rho - \eta|$, to be consistent with our examples, where we usually assumed for simplicity that $\rho = \eta$. This selection is shown in Figure 4 by markers and the red diagonal.

One thing to note in Figure 4 is that the curves for the three MT systems differ from each other. This means that the specific ρ and η of BleuRT when used as a binary metric depend on the systems that produced a given output. In our framework laid out in Section 3 we assumed that ρ and η are independent of how a given output o is produced. This calls for further analysis in future work.

D Full Show Cases Tables

In this Appendix, we show the full tables for the show cases with all the systems from the WMT and STB setting.

D.1 WMT21

For the WMT task, we have 4 scenarios (see Section 5.1), for all these scenarios we show the pairwise comparisons in Tables 4, 5, 6, and 7. Each table shows for each system the estimated α value in parentheses, and in each cell the ϵ value with $P(\alpha_1 > \alpha_2)$ in parentheses. All the pairs that are significantly distinguishable are put in bold. The WMT systems are Facebook-AI (FBAI) (Tran et al., 2021), VolcTrans-GLAT (VT-G) (Qian et al., 2021), Online-W (OW) [8] (), Nemo (NE), VolcTrans-AT (VT-A) (Qian et al., 2021), UEdin (UE) (Chen et al., 2021), and HuaweiTSC (HU) (Wei et al., 2021). Table 4 where all the human ratings are used shows that FBAI, VT-G, OW, NE, and VT-A are not significantly distinguishable from eachother as their $\epsilon < 0.06$. For the other three scenarios none of the systems are distinguishable. This is consistent with the theoretical predictions. From Table 1, we see that at least 5000 human ratings are needed to be able to significantly distinguish all the pairs of systems (i.e., for $\epsilon < 0.02$). Thus, in this case the problem is that the TG systems are too close to eachother in terms of performance and the automated metrics are too weak to boost the evaluation with low cost.

D.2 STB

Tables 8, 9, and 10 show the full evaluation of the three STB scenarios (see Section 5.2). Each table shows for each system the estimated α value in parentheses, and in each cell the ϵ value with $P(\alpha_1 > \alpha_2)$ in parentheses. All the pairs that are significantly distinguishable are put in bold. For the STB case, the six systems from the original paper are used: Blenderbot (BL) (Roller et al., 2021), Lost in Conversation (LiC) [9], KVMemNN (KV) (Dinan et al., 2020), Huggingface (HF) [10], Bert-Rank (BR) (Deriu et al., 2020), and Seq2Seq-NN (S2S) (Deriu et al., 2020). Note that BR and S2S were custom trained baseline by the STB authors. In the STB case almost all pairs of systems are significantly distinguished, which is in line with the theory and the original STB paper. Our theory reveals that this is mostly due to the fact that the difference in α between the TGs is large and not many samples are needed for discriminating.

E Full Theory Tables

Tables 11, 12, 13, and 14 show the distinguishable ϵ values for different combinations of $|\mathcal{T}_\Phi|$ and $|\mathcal{T}_M|$. Each table has different combinations of ρ, and η values. For each table, we assume that $|\mathcal{T}_\Phi| = |\mathcal{T}_{\rho,\eta}|$. These tables can be used as guidelines for deciding on the number of human and

[8] Anonymous System

[9] https://github.com/atselousov/transformer_chatbot

[10] https://github.com/huggingface/transfer-learning-conv-ai

	FBAI (0.67)	VT-G (0.64)	OW (0.64)	NE (0.64)	VT-A (0.61)	UE (0.59)	HU (0.58)
FBAI (0.67)	-	0.02 (0.798)	0.03 (0.848)	0.03 (0.862)	0.05 (0.968)	**0.08 (0.997)**	**0.09 (0.998)**
VT-G (0.64)	-0.02 (0.197)	-	0.01 (0.573)	0.01 (0.598)	0.03 (0.844)	0.06 (0.971)	**0.06 (0.978)**
OW (0.64)	-0.03 (0.148)	-0.01 (0.420)	-	0.00 (0.522)	0.02 (0.794)	0.05 (0.955)	0.05 (0.966)
NE (0.64)	-0.03 (0.134)	-0.01 (0.395)	-0.00 (0.471)	-	0.02 (0.775)	0.05 (0.949)	0.05 (0.961)
VT-A (0.61)	-0.05 (0.031)	-0.03 (0.152)	-0.02 (0.202)	-0.02 (0.220)	-	0.03 (0.808)	0.03 (0.840)
UE (0.59)	**-0.08 (0.003)**	-0.06 (0.028)	-0.05 (0.043)	-0.05 (0.049)	-0.03 (0.187)	-	0.00 (0.546)
HU (0.58)	**-0.09 (0.002)**	**-0.06 (0.021)**	-0.05 (0.033)	-0.05 (0.038)	-0.03 (0.156)	-0.00 (0.447)	-

Table 4: Full WMT scenario with $|\mathcal{T}_\Phi| = 527$, and $|\mathcal{T}_M| = 0$

	FBAI (0.67)	VT-G (0.65)	NE (0.64)	OW (0.64)	VT-A (0.61)	UE (0.59)	HU (0.58)
FBAI (0.67)	-	0.02 (0.615)	0.03 (0.670)	0.03 (0.670)	0.06 (0.809)	0.08 (0.877)	0.09 (0.904)
VT-G (0.65)	-0.02 (0.382)	-	0.01 (0.557)	0.01 (0.557)	0.04 (0.719)	0.06 (0.807)	0.07 (0.843)
NE (0.64)	-0.03 (0.327)	-0.01 (0.440)	-	0.00 (0.499)	0.03 (0.667)	0.05 (0.764)	0.06 (0.805)
OW (0.64)	-0.03 (0.327)	-0.01 (0.440)	0.00 (0.499)	-	0.03 (0.667)	0.05 (0.764)	0.06 (0.805)
VT-A (0.61)	-0.06 (0.189)	-0.04 (0.279)	-0.03 (0.330)	-0.03 (0.330)	-	0.02 (0.612)	0.03 (0.665)
UEdin (0.59)	-0.08 (0.121)	-0.06 (0.191)	-0.05 (0.234)	-0.05 (0.234)	-0.02 (0.386)	-	0.01 (0.555)
HU (0.58)	-0.09 (0.095)	-0.07 (0.155)	-0.06 (0.193)	-0.06 (0.193)	-0.03 (0.332)	-0.01 (0.442)	-

Table 5: Full WMT scenario with $|\mathcal{T}_\Phi| = 100$, and $|\mathcal{T}_M| = 0$

	FBAI (0.67)	VT-G (0.64)	NE (0.63)	OW (0.63)	VT-A (0.61)	UE (0.58)	HU (0.57)
FBAI (0.67)	-	0.02 (0.631)	0.03 (0.689)	0.04 (0.713)	0.06 (0.817)	0.08 (0.896)	0.09 (0.918)
VT-G (0.64)	-0.02 (0.366)	-	0.01 (0.560)	0.01 (0.588)	0.04 (0.713)	0.06 (0.821)	0.07 (0.854)
NE (0.63)	-0.03 (0.309)	-0.01 (0.437)	-	0.00 (0.527)	0.03 (0.658)	0.05 (0.779)	0.06 (0.817)
OW (0.63)	-0.04 (0.284)	-0.01 (0.409)	-0.00 (0.470)	-	0.02 (0.631)	0.05 (0.757)	0.06 (0.798)
VT-A (0.61)	-0.06 (0.181)	-0.04 (0.284)	-0.03 (0.339)	-0.02 (0.366)	-	0.02 (0.643)	0.03 (0.694)
UE (0.58)	-0.08 (0.103)	-0.06 (0.177)	-0.05 (0.219)	-0.05 (0.240)	-0.02 (0.354)	-	0.01 (0.555)
HU (0.57)	-0.09 (0.081)	-0.07 (0.144)	-0.06 (0.181)	-0.06 (0.200)	-0.03 (0.303)	-0.01 (0.442)	-

Table 6: Full WMT scenario with $|\mathcal{T}_\Phi| = 100$, and $|\mathcal{T}_M| = 1000$

	FBAI (0.66)	VT-G (0.62)	NE (0.61)	OW (0.61)	VT-A (0.58)	UE (0.57)	HU (0.56)
FBAI (0.66)	-	0.04 (0.742)	0.05 (0.787)	0.05 (0.779)	0.07 (0.888)	0.09 (0.916)	0.09 (0.933)
VT-G (0.62)	-0.04 (0.256)	-	0.01 (0.538)	0.01 (0.548)	0.03 (0.705)	0.05 (0.768)	0.05 (0.801)
NE (0.61)	-0.05 (0.211)	-0.01 (0.458)	-	0.00 (0.512)	0.03 (0.684)	0.04 (0.753)	0.05 (0.790)
OW (0.61)	-0.05 (0.219)	-0.01 (0.449)	-0.00 (0.485)	-	0.02 (0.658)	0.04 (0.727)	0.04 (0.762)
VT-A (0.58)	-0.07 (0.111)	-0.03 (0.292)	-0.03 (0.313)	-0.02 (0.339)	-	0.01 (0.587)	0.02 (0.629)
UE (0.57)	-0.09 (0.083)	-0.05 (0.230)	-0.04 (0.244)	-0.04 (0.271)	-0.01 (0.410)	-	0.01 (0.539)
HU (0.56)	-0.09 (0.066)	-0.05 (0.197)	-0.05 (0.208)	-0.04 (0.236)	-0.02 (0.368)	-0.01 (0.458)	-

Table 7: Full WMT scenario with $|\mathcal{T}_\Phi| = 100$, and $|\mathcal{T}_M| = 1000$

	BL (0.38)	LiC (0.30)	KV (0.24)	HF (0.18)	BR (0.07)	S2S (0.04)
BL (0.38)	-	**0.08 (0.999)**	**0.13 (1.000)**	**0.20 (1.000)**	**0.31 (1.000)**	**0.34 (1.000)**
LiC (0.30)	**-0.08 (0.001)**	-	**0.06 (0.989)**	**0.12 (1.000)**	**0.23 (1.000)**	**0.26 (1.000)**
KV (0.24)	**-0.13 (0.000)**	**-0.06 (0.010)**	-	**0.07 (0.998)**	**0.18 (1.000)**	**0.20 (1.000)**
HF (0.18)	**-0.20 (0.000)**	**-0.12 (0.000)**	**-0.07 (0.002)**	-	**0.11 (1.000)**	**0.14 (1.000)**
BR (0.07)	**-0.31 (0.000)**	**-0.23 (0.000)**	**-0.18 (0.000)**	**-0.11 (0.000)**	-	0.02 (0.974)
S2S (0.04)	**-0.34 (0.000)**	**-0.26 (0.000)**	**-0.20 (0.000)**	**-0.14 (0.000)**	-0.02 (0.024)	-

Table 8: Full STB scenario with $|\mathcal{T}_\Phi| = 600$, and $|\mathcal{T}_M| = 0$

	BL (0.38)	LiC (0.30)	KV (0.25)	HF (0.19)	BR (0.07)	S2S (0.05)
BL (0.38)	-	0.08 (0.882)	**0.14 (0.983)**	**0.20 (0.999)**	**0.31 (1.000)**	**0.33 (1.000)**
LiC (0.30)	-0.08 (0.117)	-	0.06 (0.827)	**0.12 (0.976)**	**0.24 (1.000)**	**0.25 (1.000)**
KV (0.25)	**-0.14 (0.016)**	-0.06 (0.170)	-	0.06 (0.848)	**0.18 (1.000)**	**0.20 (1.000)**
HF (0.19)	**-0.20 (0.001)**	**-0.12 (0.024)**	-0.06 (0.150)	-	**0.12 (0.995)**	**0.14 (0.999)**
BR (0.07)	**-0.31 (0.000)**	**-0.24 (0.000)**	**-0.18 (0.000)**	**-0.12 (0.005)**	-	0.02 (0.729)
S2S (0.05)	**-0.33 (0.000)**	**-0.25 (0.000)**	**-0.20 (0.000)**	**-0.14 (0.001)**	-0.02 (0.266)	-

Table 9: Full STB scenario with $|\mathcal{T}_\Phi| = 100$, and $|\mathcal{T}_M| = 0$

	BL (0.36)	LiC (0.28)	KV (0.22)	HF (0.15)	BR (0.06)	S2S (0.05)
BL (0.36)	-	0.08 (0.889)	**0.13 (0.989)**	**0.21 (1.000)**	**0.30 (1.000)**	**0.31 (1.000)**
LiC (0.28)	-0.08 (0.109)	-	0.06 (0.851)	**0.13 (0.994)**	**0.22 (1.000)**	**0.23 (1.000)**
KV (0.22)	**-0.13 (0.010)**	-0.06 (0.147)	-	0.07 (0.935)	**0.16 (1.000)**	**0.17 (1.000)**
HF (0.15)	**-0.21 (0.000)**	**-0.13 (0.006)**	-0.07 (0.064)	-	**0.09 (0.994)**	**0.10 (0.997)**
BR (0.06)	**-0.30 (0.000)**	**-0.22 (0.000)**	**-0.16 (0.000)**	**-0.09 (0.006)**	-	0.01 (0.628)
S2S (0.05)	**-0.31 (0.000)**	**-0.23 (0.000)**	**-0.17 (0.000)**	**-0.10 (0.003)**	-0.01 (0.365)	-

Table 10: Full STB scenario with $|\mathcal{T}_\Phi| = 100$, and $|\mathcal{T}_M| = 10000$

automated ratings needed for different automated metric performances.

		$\lvert\mathcal{T}_{\mathcal{M}}\rvert$						
		0	1000	2500	5000	10000	50000	100000
$\lvert\mathcal{T}_{\rho,\eta}\rvert = \lvert\mathcal{T}_{\mathcal{O}}\rvert$	0	1.000	0.734	0.733	0.733	0.733	0.733	0.733
	100	0.134	0.124	0.124	0.123	0.123	0.123	0.123
	250	0.085	0.080	0.079	0.079	0.079	0.079	0.079
	500	0.061	0.057	0.057	0.056	0.056	0.056	0.056
	1000	0.043	0.041	0.040	0.040	0.040	0.040	0.039
	2500	0.027	0.027	0.026	0.026	0.025	0.025	0.025
	5000	0.019	0.019	0.019	0.018	0.018	0.018	0.018
	10000	0.014	0.013	0.013	0.013	0.013	0.013	0.013

Table 11: Estimated ϵ_γ for $\alpha = 0.60$, $\rho = 0.70$, $\eta = 0.70$, and $\gamma = 0.05$

		$\lvert\mathcal{T}_{\mathcal{M}}\rvert$						
		0	1000	2500	5000	10000	50000	100000
$\lvert\mathcal{T}_{\rho,\eta}\rvert = \lvert\mathcal{T}_{\mathcal{O}}\rvert$	0	1.000	0.739	0.738	0.738	0.738	0.738	0.738
	100	0.134	0.091	0.088	0.087	0.086	0.086	0.086
	250	0.085	0.061	0.057	0.055	0.054	0.053	0.053
	500	0.061	0.046	0.042	0.040	0.039	0.038	0.037
	1000	0.043	0.036	0.032	0.030	0.028	0.027	0.026
	2500	0.027	0.025	0.022	0.021	0.019	0.017	0.017
	5000	0.019	0.018	0.017	0.016	0.015	0.013	0.012
	10000	0.014	0.013	0.013	0.012	0.011	0.009	0.009

Table 12: Estimated ϵ_γ for $\alpha = 0.60$, $\rho = 0.90$, $\eta = 0.90$, and $\gamma = 0.05$

		$\lvert\mathcal{T}_{\mathcal{M}}\rvert$						
		0	1000	2500	5000	10000	50000	100000
$\lvert\mathcal{T}_{\rho,\eta}\rvert = \lvert\mathcal{T}_{\mathcal{O}}\rvert$	0	1.000	0.742	0.742	0.742	0.742	0.742	0.742
	100	0.134	0.059	0.051	0.048	0.046	0.045	0.045
	250	0.085	0.044	0.035	0.030	0.027	0.025	0.024
	500	0.061	0.037	0.028	0.023	0.019	0.016	0.015
	1000	0.043	0.031	0.024	0.020	0.016	0.011	0.010
	2500	0.027	0.023	0.020	0.016	0.013	0.008	0.007
	5000	0.019	0.018	0.016	0.014	0.012	0.007	0.006
	10000	0.014	0.013	0.012	0.011	0.010	0.006	0.005

Table 13: Estimated ϵ_γ for $\alpha = 0.60$, $\rho = 0.99$, $\eta = 0.99$, and $\gamma = 0.05$

		$\lvert\mathcal{T}_{\mathcal{M}}\rvert$						
		0	1000	2500	5000	10000	50000	100000
$\lvert\mathcal{T}_{\rho,\eta}\rvert = \lvert\mathcal{T}_{\mathcal{O}}\rvert$	0	1.000	0.732	0.732	0.732	0.732	0.732	0.732
	100	0.134	0.133	0.133	0.133	0.133	0.133	0.133
	250	0.085	0.085	0.085	0.085	0.085	0.085	0.085
	500	0.061	0.061	0.060	0.060	0.060	0.060	0.060
	1000	0.043	0.043	0.043	0.043	0.043	0.043	0.043
	2500	0.027	0.027	0.027	0.027	0.027	0.027	0.027
	5000	0.019	0.019	0.019	0.019	0.019	0.019	0.019
	10000	0.014	0.014	0.014	0.014	0.014	0.014	0.014

Table 14: Estimated ϵ_γ for $\alpha = 0.60$, $\rho = 0.51$, $\eta = 0.51$, and $\gamma = 0.05$

Improving NL-to-Query Systems through Re-ranking of Semantic Hypothesis

Pius von Däniken[1] and **Jan Deriu**[1] and **Eneko Agirre**[2] and **Ursin Brunner**[1] and **Mark Cieliebak**[1] and **Kurt Stockinger**[1]
[1] ZHAW Zurich University of Applied Sciences
{vode,deri,brrn,ciel,stog}@zhaw.ch
[2] HiTZ Center - Ixa, University of the Basque Country UPV/EHU
e.agirre@ehu.eus

Abstract

Natural Language-to-Query systems translate a natural language question into a formal query language such as SQL. Typically the translation results in a set of candidate query statements due to the ambiguity of natural language. Hence, an important aspect of NL-to-Query systems is to rank the query statements so that the most relevant query is ranked on top. We propose a novel approach to significantly *improve the query ranking and thus the accuracy of such systems*. First, we use existing methods to translate the natural language question (NL_{in}) into k query statements and rank them. Then we translate each of the k query statements back into a natural language question (NL_{gen}) and use the *semantic similarity* between the original question NL_{in} and each of the k generated questions NL_{gen} to *re-rank the output*. Our experiments on two standard datasets, OTTA and Spider, show that this technique improves even strong state-of-the-art NL-to-Query systems by up to 9 percentage points. A detailed error analysis shows that our method correctly down-ranks queries with missing relations and wrong query types. While this work is focused on NL-to-Query, our method could be applied to any other semantic parsing problems as long as a text generation method is available.

1 Introduction

NL-to-Query describes the task of translating natural language questions to meaningful representations, such as logical forms, executable code, or structured query languages like SQL. The application of neural networks and the introduction of larger datasets (Yin and Neubig, 2017; Yu et al., 2018; Brunner and Stockinger, 2021) has increased performance, but the task is far from solved.

Re-ranking of candidate query statements allows introducing additional information in the process (Yin and Neubig, 2019). For a given natural language question (NL_{in}), neural networks keep a beam search and produce k candidate query statements (QS). Our analysis shows that an oracle selecting the correct query among the top-scoring 15 candidates would improve the performance of publicly available systems by up to 10 accuracy points on the Spider benchmark (Yu et al., 2018).

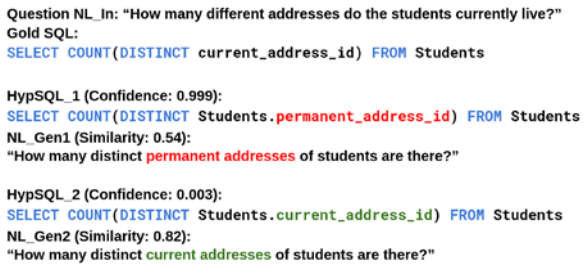


Figure 1: Example illustrates how semantic similarity is used to extract the correct hypothesis. `NL_In` is the input question, `Gold SQL` is the gold SQL query, `HypSQL_1` and `HypSQL_2` are generated by an NL-to-Query system (with confidence scores), and `NL_Gen1` and `NL_Gen2` are back-translated from the `HypSQL` statements, with scores by a similarity system. See text for further details.

Inspired by the success of back-translation in machine translation (Sennrich et al., 2016), we propose to *re-rank the candidate queries* according to the semantic similarity between the original question NL_{in} and the k synthetic questions NL_{gen} obtained via back-translating each of the k candidate queries into natural language. Figure 2 depicts the pipeline of our proposed system.

Figure 1 shows an example from the Spider dataset. For the question "How many different addresses do the students currently live?". The highest-ranked query according to the beam search ranking is HypSQL_1 with a confidence score of 0.999. However, this query returns the *permanent* addresses, which does not refer to the correct attribute, which would be the *current* addresses. In the example, the second hypothesis (i.e., HypSQL_2) has a much lower confidence of 0.003

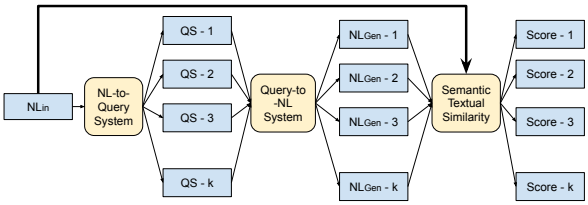


Figure 2: Pipeline of our system. NL_{in} = original natural language, QS = query statement, NL_{gen} = generated natural language.

although it fits the input question perfectly. On the other hand, the *semantic similarity* score between NL_{in} and the generated questions NL_{gen} shows a different picture: The back-translation of the correct hypothesis, i.e., NL_{gen2}, has a higher semantic similarity (0.82) than the back-translation of the incorrect hypothesis (0.54). Hence, semantic similarity would help to identify the correct query. This paper makes the following contributions:

- We present a novel method to improve NL-to-Query systems using re-ranking according to *Query-to-NL back-translation and semantic similarity*.
- We showcase improvements in two datasets using three systems, around $5-9$ points in OTTA (Deriu et al., 2020) and $2-3$ points in Spider (Yu et al., 2018).
- The error analysis shows that our method down-ranks hypotheses with missing relations or with incorrect query types.

2 Related Work

NL-to-Query (also referred to as Natural Language to Databases NLIDB) describes the task of translating natural language questions into structured queries (e.g., SQL). Most current approaches are based on sequence-to-sequence architectures (Yin and Neubig, 2017; Dong and Lapata, 2018; Suhr et al., 2018; Deriu et al., 2020), where the encoder is a recurrent neural network that generates a hidden representation of the natural language question, and the decoder is a recurrent neural network that generates the query. Alternatively, some approaches combine symbolic reasoning with information retrieval techniques (Sen et al., 2020). For a more in-depth treatment, we refer the reader to Affolter et al. (2019) and Odzcan et al. (2020).

In this work, we focus on the translation from natural language questions to database queries, where most recent approaches were proposed in the context of the text-to-SQL Spider dataset (Yu et al., 2018)[1]. Instead of working directly on SQL, some authors propose to use simpler and more general abstract syntax trees. For instance, Deriu et al. (2020) propose to use so-called Operation Trees, which we also used for this work.

Hypothesis Re-ranking is the task of creating an alternative ranking of k candidate solutions for a given task. The k candidates are usually the output of a beam search. In our case, the candidates are queries for the given natural language question. However, the problem of hypothesis re-ranking arises in many different generation tasks, not only NL-to-Query. For instance, Dušek and Jurcicek (2016) train a re-ranking network to score the generated hypotheses of their natural language generation model. Alternatively, (Deriu and Cieliebak, 2018; Agarwal et al., 2018) trained classifiers to predict the correctness of the hypotheses produced by their natural language generation system and select the hypothesis with the highest correctness score. Most of these approaches are developed in the field of natural language generation from structured data. For code generation, Yin and Neubig (2019) perform re-ranking by reconstructing the original utterance for the generated code. They use the reconstruction error as a measure for re-ranking. We are not aware of prior research on using textual semantic similarity to re-rank hypotheses in the field of NL-to-Query or Semantic Parsing in general.

Semantic Textual Similarity assesses to what degree two chunks of text are similar, usually on a 0-5 scale, which ranges from unrelated (0) to semantically equivalent (5) (Agirre et al., 2013). The advent of transformer-based models such as RoBERTa (Liu et al., 2019) has improved automatically assessing semantic textual similarity. Recently (Kane et al., 2020) introduced *NUBIA (NeUral Based Interchangeability Assessor for Text Generation)*. It extracts features from RoBERTa and GPT-2 (Radford et al., 2019) and fine-tunes a fully connected neural network to output a score between 0 and 1, indicating how interchangeable two input sentences are. Throughout this work, we will use *NUBIA* to automatically score the similarity between a natural question (NL_{in}) and a back-translated question (NL_{gen}).

Query-to-NL has the goal of translating a struc-

[1] https://yale-lily.github.io/spider

tured query into natural language and to provide a lay user with an explanation of the meaning of the query. A simple approach is to define production rules applied to the nodes of the abstract syntax tree (AST) of the query. Systems based on this idea have been developed for SQL (Koutrika et al., 2010), SPARQL (Ngonga Ngomo et al., 2013), Operation Trees (von Däniken, 2021), and queries expressed in lambda calculus (Wang et al., 2015). There are also systems based on neural networks such as (Xu et al., 2018). In this work, we leverage one of those systems to post-process the output of an NL-to-Query system. Others have also used query explanations to incorporate corrective feedback from the user in the NL-to-Query workflow (Elgohary et al., 2020; Labutov et al., 2018; Yao et al., 2019, 2020).

3 Method: Similarity for Re-ranking

The proposed method works in three steps (see also Figure 2): first, the NL-to-Query system translates the natural language input NL_{in} into a set of k candidate query statements QS - called our hypotheses. This is achieved by applying beam search during the decoding stage of a recurrent neural network. In the second step, each of the k hypotheses QS is translated back into natural language NL_{gen} using a Query-to-NL system. In the last step, each of the k back-translations NL_{gen} is compared to the original input using an off-the-shelf semantic textual similarity algorithm. We use the semantic similarity score to rank the hypotheses. For each NL_{in}, the top-scoring hypothesis is returned as the answer of the system.

3.1 Ranking Hypotheses based on Semantic Textual Similarity

Let NL_{in} be the user input (i.e., the natural language question) and $H = \{QS_1, ..., QS_k\}$ be the set of k hypotheses, i.e. candidate query statements QS_i, that are the output of the NL-to-Query system. In most cases, this set is the result of applying beam search for decoding. However, other approaches result in a set of hypotheses, for instance an ensemble of different NL-to-Query systems. In this work, we focus only on beam search-based hypothesis sets. Thus, each of the hypotheses has a confidence score c_i, which is used to rank the set of hypotheses, i.e., the candidate queries. We refer to this ranking as *Confidence*.

In a second step, each of the hypotheses QS_i is back-translated into a natural language question NL^i_{gen} using a Query-to-NL engine. Thus, we end up with a set of back-translated hypotheses $H_Q = \{NL^1_{gen}, ..., NL^k_{gen}\}$.

In a third step, we compute for each back-translated hypothesis the semantic textual similarity score with the user input NL_{in}, i.e., $s_i = \mathit{SemSim}(NL_{in}, NL^i_{gen})$. The set of hypotheses can be ranked according to the semantic similarity scores. We refer to this ranking as *Semantic*.

3.2 Weighting Strategies

Since the two rankings, *Confidence* and *Semantic* may disagree on the top hypothesis in some cases (as we have shown in the example in Figure 1), we combine the two scores c_i and s_i into a new ranking. For this, we propose the following weighting strategies:

Equal Weighting. The naive strategy is based on simply multiplying the two scores, that is $m_i^{equal} = c_i * s_i$, and we rank the set of hypotheses according to m_i^{equal}. We refer to this ranking as *Equal Weighting*.

Calibrated Weighting. Since the confidence scores and the semantic similarity scores have different distributions, the influence of each score in the *Equal Weighting* is not equal. For instance, in some cases, the influence of c_i is stronger than s_i and vice-versa. To counteract this effect, we decided to *calibrate both scores before multiplying*. A calibrated score should reflect the proportion of correct hypotheses selected, e.g., when a calibrated system assigns a score of 0.8 to a hypothesis, this hypothesis will be correct in 80% of the cases.

We use *Platt Scaling* (Platt, 2000) to calibrate both scores. This works by training a logistic regression model on the outputs of a model to transform these outputs into probability distributions. More precisely, for the confidence scores and the semantic scores respectively, a logistic regression model is trained. For this, we have to set aside a few hypotheses (more details later on). For the confidence calibration, a logistic regression model is trained on a set of pairs of confidence score and a label that indicates if the query is correct, i.e., $\mathbf{D} = \{(c_i, I^i_{\text{corr}}\}$. Analogously, we train a logistic regression model for the semantic similarity score s_i. Thus, the calibrated scores can be interpreted as the probability of the query being correct, i.e., $c_i^{calib} = Pr(I^i_{\text{corr}} = 1|c_i)$ and $s_i^{calib} = Pr(I^i_{\text{corr}} = 1|s_i)$. We call the score after

calibration c_i^{calib} and s_i^{calib} and the resulting mixed score $m_i^{calib} = c_i^{calib} * s_i^{calib}$. The resulting ranking is called *Calibrated Weighting*.

Learned Weighting. A natural extension of the calibration idea is to *train a logistic regression on both scores* at the same time, instead of independently. That is, we train a logistic regression model on pairs of confidence and semantic-scores[2], i.e., $\mathbf{D} = \{((c_i, s_i), I^i_{\text{corr}}\}$. This way, the model can learn the mixed proportions directly. Thus, $m_i^{learned} = Pr(I^i_{\text{corr}} = 1|c_i, s_i)$. For this, we again have to set aside a few hypotheses. We use the predicted probabilities from the logistic regression model to rank hypotheses and call the resulting ranking *Learned Weighting*.

Threshold Weighting. We observed that the confidence scores c_i are high in most cases in which the *Confidence* ranking yields a correct query. In many cases where the *Confidence* ranking yields wrong queries, the confidence scores are low. However, the *Semantic* scores tend to be higher. Thus, we propose the following strategy: If the maximum confidence score of the hypotheses set is above a threshold, we use the *Confidence* ranking, otherwise, we use the *Semantic* ranking. We refer to this ranking as *Threshold Weighting*. The threshold is calculated by first determining the 90th percentile over the confidence scores of all training hypotheses and then finding the lowest confidence of a correct hypothesis that lies above that.

Upper Bounds. To determine the theoretical upper bounds of our approach, we introduce two oracles. The first oracle selects the correct hypothesis from the candidates if there is one. The second oracle selects the correct hypothesis between the two top-ranked hypotheses by *Confidence* or *Semantic* if there is one. The first oracle determines the potential of re-ranking in general (we refer to it as *Oracle*). The second oracle determines the maximum contribution that the semantic similarity could do to Confidence (we refer to it as *Oracle-Sem*).

4 Experimental Setup

In this section, we describe the experimental setup, the datasets, the NL-to-Query models, the Query-to-NL model, and the semantic textual similarity model.

[2]Using more features, e.g., the length of the generated query or m_i^{equal} did not yield any improvements.

4.1 Datasets

We analyzed our approach on two different datasets used as benchmarks for evaluating NL-to-Query systems: Spider (Yu et al., 2018) and OTTA (Deriu et al., 2020). Both datasets contain complex queries and cover large amounts of attributes of the databases. Spider contains around 10K queries against 200 different databases. The dataset is used to study NL-to-SQL translations. OTTA contains around 3.8K queries over 5 databases. OTTA is used to study translations from NL-to-OT (Operation Trees) which are similar to abstract syntax trees (AST), i.e., an intermediate query language can be translated to other query languages such as SQL or SPARQL. OTTA contains more complex queries with longer join paths than Spider. From the OTTA corpus, we used only queries against the databases *Moviedata* and *Chinook* since they contain the largest amounts of queries. Details about the queries used for each dataset are given below.

4.2 NL-to-Query Models

We applied publicly available machine learning models trained for the datasets, which produce queries with filter values in the WHERE-clauses as otherwise there would be placeholder tokens in the back-translations. For all models, we use a beam size[3] of $k = 15$. For the OTTA corpus, we used the pre-trained *GrammarNet* by (Deriu et al., 2020). The output of *GrammarNet* is a set of Operation Tree (OT) hypotheses, which represent the query. OTs can be translated to SQL and executed on an SQL database. For each of the two domains in OTTA (i.e., *Moviedata* and *Chinook*), we use a specifically trained *GrammarNet*. We refer to these models as *GrammarNet-Moviedata* and *GrammarNet-Chinook*. For the Spider dataset, we apply two strong NL-to-SQL systems that are publicly available. The first system is *BridgeV2* (Lin et al., 2020), which returns a set of hypothesis SQL queries from a beam search decoder. We refer to this model as *Spider-BridgeV2* [4]. The second system is *ValueNet* by Brunner and Stockinger (2021), which also returns a set of SQL hypotheses from a beam search decoder[5]. We refer to this model as

[3]In preliminary experiments, we noted that using a larger beam size does not impact the scores significantly.

[4]We chose these systems for their strong performance, code availability and quality of code.

[5]The API provided by the authors included confidence scores based on the sum per-token-confidence instead of average. We approximated the average by dividing the provided score by the number of characters in the SQL hypothesis.

Spider-ValueNet.

4.3 Query-to-NL Model

For back-translating queries to natural language, we use the *Operation Tree-to-Text (OT3)* system kindly made available by von Däniken (2021). It translates OTs into natural language questions in a rule-based manner, which ensures that most OTs are translated correctly, i.e., no nodes are left out or added during translation. OT3 is domain-agnostic, which allows it to be easily adapted to a new domain by just defining domain-specific metadata, i.e., the canonical names of the tables, attributes and types. The main advantage of OT3 is the ability to express relationships naturally, which results in more fluent back-translations. There are currently some limitations with the state-of-the-art Query-to-NL models, which do not handle more complex constructs [6] well. Thus, we perform the evaluation only on the queries that are handled by OT3. More details can be found in Appendix A.

4.4 Semantic Textual Similarity Model

In order to compute textual semantic similarity between two questions, i.e., between NL_{in} and NL_{gen}, we apply NUBIA (Kane et al., 2020), a pre-trained model that scores a pair of sentences based on their interchangeability. We use NUBIA[7] out-of-the-box without any fine-tuning.

4.5 Mixed Strategies

For the *Calibrated Weighting*, the *Learned Weighting*, and the *Threshold Weighting* rankings, labeled data points are needed for setting up the mixed strategy. The samples are used to train the logistic regression models for the *Calibrated Weighting* and the *Learned Weighting*. We use the implementation provided by *scikit-learn* (Pedregosa et al., 2011) with balanced class weights and all other parameters as default. For the *Threshold Weighting*, these samples are used to determine the threshold for when to select the *Confidence* ranking or the *Semantic* ranking. We use k-fold cross-validation with k[8] chosen such that there are 20 samples in each fold[9] for training the strategies for each split. We report accuracies averaged over the k test splits for all strategies.

[6] E.g., GroupBy, SetOperations, or Nested Quieries

[7] https://github.com/wl-research/nubia

[8] Concretely, $k = 22$ for Spider, $k = 11$ for Moviedata, and $k = 12$ for Chinook.

[9] This results in $20 * 15 = 300$ data points for training the logistic regression models.

5 Results

As explained in the previous section, we evaluate the effectiveness of our approach over two different datasets consisting of 22 databases using three different systems, as shown in Table 1. We evaluate the systems using the *component equality* proposed by Yu et al. (2018). We can see that for all datasets one of our *re-ranking approaches outperforms the baseline* without re-ranking up to 9%. We will now analyze our re-ranking approaches in more detail.

Semantic Re-ranking. In all cases, except for *Chinook*, the *Semantic-based* re-ranking performs worse than the baseline system ranking (Confidence), showing that our method alone has not enough information to select the correct hypothesis.

Mixed Re-ranking (i.e. Equal, Calibrated, Learned, Threshold). On the contrary, the combination of the Confidence and Similarity scores improves over *Confidence* alone in all mixed strategies (with a minor exception for Threshold for ValueNet in Spider). The improvement ranges from $2-3\%$ on *Spider* to $5-9\%$ on *OTTA*. These results show that our method injects new information and improves over the base systems. In all cases, the simple *Equal Weighting* performs well, making it a great default mixed strategy. The results or other mixed strategies are better in some cases, although the best mixed strategy varies in each column. For instance, for *Spider-Bridge* the *Threshold Weighting* strategy works best, yielding an improvement of 2.56 points in accuracy.

Oracle. The difference between *Confidence* and *Oracle*, i.e. the optimal re-ranking, lies at around 18% for both *OTTA* subcorpus and $8-10\%$ for Spider, depending on the system. The differences in margins between *Spider* and *OTTA* can be explained by the fact that the *Spider*-based models achieve higher *Confidence* accuracies, which decreases the margin for improvement.

Oracle-Sem. The difference between the best mixed strategy and *Oracle-Sem* is around 3 points. Thus, there is a potential improvement of around 3 points left for all systems using semantic similarity. However, the difference between the *Oracle-Sem* score and the *Oracle* score differs between the *Spider*-based systems and the *OTTA*-based systems. While the difference in the *Spider*-based systems is between 3 to 4 points, the difference for the *OTTA*-based systems is between 6 to 7 points.

Dataset System	OTTA-Chin. GrammarNet	OTTA-Movie GrammarNet	Spider Bridge	Spider ValueNet
Confidence	42.89	52.24	71.46	74.31
Semantic	48.16 (+5.27)	45.25 (−6.99)	62.70 (−8.76)	68.01 (−6.30)
Equal	**51.84** (+8.95)	59.23 (+6.99)	73.03 (+1.57)	76.83 (+2.52)
Calibrated	51.44 (+8.55)	**59.60** (+7.36)	73.78 (+2.32)	**77.22** (+2.91)
Learned	51.48 (+8.59)	59.44 (+7.20)	73.93 (+2.47)	77.09 (+2.78)
Threshold	46.90 (+4.01)	54.71 (+2.47)	**74.05** (+2.59)	71.30 (−3.01)
Oracle-Sem	54.94 (+12.05)	62.38 (+10.14)	77.30 (+5.84)	80.35 (+6.04)
Oracle	61.32 (+18.43)	69.98 (+17.74)	81.12 (+9.66)	83.12 (+8.81)

Table 1: Accuracy of our approach for translating NL questions to OTs and SQL, respectively, using three different systems and two different datasets. The deltas with respect to the *Confidence* ranking (baseline) are shown in parentheses. *Oracle-Sem* and *Oracle* are theoretical upper bounds.

6 Discussion

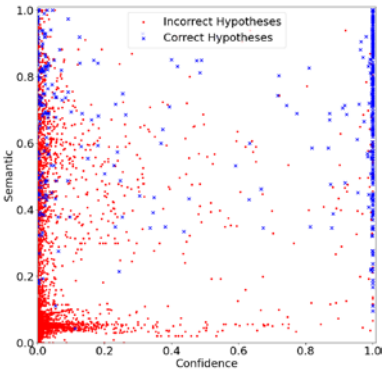
Based on the results, we see that including semantic similarity for re-ranking works better than using the *Confidence* scoring only. In this section, we explore the potential and limitations of this approach in more detail.

6.1 Confidence Score vs. Semantic Similarity Score

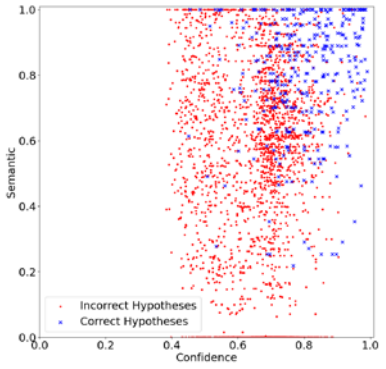
To better understand the results, we analyze the relationship between the confidence scores and the semantic scores. In Figure 3, the confidence scores are plotted against the semantic similarity scores, where blue dots denote correct hypotheses, and red dots denote incorrect ones. We perform the analysis on the *Bridge* system over Spider and the GrammarNet system over *Moviedata*, as they show the clearest difference in score distributions.

First, we note that the distributions for the two systems look different. For *Bridge* the confidence scores mostly lie at the edges, either at 0.0 or 1.0. The *Moviedata* confidence scores are more evenly distributed between 0.4 and 1.0. On the other hand, the semantic similarity scores are evenly distributed in both cases.

Second, we note that for the *Bridge* system, confidence scores close to 1.0 are reliable, i.e., a hypothesis with confidence close to 1.0 tends to be correct. On the other hand, correct hypotheses with low confidence tend to have higher semantic scores (see upper left corner). This explains the strong performance of *Threshold Weighting* for *Bridge*. For *Moviedata*, the picture is different. The correct samples tend to have both high confidence and high semantic scores (upper right corner). Thus, the other weighing strategies tend to perform well, while *Threshold Weighting* under-performs.



(a) Spider-BridgeV2



(b) Moviedata-GrammarNet

Figure 3: Confidence scores and semantic similarity scores for hypotheses produced by *Spider-BridgeV2* and *Moviedata-GrammarNet*. Every cross corresponds to a hypothesis. Blue indicates correct hypotheses and red incorrect ones.

Third, we note that semantic scoring alone is not sufficient. For *Bridge*, the semantic score tends to score correct hypotheses as low as the incorrect ones (see lower part). However, it works well for finding incorrect hypotheses. Although the distributions for *Bridge* and *Moviedata* have great differences, our approach works in both cases.

Error Type	Missing Join			
Original Question	List all singer names in concerts in year 2014.			
Ranking	**SQL**	**Back-translated Question**	c_i	s_i
Gold	SELECT T2.name FROM singer_in_concert AS T1 JOIN singer AS T2 ON T1.singer_id = T2.singer_id JOIN concert AS T3 ON T1.concert_id = T3.concert_id WHERE T3.year = 2014	What are the names of singers who performed in concerts whose year is 2014?	-	-
Baseline	SELECT singer.Name FROM singer_in_concert JOIN singer ON singer_in_concert.Singer_ID = singer.Singer_ID WHERE singer.Song_release_year = 2014 (**missing table "concert"**)	What are the names of singers who were released in 2014 who performed in concerts?	0.020	0.792
Semantic	SELECT singer.Name FROM singer_in_concert JOIN singer ON singer_in_concert.Singer_ID = singer.Singer_ID JOIN concert ON singer_in_concert.concert_ID = concert.concert_ID WHERE concert.Year = 2014	What are the names of singers who performed in concerts whose year is 2014?	0.015	0.823

Error Type	Wrongly added Filter			
Original Question	Find the pixel aspect ratio and nation of the tv channels that do not use English.			
Ranking	**SQL**	**Back-translated Question**	c_i	s_i
Gold	SELECT Pixel_aspect_ratio_PAR , country FROM tv_channel WHERE LANGUAGE $\neq$ 'English'	What are the aspect ratios and countries of tv channels whose language is not English?	-	-
Baseline	SELECT TV_Channel.Pixel_aspect_ratio_PAR, TV_Channel.Country FROM TV_Channel WHERE TV_Channel.Language $\neq$ "English"	What are the aspect ratios and countries of tv channels whose language is not english?	1.000	0.654
Semantic	SELECT TV_Channel.Pixel_aspect_ratio_PAR, TV_Channel.Country FROM TV_Channel WHERE TV_Channel.Language $\neq$ "English" AND TV_Channel.Country $\neq$ "English" (**wrong additional filter**)	What are the aspect ratios and countries of tv channels whose country is not english and whose language is not english?	0.008	0.673

Table 2: Examples of types of errors due to re-ranking. For each error type, we show the natural language question and the corresponding SQL gold standard. Next we show the top candidates according to the *Confidence* ranking and the *Semantic* ranking. c_i and s_i refer to confidence score of the NL-to-query translation and the similarity score between the natural language questions, respectively.

NL_{in}:	Whats the average track size of tracks purchased from 120 S Orange Ave?				
i	NL_{gen}	c_i	s_i	m_i^{equal}	OK
1	What is the average size of all tracks on invoice lines which are part of invoices?	0.669	0.49	0.327	F
2	What is the average size of all tracks on invoice lines which are part of invoices whose billing street is 120 S Orange Ave?	0.668	0.61	0.407	T
9	What is the average size of all tracks on Albums on invoice lines which are part of invoices whose billing street is 120 S Orange Ave?	0.632	0.3	0.1896	F
NL_{in}:	Which companies from Mexico produced their films in Mexico ?				
i	NL_{gen}	c_i	s_i	m_i^{equal}	OK
1	What are the names of companies which produced movies whose status is Mexico?	0.729	0.676	0.492	F
3	What are the names of companies which produced movies which were produced in countries whose name is Mexico?	0.712	0.751	0.534	T
5	What are the names of companies which produced movies whose name is Mexico?	0.664	0.741	0.492	F
NL_{in}:	What are the distinct template type descriptions for the templates ever used by any document?				
i	NL_{gen}	c_i	s_i	m_i^{equal}	OK
1	What are the distinct descriptions of template types for templates?	0.494	0.686	0.338	F
2	What are the distinct descriptions of template types for templates used for documents?	0.091	0.973	0.166	T
3	Show me everything about template types.	0.031	0.133	0.050	F

Table 3: Illustrative examples of the impact of re-ranking. We show three original questions (NL_{in}) and the corresponding back-translated examples (NL_{gen}). Value i denotes the rank in the *Confidence* ranking, c_i is the confidence score of the decoder, s_i is the similarity score, m_i^{equal} is the combination of c_i and s_i, OK indicates whether the generated query is correct (T = true, F = false).

6.2 Error Analysis: Confidence vs. Semantic Ranking

To better understand the differences between the *Semantic* and *Confidence* rankings, we analyze the cases in which one of the two ranking schemes returns a correct query, and the other one does not. This analysis is performed on the *Bridge* output where in 19.2% of the cases, only one of the two ranking schemes returns the correct hypothesis. In 25% of the cases in which only the *Confidence*

ranking returns a correct query, the *Semantic* ranking returned a query with a redundant *WHERE*-clause, and in 20% of cases, the *Semantic* ranking returned a wrong attribute in the projection. This suggests that the *Semantic* ranking is not stable against redundant information in the query and slight variations in the return attributes.

In the cases where only the *Semantic* ranking returns a correct query, the query returned by the *Confidence* ranking contains missing or redundant *Join*-clauses in 47% of cases and wrong query types in 21% of cases. This suggests that the *Semantic* ranking's strength lies in detecting missing relations and detecting wrong query types (i.e., SUM instead of COUNT).

In Table 2 two examples of errors are shown. The first example shows a missing join operation of *Confidence*. In particular, the table "concert" is missing in the SQL statement. In this case the confidence score of the wrong *Confidence* query, i.e. $c_i = 0.02$, is higher than the confidence of the correct *Semantic* query, i.e. 0.015. On the other hand, the semantic textual similarity score s_i of the correct *Semantic* query, i.e., 0.823, is higher than the score of the incorrect *Confidence* query, i.e., 0.792. We note that although the confidence score of the incorrect query is the highest of all hypotheses, it is a low score. Usually, the confidence scores are around 1.0.

The second example shows the problem of an additional filter (TV_Channel.Country $\neq$ "English"), which confuses the semantic similarity score. The *Confidence* ranking selects the correct query with high confidence, i.e. 1.0. However, the semantic score of the incorrect *Semantic* query, i.e., 0.673, is higher than the semantic score of the correct query, i.e., 0.654.

This phenomenon motivates the *Threshold Weighting*. The reason is that high confidence scores from the NL-to-Query system are more trustworthy than the semantic scores. However, in cases where the NL-to-Query system is not confident, the semantic score performs well. The automatically determined threshold in our experiments lies at around 0.9.

6.3 Qualitative Analysis

In Table 3, we show examples of the different rankings. We show three representative examples of a 15-best list. In the first example, we note that the hypothesis with the best confidence score, i.e., c_1 = 0.669, is incorrect. The second best hypothesis, according to the confidence score, is correct and has a very similar score to the hypothesis placed first (0.669 *vs.* 0.668). The hypothesis that is placed 9^{th} adds an unnecessary relation. However, the confidence score is still close to the hypothesis placed first. The semantic score, on the other hand, is more accurate. The correct hypothesis is placed 1^{st} with a large margin (0.61 *vs.* 0.49) and an even larger difference with the score of the 9^{th} place. Finally, the combined score m_2^{equal} of 0.407 clearly identifies result 2 as the correct one.

The second example shows a similar pattern: the first hypothesis with a confidence score c_1 of 0.729 is obviously wrong. The second hypothesis, which is correct, has a slightly lower confidence score c_2 of 0.712. The *Semantic* score s_3 of 0.751 ranks the set of hypotheses correctly. However, *Semantic* re-ranking alone is not enough since the 5^{th} ranked example has a very high semantic similarity score while being incorrect. In this case the *Equal Weighing* approach m_i^{equal} helps differentiating: While s_3 and s_5 are very close, m_3^{equal} and m_5^{equal} have a bigger margin.

The last example shows a case where the *Equal Weighting* does not work. Although the semantic score s_2 of 0.933 works to find the correct answer, the confidence score c_2 of 0.091 of the correct hypothesis is much lower than the confidence of the incorrect hypothesis, $c1$ of 0.494. In this case, the *Threshold Weighting* would work well as it relies on s_i for the cases where the maximum confidence score is too low.

7 Conclusion

We proposed a novel approach to improve semantic NL-to-Query systems based on back-translating the generated query into a natural language question, and re-ranking the top hypothesis of the NL-to-Query system according to the semantic similarity of the generated questions with regard to the original question. Our approach improves over strong, publicly available systems by up to 3 percentage points on the Spider dataset and up to 9 points on the OTTA dataset.

Our results clearly show the potential of back-translation for improving NL-to-Query systems, and it could be applied to more general semantic parsing problems as long as a generation method is available.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 863410.

References

Katrin Affolter, Kurt Stockinger, and Abraham Bernstein. 2019. A comparative survey of recent natural language interfaces for databases. *The VLDB Journal*, 28(5):793–819.

Shubham Agarwal, Marc Dymetman, and Eric Gaussier. 2018. Char2char generation with reranking for the e2e nlg challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 451–456.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Angela Bonifati, Wim Martens, and Thomas Timm. 2017. An analytical study of large SPARQL query logs. *Proc. VLDB Endow.*, 11(2):149–161.

Ursin Brunner and Kurt Stockinger. 2021. Valuenet: A natural language-to-sql system that learns from database information. *International Conference on Data Engineering (ICDE)*.

Jan Deriu, Katsiaryna Mlynchyk, Philippe Schläpfer, Alvaro Rodrigo, Dirk von Grünigen, Nicolas Kaiser, Kurt Stockinger, Eneko Agirre, and Mark Cieliebak. 2020. A methodology for creating question answering corpora using inverse data annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 897–911, Online. Association for Computational Linguistics.

Jan Milan Deriu and Mark Cieliebak. 2018. Syntactic manipulation for generating more diverse and interesting texts. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 22–34.

Li Dong and Mirella Lapata. 2018. Coarse-to-fine decoding for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–742, Melbourne, Australia. Association for Computational Linguistics.

Ondřej Dušek and Filip Jurcicek. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51.

Ahmed Elgohary, Saghar Hosseini, and Ahmed Hassan Awadallah. 2020. Speak to your parser: Interactive text-to-SQL with natural language feedback. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2065–2077, Online. Association for Computational Linguistics.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. Nubia: Neural based interchangeability assessor for text generation.

Andreas Kokkalis, Panagiotis Vagenas, Alexandros Zervakis, Alkis Simitsis, Georgia Koutrika, and Yannis Ioannidis. 2012. Logos: A system for translating queries into narratives. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, page 673–676, New York, NY, USA. Association for Computing Machinery.

G. Koutrika, A. Simitsis, and Y. E. Ioannidis. 2010. Explaining structured queries in natural language. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pages 333–344.

Igor Labutov, Bishan Yang, and Tom Mitchell. 2018. Learning to learn semantic parsers from natural language supervision. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1676–1690, Brussels, Belgium. Association for Computational Linguistics.

Xi Victoria Lin, Richard Socher, and Caiming Xiong. 2020. Bridging textual and tabular data for cross-domain text-to-sql semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, November 16-20, 2020*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Axel-Cyrille Ngonga Ngomo, Lorenz Bühmann, Christina Unger, Jens Lehmann, and Daniel Gerber. 2013. Sorry, i don't speak sparql: Translating sparql queries into natural language. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 977–988, New York, NY, USA. Association for Computing Machinery.

Fatma Odzcan, Abdul Quamar, Jaydeep Sen, Chuan Lei, and Vasilis Efthymiou. 2020. State of the art and open challenges in natural language interfaces to data. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, SIGMOD '20, page 2629–2636, New York, NY, USA. Association for Computing Machinery.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer,

R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

John Platt. 2000. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jaydeep Sen, Chuan Lei, Abdul Quamar, Fatma Özcan, Vasilis Efthymiou, Ayushi Dalmia, Greg Stager, Ashish Mittal, Diptikalyan Saha, and Karthik Sankaranarayanan. 2020. Athena++: Natural language querying for complex nested sql queries. *Proc. VLDB Endow.*, 13(12):2747–2759.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Alane Suhr, Srinivasan Iyer, and Yoav Artzi. 2018. Learning to map context-dependent sentences to executable formal queries. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2238–2249, New Orleans, Louisiana. Association for Computational Linguistics.

Pius von Däniken. 2021. Improving a semantic parser through user interaction. *Publikationen School of Engineering*.

Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a semantic parser overnight. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1332–1342, Beijing, China. Association for Computational Linguistics.

Kun Xu, Lingfei Wu, Zhiguo Wang, Yansong Feng, and Vadim Sheinin. 2018. SQL-to-text generation with graph-to-sequence model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 931–936, Brussels, Belgium. Association for Computational Linguistics.

Ziyu Yao, Yu Su, Huan Sun, and Wen-tau Yih. 2019. Model-based interactive semantic parsing: A unified framework and a text-to-SQL case study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5447–5458, Hong Kong, China. Association for Computational Linguistics.

Ziyu Yao, Yiqi Tang, Wen-tau Yih, Huan Sun, and Yu Su. 2020. An imitation game for learning semantic parsers from user interaction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6883–6902, Online. Association for Computational Linguistics.

Pengcheng Yin and Graham Neubig. 2017. A syntactic neural model for general-purpose code generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 440–450, Vancouver, Canada. Association for Computational Linguistics.

Pengcheng Yin and Graham Neubig. 2019. Reranking for neural semantic parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4553–4559, Florence, Italy. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

A On Query-to-NL

While Query-to-Text is not a contribution of our work, we discuss and motivate our choice of OT3 as our Query-to-Text engine. We adapted OT3 to handle all the domains in the Spider development set, which comprises 20 databases. In order to handle SQL queries, we translate SQL queries into OTs using a rule-based approach. The main advantage over statistical methods is that we can be sure that the queries are correctly back-translated to text. This is due to the rule-based nature of OT3.

Sanity Check. In order to show that OT3 correctly renders the semantics of a query, we first perform a sanity check, where we backtranslated the gold-standard tree for a given question. Thus, we need to show that the original question and the back-translation are semantically equivalent. As negative examples, we also mix in randomly sampled human questions, thus the original question and the negative back-translation should never be semantically equivalent. We let humans annotate this data, that is, we showed humans pairs of original questions and either a positive or negative back-translation. In this setting, humans agree in 94% of cases with the parsing ground-truth. This shows that the synthetic questions are understandable and

generally maintain the semantics of the underlying OT. The experiments show that the synthetic questions are of high quality and can be used as basis for re-ranking.

Limitations. OT3 does not handle *GROUP BY*, *sub-queries* and *set operations*, thus, we discard these samples from the Spider and OTTA development sets, keeping 82% of *OTTA-Moviedata*, 76% of *OTTA-Chinook* and 43% of *Spider*. The reported results are on these subsets of the datasets. Note that several studies on natural language query logs (Bonifati et al., 2017; Affolter et al., 2019) show that typical queries in real-world applications are far less complex than the ones contained in the Spider dataset. Hence, not supporting GROUP BYs, sub queries or set queries is not a significant issue in a real-world scenario. Note that our method can still be applied to the full datasets, defaulting to the Confidence ranking when none of the hypotheses could be back-translated. The positive results are consistent, but the improvement is lower, correlated with the coverage. E.g. an overall improvement of 0.67% for the whole Spider (with the Bridge system using equal mixed re-ranking), which roughly corresponds to the 1.57% improvement obtained on the 43% subset of Spider which does not contain complex SQL operations.

Selection. The choice of OT3 is motivated by the fact that it renders relationships between entities naturally. For instance, the relationship between persons and movies, which is modelled via the cast table, is expressed as "Persons that play in movies". For instance, Logos (Kokkalis et al., 2012) expresses the same relationship as "Persons associated with movies", which is not natural and cannot be handled by our semantic textual similarity tool. We also evaluated statistical models, which suffer from hallucinations (i.e., adding text that is not semantically related to the query) and are generally unreliable. Thus, we are not aware of any Query-to-Text solution, that handles all types of queries (Group By, Set Operations, Nested Queries) such that the generated texts read naturally. Thus, OT3 has proved to be best suited for our task.

B On Evaluation

We adapted the *Component Equality* measure for operation trees (OTs) since we translate the SQL queries of the *Spider*-based systems to OTs. For OTs, this measure checks if the nodes of the predicted tree correspond to the nodes of the gold standard tree. This allows measuring query equality independently of the order of the nodes. Furthermore, we adapted this analysis also to measure if the *Join* attributes are rendered correctly. We decided against a result-based evaluation since it is impossible to reasonably evaluate queries that return an empty result set, often leading to over-estimating the quality of NL-to-Query systems. This happens often in cases where the result set is empty or for count questions. For Spider the databases are very small and do not contain much data, thus, queries tend to return empty results. For OTTA, which uses Yes/No questions, this problem is even more pronounced. Thus, the result-based evaluation is not reliable, and we opted for the component-based evaluation, which is now the standard evaluation for the Spider dataset.

Appendix

A.1 List of Publications

Computer Vision, Perception and Cognition Group

J.E. Huggins, D. Krusienski, M.J. Vansteensel, D. Valeriani, A. Thelen, S. Stavisky, J.J.S. Norton, A. Nijholt, G. Müller-Putz, N. Kosmyna, L. Korczowski, C. Kapeller, C. Herff, S. Halder, C. Guger, M. Grosse-Wentrup, R. Gaunt, A.N. Dusang, P. Clisson, **R. Chavarriaga**, C.W. Anderson, B. Allison, T. Aksenova, E. Aarnoutse. *"Workshops of the eighth international brain-computer interface meeting: BCIs: the next frontier"*. Brain-Computer Interfaces, 9.2, 69-101, February 08, 2022, DOI: 10.1080/2326263X.2021.2009654.

T. Stadelmann, T. Klamt, P.H. Merkt. *"Data centrism and the core of Data Science as a scientific discipline"*. Archives of Data Science, Series A, 8.2, March 10, 2022, DOI: 10.5445/IR/1000143637.

C. v.d.Malsburg, **T. Stadelmann**, B.F. Grewe. *"A Theory of Natural Intelligence"*. arXiv, April 22, 2022, DOI: 10.48550/ARXIV.2205.00002.

M. Ienca, J.J. Fins, R.J. Jox, F. Jotterand, S. Voeneky, R. Andorno, T. Ball, C. Castelluccia, **R. Chavarriaga**, H. Chneiweiss, A. Ferretti, O. Friedrich, S. Hurst, G. Merkel, F. Molnár-Gábor, J.-M. Rickli, J. Scheibner, E. Vayena, R. Yuste, P. Kellmeyer. *"Towards a governance framework for brain data"*. Neuroethics, 15.2, 20, June 03, 2022, DOI: 10.1007/S12152-022-09498-8.

F. Dell'Agnola, P.-K. Jao, A. Arza, **R. Chavarriaga**, J.d.R. Millán, D. Floreano, D. Atienza. *"Machine-Learning Based Monitoring of Cognitive Workload in Rescue Missions With Drones"*. IEEE Journal of Biomedical and Health Informatics, 26.9, 4751-4762, June 27, 2022, DOI: 10.1109/JBHI.2022.3186625.

P. Sager, **S. Salzmann**, F. Burn, **T. Stadelmann**. *"Unsupervised Domain Adaptation for Vertebrae Detection and Identification in 3D CT Volumes Using a Domain Sanity Loss"*. Journal of Imaging, 8.8, August 19, 2022, DOI: 10.3390/JIMAGING8080222.

F.M. Schmitt-Koopmann, E.M. Huang, H.-P. Hutter, **T. Stadelmann**, A. Darvishy. *"FormulaNet: A Benchmark Dataset for Mathematical Formula Detection"*. IEEE Access, 10, 91588-91596, August 29, 2022, DOI: 10.1109/ACCESS.2022.3202639.

C. v.d.Malsburg, B.F. Grewe, **T. Stadelmann**. *"Making Sense of the Natural Environment"*. The Biannual Conference of the German Cognitive Science Society, September 05, 2022.

L. Tuggener, J. Schmidhuber, **T. Stadelmann**. *"Is it enough to optimize CNN architectures on ImageNet?"*. Frontiers in Computer Science, 4, November 15, 2022, DOI: 10.3389/FCOMP.2022.1041703.

Explainable Artificial Intelligence Group

L. Wertz, **J. Bogojeska**, K. Mirylenka, J. Kuhn. *"Evaluating Pre-Trained Sentence-BERT with Class Embeddings in Active Learning for Multi-Label Text Classifcation"*. Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, 2, 366-372, November, 2022, DOI: 10.21256/ZHAW-26577.

Intelligent Vision Systems Group

F.-P. Schilling, D. Flumini, R.M. Füchslin, E. Gavagnin, A. Geller, S. Quarteroni, **T. Stadelmann**. *"Foundations of Data Science: a comprehensive overview formed at the 1st International Symposium on the Science of Data Science"*. Archives of Data Science, Series A, 8.2, May 05, 2022, DOI: 10.5445/IR/1000146422.

I. Herzig, P. Paysan, S. Scheib, A. Zuest, **F-P. Schilling**, J. Montoya, **M. Amirian**, **T. Stadelmann**, P. Eggenberger, R. Füchslin, L. Lichtensteiger. *"Deep learning-based simultaneous multi-phase deformable image registration of sparse 4D-CBCT"*. Medical Physics, 49.6, 325-326, June 09, 2022, DOI: 10.21256/ZHAW-25181.

Natural Language Processing Group

J. Deriu, D. Tuggener, P. v.Däniken, M. Cieliebak. *"Probing the Robustness of Trained Metrics for Conversational Dialogue Systems"*. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2, 750-761, May, 2022, DOI: 10.21256/ZHAW-26130.

B. Wang, K. Wang, S. Li, **M. Cieliebak**. *"Detection of Typical Sentence Errors in Speech Recognition Output"*. June, 2022.

M. Plüss, **M. Hürlimann**, M. Cuny, A. Stöckli, Nikolaos Kapotis, J. Hartmann, M.A. Ulasik, C. Scheller, Y. Schraner, A. Jain, **J. Deriu**, **M. Cieliebak**, M. Vogel. *"SDS-200: A Swiss German Speech to Standard German Text Corpus"*. Proceedings of the 13th Conference on Language Resources and Evaluation, 3250-3256, June, 2022, DOI: 10.21256/ZHAW-26131:

M. Hürlimann, R. Mastropietro, D. Puccinelli, F. Rinaldi, **M. Cieliebak**. *"Welcome to the 7th Swiss Text Analytics Conference"*, Proceedings of the 7th Swiss Text Analytics Conference, June, 2022:

M. Hürlimann, J. Galbier, **M. Cieliebak**. *"Speech-to-text technology for hard-of-hearing people"*. ERCIM News 130, 15-16, July 07, 2022, DOI: 10.21256/ZHAW-27071.

P. v.Däniken, **J. Deriu**, **D. Tuggener**, **M. Cieliebak**. *"On the Effectiveness of Automated Metrics for Text Generation Systems"*. arXiv, October 24, 2022, DOI: 10.48550/ARXIV.2210.13025.

P. v.Däniken, **J. Deriu**, E. Agirre, U. Brunner, **M. Cieliebak**, K. Stockinger. *"Improving NL-to-Query systems through re-ranking of semantic hypothesis"*. 5th International Conference on Natural Language and Speech Processing, December, 2022, DOI: 10.21256/ZHAW-26147.

A.2 CAI Team as of 31.12.2022

Name	Function	Email
Mohammadreza Amirian	Doctoral Student	amir@zhaw.ch
Daniel Barco	Doctoral Student	baoc@zhaw.ch
Dr. Jasmina Bogojeska	Senior Lecturer, Head of XAI Group	bogo@zhaw.ch
Dr. Ricardo Chavarriaga	Senior Researcher, MSc Program Coordinator	char@zhaw.ch
Prof. Dr. Mark Cieliebak	Senior Lecturer, Head of NLP Group	ciel@zhaw.ch
Maggie Deller Haight	Administrative Assistant	deel@zhaw.ch
Dr. Philipp Benedikt Denzel	Research Associate	denp@zhaw.ch
Dr. Jan Milan Deriu	Research Associate	deri@zhaw.ch
Dr. Manuel Dömer	Associate Faculty Member	doem@zhaw.ch
Raphael Emberger	Research Assistant	embe@zhaw.ch
Prof. Dr. Rudolf Marcel Füchslin	Associate Faculty Member	furu@zhaw.ch
Dr. Elena Gavagnin	Associate Faculty Member	gava@zhaw.ch
Nicola Good	Research Assistant, Master Student	goon@zhaw.ch
Prof. Dr. Christoph Heitz	Associate Faculty Member	heit@zhaw.ch
Simona Hovançikova	Research Intern	hova@zhaw.ch
Manuela Hürlimann	Research Associate	hueu@zhaw.ch
Paul-Philipp Luley	Research Assistant, Master Student	lule@zhaw.ch
Benjamin Meyer	Research Assistant, Master Student	mebr@zhaw.ch
Katsiaryna Mlynchyk	Research Assistant	mlyn@zhaw.ch
Daniel Sebastian Neururer	Research Assistant	neud@zhaw.ch
Pascal Sager	Research Assistant, Master Student	sage@zhaw.ch
Prof. Dr. Frank-Peter Schilling	Senior Lecturer, Head of IVS Group	scik@zhaw.ch
Prof. Dr. Thilo Stadelmann	Senior Lecturer, Head of CVPC Group, Director of Centre	stdm@zhaw.ch
Dr. Don Tuggener	Senior Researcher	tuge@zhaw.ch
Lukas Tuggener	Doctoral Student	tugg@zhaw.ch
Pius von Däniken	Research Assistant, Master Student	vode@zhaw.ch
Prof. Dr. Christoph von der Malsburg	Visiting Professor	malsburg@fias.uni-frankfurt.de
Ali Waqar	Joint Doctoral Student with University of Venice	xalw@zhaw.ch
Peng Yan	Doctoral Student	yanp@zhaw.ch

A.3 Location

CAI Centre for Artificial Intelligence

Technikumstrasse 71
PO Box
CH-8401 Winterthur
www.zhaw.ch/en/engineering/institutes-centres/cai/

Contact

Prof. Thilo Stadelmann
Phone +41 58 934 72 08
thilo.stadelmann@zhaw.ch

Administration

Maggie Deller Haight
Phone +41 58 934 76 48
office.cai@zhaw.ch

Zurich University
of Applied Sciences

School of Engineering

CAI Centre for Artificial Intelligence
Technikumstrasse 71
PO Box
CH-8401 Winterthur
www.zhaw.ch/cai/en

Contact
Prof. Dr. Thilo Stadelmann
Head of Centre
+41 58 934 72 08
thilo.stadelmann@zhaw.ch

Administration
+41 58 934 76 48
office.cai@zhaw.ch