PRAGUE

3-5 JULY 2023

IACAP

International Association of Computing and Philosophy

This program was compiled from LaTeX source code on July 2, 2023.

This template originates from LaTeXTemplates.com and is based on the original version at: https://github.com/maximelucas/AMCOS_booklet.

Front cover is designed by Arzu Formánek and uses a photograph by Andrew Friedrich, under the Unsplash License.

Contents

About

The board of IACAP welcomes you to the IACAP Conference 2023 in Prague.

The International Association of Computing and Philosophy

The International Association of Computing and Philosophy (IACAP) has a long-lasting tradition of promoting philosophical dialogue and interdisciplinary research on all aspects of the digital turn. IACAP's members have contributed to shaping the philosophical and ethical debate about computing, information technologies, and artificial intelligence. The 2023 annual meeting continues this tradition and gathers philosophers, ethicists, roboticists, and computer scientists and engineers interested in the following topics:

- Artificial Intelligence and Machine Learning
- Algorithmic Opacity and Bias
- Artificial Life and Moral Agency
- Autonomous Weapon Systems
- Computation, Cognition, and Cognitive Science
- Computational Modelling in Science and Social Science
- Computer-Mediated Communication
- Ethical Problems and Societal Impact of Computation and Information
- History of Computing
- Information Culture and Society
- Metaphysics of Computing
- Philosophy of Information
- Philosophy of Information Technology
- Robotics
- Virtual Reality

... and related issues

The Karel Čapek Centre

The 2023 meeting is hosted in collaboration with the Karel Čapek Centre for Values in Science and Technology (CEVAST). The conference is co-organised with the generous support of the AV21 Philosophy and Future of Artificial Intelligence programme. CEVAST was formed in 2018 under the Czech Academy of Sciences in the historic city of Prague. CEVAST has a reputation as a leading European institution dealing with emerging issues in technology ethics.

Programme Chairs

IACAP:	Steve McKinlay (president)	Björn Lundgren (vice-president)
	Arzu Formánek	Ramón Alvarado
	Ahmed Amer	Brian Ballsun-Stanton (technical director)
	Hajo Greif	Thomas M. Powers
CEVAST:	Tomáš Hříbek	
	David Černý	

Timetable

Monday, 3 July

09:00–11:00	**Registration**	
11:00–12:30	Room 131	**Symposium:** New Directions in the Philosophy of Computation
	Room 300	Human–Robot Interaction
	Room 301	Explainability
12:30–13:30	**Lunch**	
13:30–15:00	Room 131	**Invited Keynote:** Professor Mark Coeckelbergh
15:00–15:30	**Afternoon Tea**	
15:30–17:30	Room 300	Minds and Machines SIG
	Room 301	Agency and Consciousness
	Room 325	Epistemic Problems
17:30–18:30	**Welcome Drinks**	
19:00–20:00	**Social Outing: boat trip on the Vltava river**	

Tuesday, 4 July

08:30–10:30	Room 300	Minds and Machines SIG
	Room 301	Digital Governance
	Room 325	AI Ethics
	Room 326	Healthcare–1
10:30–11:00	**Coffee Break**	
11:00–12:30	Room 300	Healthcare–2
	Room 301	**Symposium:** Sketching Introductory Courses in the Philosophy of Computing
	Room 325	AI Legislation
	Room 326	Simulation
12:30–13:30	**Lunch**	
13:30–15:00	Room 300	**Covey Award Keynote:** Professor Oron Shagrir
15:00–15:30	**Afternoon Tea**	
15:30–17:30	Room 300	**Workshop:** An afternoon with Large Language Models
	Room 301	The Ethics of Autonomous Military Robots SIG
	Room 325	Epistemology
	Room 326	Ethics and Applications
19:00–21:00	**Conference Dinner**	

Wednesday, 5 July

09:00–10:30	Room 300 Room 301 Room 325 Room 326	AI Impact **Symposium:** Turing and Ashby on Computation, Mechanisms and Intelligence Data Decision-making
10:30–11:00	**Coffee Break**	
11:00–12:30	Room 300	**Simon Award Keynote:** Assistant Professor Kathleen A. Creel
12:30–13:30	**Lunch**	
13:30–14:30	Room 300 Room 301 Room 325	Information Minds Digital Platforms
14:45–15:15	Room 300	Presidental Lecture and Conference Farewell
15:30–16:30	Room 300	IACAP Members General Meeting

Papers presented during Monday, July 3 11:00–12:30

Room 131 - Aula: Symposium: New Directions in the Philosophy of Computation

- New Directions in the Philosophy of Computation by Andre Curtis-Trudel, John Symons, Paula Quinon, Gualtiero Piccinini

Room 300: Human–Robot Interaction

- New Take on Robots Ethical by Design by Gordana Dodig Crnkovic, Baran Çürüklü, Jathoosh Thavarasa, Tobias Holstein
- Humanoid robots as socially disruptive technologies: three ways in which humanoid robots disrupt our human relational experiences by Cindy Friedman
- Is it wrong to kick Kickable 3.0? An affordance based approach to ethics of human-robot interaction by Arzu Formánek

Room 301: Explainability

- Genuine Understanding or Mere Rationalizations? Approximations and Idealizations in Science and XAI by Luis Lopez
- Mechanistic Computation and its Problems by Luke Kersten
- Understanding Mechanistic Explanations of Epistemically Opaque Deep Learning Systems by Marcin Rabiza

Papers presented during Monday, July 3 15:30–17:30

Room 300: Minds and Machines SIG

- Diffusing the Creator: Attributing Credit for Generative AI Outputs by Donal Khosrowi, Finola Finn, Elinor Clark
- How to Accelerate Ethics for Innovation and Against Precaution in Generative AI by James Brusseau
- From thinking to how to think by Nicola Angius, Alessio Plebe, Pietro Perconti, Alessandro Acciai
- Belief revision for language models by Thomas Hofweber

Room 301: Agency and Consciousness

- Proxy Assertions and Agency: the case of machine-assertions by Chirag Arora
- Thinking Without Phenomenal Character: How are Artificial Intelligence, Cognitive Phenomenology, and Extended Cognition Related? by Max Parks
- Is 'Responsible AI' Accountable?: A Critical Analysis by Riya Manna
- Moral Agency, Consciousness and Mental States by Zacharus Gudmunsen

Room 325: Epistemic Problems

- The theory and practice of computational errors by Nico Formánek
- AI as an Epistemic Technology by Ramon Alvarado
- Disinformation and Epistemic Injustice: Implications for the Philosophy of Information and Ethics. by Steve McKinlay

Papers presented during Tuesday, July 4 08:30–10:30

Room 300: Minds and Machines SIG

- Smooth Answers and Fabulations. Engaging with Processed Language by Leonie Möck, Sven Thomas
- Gorgias-tp4: Socratic critique of rhetoric and language models by Stephen Rainey
- LLMs in academia: a thread or an opportunity? by Vojtěch Kolomý
- Measuring scientific understanding in Large Language Models by Kristian González Barman, Henk de Regt, Sascha Caron, Tom Claassen

Room 301: Digital Governance

- How to understand "user autonomy" for efficient platform regulation by Eloise Soulier
- Notes on the Formative Power of Concepts: The Case of Digital Sovereignty and COVID-19 by Gernot Rieder
- The Politics of Platform Governance by Laura Fichtner

Room 325: AI Ethics

- Exploratory Sandboxes and Experimental AI ethics: Performing Regulations over Prompts, Model Chaining, and Guidance programs by Denisa Reshef Kera, Brian Ballsun-Stanton, Frantisek Kalvas
- Humans and Machines: Challenges from Artificial Intelligence - Presenting a very recent report by the German Ethics Council by Judith Simon
- The case for e-vigilance: To what extent can and should we trust AI? by Rico Hauswald

Room 326: Healthcare–1

- Conversational Artificial Intelligence used in psychotherapy as a special kind of cognitive artifact by J. P. Grodniewicz, Mateusz Hohol
- AI for psychiatry: close encounters of the algorithmic kind by Y J Erden
- In patients' interest? Ethical and social considerations of (intelligent) virtual agents in the healthcare sector by Catharina Rudschies

Papers presented during Tuesday, July 4 11:00–12:30

Room 300: Healthcare–2

- What kind of explanations are requested in the clinical deployment of AI to ensure patients' trust in medical decisions? by Anne Gerdes
- Human sovereignty when a disease is controlled through restrictions on persons: Citizens' views on whether scientific evidence for restrictions is necessary by Ingvar Tjostheim, John A. Waterworth
- Trust and Responsibility in Joint Human-AI Decision-making in Medicine by Vilius Dranseika

Room 301: Symposium: Sketching Introductory Courses in the Philosophy of Computing

- Sketching Introductory Courses in the Philosophy of Computing by Robin Hill, Ramón Alvarado

Room 325: AI Legislation

- The Open Texture of 'Algorithm' in Legal Language by Davide Baldini, Matteo De Benedetto
- Artificial Intelligence, Radical Ignorance, and the Institutional Context of Consent by Etye Steinberg
- A Causal Analysis of Harm by Sander Beckers, Hana Chockler, Joseph Halpern

Room 326: Simulation

- All that glitters is not a deduction: Non-deductive methods in computational modelling by Michal Hladky
- Deep Learning in Simulative Sciences by Nicola Angius, Alessio Plebe

Papers presented during Tuesday, July 4 15:30–17:30

Room 301: The Ethics of Autonomous Military Robots SIG

- Utopia's Killer Robots by Ahmed Amer
- Responsibility gaps: old wine in new bottles by Ann-Katrien Oimann, Fabio Tollon
- AI as a threat to democracy across defence and civilian sectors by May Thorseth
- Bombs, Bots, and the Principle of Distinction by Nathan Wood

Room 325: Epistemology

- The Complementary Minds: A Logical Framework for Belief Formation by Dimana Orlinova Anastassova, Massimiliano Badino, Fabio Aurelio D'Asaro
- Probability Space for First-Order Predicate Logic by Kenneth Presting
- Understanding the analogical roots of Agent-Based Models in economics and social sciences by Massimo Rusconi, Davide Secchi, Raffaello Seri
- The Meta-Ontology of AI systems AI by Roman Krzanowski, Pawel Polak

Room 326: Ethics and Applications

- Experiencing AI and the Relational 'Turn' in AI Ethics by Jason Branford
- A revision and extension of List & Goodin's model of epistemic democracy by Pablo Rivas-Robledo
- War or peace between humanity and artificial intelligence by Wolfhart Totschnig

Papers presented during Wednesday, July 5 09:00–10:30

Room 300: AI Impact

- Digital Transformations of Democracy: How to Successfully Solve Shared Problems in the Age of Anthropocene by Jan-Philipp Kruse
- Deepfakes, Brain Emulation, and Things We Are: Asserting Rights of Origination by Jeremy Hansen
- Ethical Issues in Generating Robot Nudgers by Stefano Calboli

Room 301: Symposium: Turing and Ashby on Computation, Mechanisms and Intelligence

- Turing and Ashby on Computation, Mechanisms and Intelligence by Hajo Greif, Adam Kubiak, Paula Quinon, Paweł Stacewicz

Room 325: Data

- The Colour of Offline Data by Brian Ballsun-Stanton
- Topical alignment with the New York Times boosts content popularity on the social web by Kate Barnes, Roland Molontay
- Ethics of Metadata Ownership by Navid Shaghaghi, Maria Joseph Israel

Room 326: Decision-making

- AI decision-making and the problem of value-alignment by Björn Lundgren
- Paternalistic machines by Kalle Grill
- Understanding over explanation: an epistemic rights perspective on XAI by Yeji Streppel

Papers presented during Wednesday, July 5 13:30–15:00

Room 300: Information

- Continental Approches to Philosophy of Information by Ashley Woodward
- An Inquiry Concerning the Persistence of Information by Roman Krzanowski

Room 301: Minds

- Exploring and Understanding the Role of Motivation in Computational Models of the Mind by Ron Sun
- Illusionism and Machine Consciousness by Tomas Hribek

Room 325: Digital Platforms

- Digital Recording and the Hazards of Unbounded Moral Judgment by Bart Kamphorst, Elizabeth O'Neill
- Skilled agency and illegitimate technological control for Social Media users by Lavinia Marin

Special Sessions and Tracks

Invited Keynote Address

Monday, 3 July, 13:30–15:00, Room 131

The Executive Board of the International Association for Computing and Philosophy , the Czech Academy of Sciences, and Karel Čapek Center for Values in Science and Technology are pleased to announce that Mark Coeckelbergh has accepted the invitation to give an Invited Keynote Address at the 2023 IACAP conference. The board approached Professor Coeckelbergh in acknowledgement of his contributions to systematic exploration of the ethical and political ramifications of artificial intelligence technologies.

Mark Coeckelbergh is a full Professor of Media and Technology at the University of Vienna, and until recently Vice Dean of the Faculty of Philosophy and Education. Currently, he is also ERA Chair at the Institute of Philosophy of the Czech Academy of Sciences in Prague and Guest Professor at WASP-HS and University of Uppsala. Previously he was the President of the Society for Philosophy of Technology. Professor Coeckelbergh has been a member of various advisory boards and policy entities, at both national and international level, aimed at regulating the area of robotics and artificial intelligence, such as the European Commission's High Level Expert Group on Artificial Intelligence, the Austrian Council on Robotics and Artificial Intelligence, and the Austrian Advisory Council on Automated Mobility. In his research, Professor Coeckelbergh focuses on the ethical and political issues of technology, in particular robotics and artificial intelligence. Recently, he has been exploring the interface between technology ethics and environmental ethics. He is the author of nearly 20 books, most recently Introduction to Philosophy of Technology (Oxford University Press, 2019); AI Ethics (MIT Press, 2020); Green Leviathan (Routledge, 2021); The Political Philosophy of AI (Polity, 2022); Robot Ethics (MIT Press, 2022); Self-Improvement: Technologies of the Soul in the Age of Artificial Intelligence (Columbia University Press, 2022); and Digital Technologies, Temporality, and the Politics of Co-Existence (Palgrave, 2023).

AI and the foundations of democracy

Mark Coeckelbergh
Professor of Philosophy of Media and Technology, University of Vienna
Center of Environmental and Technology Ethics, Prague

As AI is used for political manipulation and authoritarian repression, the question arises what impact AI has on democracy and how we can ensure that AI is used for the common good. Political philosophy, next to empirical work on the political effects of AI, can help to answer this question. This talk warns for anti-democratic uses of AI and analyzes the potential influence of AI on the principles and knowledge that form the basis of a healthy and sustainable liberal democracy.

Covey Award Keynote Address

Tuesday, 4 July, 13:30–15:00, Room 300

The International Association for Computing and Philosophy's Executive Board has selected Oron Shagrir for the 2023 Covey Award recognizing senior scholars with a substantial record of innovative research in the field of computing and philosophy broadly conceived. The board recognised Professor Shagrir's significant contribution to our field over several decades; in particular, his contribution to theories of computation.

Oron Shagrir is the Schulman Chair in Philosophy, professor of philosophy and cognitive and brain sciences at the Hebrew University of Jerusalem. He graduated in mathematics and computer science from the Hebrew University, and received his PhD in philosophy and cognitive science from the University of California, San Diego. He was a visiting fellow at the Center for Philosophy of Science at the University of Pittsburgh, and an Erskine Fellow at the University of Canterbury, New Zealand. He has served the academic community in many different roles, and currently, since 2017, he is the vice president for international affairs of the Hebrew University. He was an associate editor of Cognitive Science and served on the editorial boards of several journals and book series. Professor Shagrir's current research focuses on the nature of computation and representation, the role of computational approaches in cognitive and brain sciences, and the history of computability. He is the author of The Nature of Physical Computation (Oxford University Press, 2022), the editor, with Jack Copeland and Carl Posy, of Computability: Turing, Gödel, Church, and Beyond (MIT 2013), and the author of numerous papers on computation and the mind.

How neural networks have altered our philosophical theories of computation

Oron Shagrir
Vice-President for International Affairs
Schulman Chair in of Philosophy
Professor of Philosophy and of Cognitive and Brain Sciences
The Hebrew University of Jerusalem

It is widely acknowledged that neural networks challenge the idea that the mind/brain is a classical Turing-machine-style computing machine. It is less appreciated that neural networks also challenge the notion of computation that dominated the philosophy of mind until about a decade ago. I will focus on the latter, what-is-computation, issue, and tell how neural networks have altered our theories of computation along the years. I start with earlier accounts of physical computation that were widespread from the 1960s to the 2000s. These theories rely on two dogmas. The logical dogma is that there is a strong linkage between the mathematical theories we find in logic and computer science (e.g., computability theory, automata theory, proof theory) and physical computation. The architectural dogma is that the difference between computing and non-computing physical systems has to do (at least in part) with a distinct abstract causal structure, e.g., discrete, digital, or stepwise structure. In the second part of the talk I will show how the re-appearance of neural networks in the 1980s has shaken the logical and architectural dogmas, and forced us to re-think about physical computation. I conclude with the more recent mechanistic, semantic and pragmatic accounts that abandoned the two dogmas and advance core notions of computation that can accommodate classical, neural and other kinds of computation.

Simon Award Keynote Address

Wednesday, 5 July, 11:00–12:30, Room 300

The International Association for Computing and Philosophy's executive board has selected Dr. Kathleen Creel for the 2023 Herbert A. Simon Award for Outstanding Research in Computing and Philosophy, which specifically recognizes scholars at an early stage of their academic career whose research is likely to reshape debates at the nexus of Computing and Philosophy.

Dr. Creel is an assistant professor at Northeastern University, cross appointed between the Department of Philosophy and the Khoury College of Computer Sciences. Her research explores the moral, political, and epistemic implications of machine learning as it is used in non-state automated decision making and in science. A current project focuses on defining, measuring, and ethically evaluating algorithm-derived outcome homogeneity, namely the extent to which monoculture among decision-making systems causes individuals to receive the same outcomes from multiple decision-makers. In other work, she has developed definitions of transparency for complex computational systems, argued that algorithmic arbitrariness is wrong at scale, and contended that ethically setting decision thresholds in medical settings requires the consideration of individual patient values.

Before Northeastern, she received her BA from Williams College in Computer Science and Philosophy. After working as a software engineer at MIT Lincoln Laboratory, she received her MA from Simon Fraser University's Philosophy Department and her Ph.D. from the University of Pittsburgh's History and Philosophy of Science Department. Most recently, she was the Embedded Ethics postdoctoral fellow at Stanford University.

Algorithmic Monoculture and the Ethics of Systemic Exclusion

Kathleen A. Creel
Assistant Professor, Department of Philosophy and Religion and Khoury College of Computer Sciences, Northeastern University

Mistakes are inevitable, but fortunately human mistakes are typically heterogenous. Using the same machine learning model for high stakes decisions creates consistency while amplifying the weaknesses, biases, and *idiosyncrasies* of the original model. When the same person re-encounters the same model or models trained on the same dataset, she might be wrongly rejected again and again. Thus *algorithmic monoculture* could lead to consistent ill-treatment of individual people by *homogenizing the decision outcomes* they experience.

Is it wrong to allow the quirks of an algorithmic system to consistently exclude a small number of people from consequential opportunities? Many philosophers have claimed or indicated in passing that consistent and arbitrary arbitrary exclusion is wrong, even when it is divorced from bias or discrimination. But why and under what circumstances it is wrong has not yet been established. This talk will formalize a measure of outcome homogenization, describe experiments that demonstrate that it occurs, then present an ethical argument for why and in what circumstances outcome homogenization is wrong.

Special Tracks

The Ethics of Autonomous Military Robots

Chair: David Černý Czech Academy of Sciences, Prague, Czech Republic

Modern wars are increasingly fought using artificial intelligence systems and military robots. Recently, the use of fully autonomous military robots that do not require human consent even to carry out a lethal attack has become the subject of widespread ethical debate. Given the topicality and importance of this issue, we have decided to hold a special session at this year's conference entitled The Ethics of Autonomous Military Robots.

Minds & Machines Special Interest Group Track

Chair: Thomas M. Powers University of Delaware, Newark, DE, USA

ChatGPT and other generative AIs have produced some impressive results that were formerly thought to be the sole domain of human intelligence. Works of poetry, art, computer code, scholarly analysis, and legal counsel are among the early products, with more sure to follow. With its ability to create novelty from data, has generative AI finally settled the debate on whether machines can think?

Special Session

An afternoon with Large Language Models

Organiser: Brian Ballsun-Stanton Macquarie University, Sydney, Australia

A special extended afternoon session where attendees will be guided through hands-on explorations of the latest Large Language Models like ChatGPT, Claude, and Bing Chat. We will explore how well these models interact with formal logic, code generation, academic research, large masses of PDFs, and literature searches. We will also demonstrate (and discuss) how these models can be used in the classroom, in ways that will subvert, be indifferent to, or enhance assessments and teaching. All attendees are encouraged to bring questions and their own demonstrations. To join us in the planning of this afternoon, join us on the IACAP Slack!

List of Abstracts

Monday, July 3

11:00–12:30: Symposium: New Directions in the Philosophy of Computation (Room 131 - Aula)

New Directions in the Philosophy of Computation

Andre Curtis-Trudel *<**curtistrudel.1@osu.edu**>*[4]
John Symons[2]
Paula Quinon[1]
Gualtiero Piccinini[3]
[1] Warsaw University of Technology, Poland, [2] University of Kansas, United States, [3] University of Missouri, St. Louis, United States, [4] Lingnan University, Hong Kong

Over the past half-century or so, the notion of computation has come to play a central theoretical role in a wide variety of scientific disciplines. More recently, the philosophy of computation has emerged alongside in an attempt to illuminate this development and situate it within the broader philosophical and scientific landscape. As the philosophy of computation grows into an independent philosophical subdiscipline in its own right, the time is ripe to step back and assess its traditional motivations and foundational assumptions. This is important not only for its own sake, but also for identifying critical avenues for future philosophical work on computation. Thus, the goal of this symposium is to take stock of the philosophy of computation: to consider where it has been, and where it might go next. Topics to be considered include the relationship between physical and mathematical notions of computation, methodological assumptions guiding the development of philosophical theories of physical computation, the relationship between mathematical models of computation and their pretheoretic, informal counterparts, and skeptical challenges levied against the notion of physical computation.

11:00–12:30: Human–Robot Interaction (Room 300)

New Take on Robots Ethical by Design

Gordana Dodig Crnkovic *<**gordana.dodig-crnkovic@mdu.se**>*[1]
Baran Çürüklü *<**baran.curuklu@mdu.se**>*[2]
Jathoosh Thavarasa *<**jathoosh.thavarasa@efrei.net**>*[3]
Tobias Holstein *<**research@tobiasholstein.de**>*[2]
[1] School of Innovation, Design and Engineering, Mälardalen University, Sweden | Chalmers University of Technology, Sweden, Sweden, [2] School of Innovation, Design and Engineering, Mälardalen University, Sweden, Sweden, [3] Software Engineering, Efrei Paris, Villejuif, France, France

The paper explores the incorporation of ethics into the design of intelligent robots and softbots via artificial morality. As these agents increase in complexity, their ethical framework must accommodate their growing agency, autonomy, and intelligence. It is furthermore introducing the concept of functional moral responsibility, emphasizing the importance of a system-level approach throughout agents' lifecycle, integrating technical, human, and societal factors. The paper further promotes anticipatory ethics and speculative design, addressing the uncertainties of long-term development and their potential ethical implications.

Humanoid robots as socially disruptive technologies: three ways in which humanoid robots disrupt our human relational experiences

Cindy Friedman *<**c.friedman@uu.nl**>* Utrecht University, Netherlands

Socially disruptive technologies (SDTs) are technologies that "disrupt human knowledge and cognition, moral norms and values, as well as fundamental concepts and categories of thought" (Hopster 2021: 3-4)". My paper explains how humanoid robots can be understood as a SDT, by examining three ways in which humanoid robots particularly disrupt our human relational experiences: (1) the way in which we relate to technology; (2) the way in which we relate to other people; and (3) the way in which we relate to ourselves. In doing so, I also engage with why it is important that we anticipate the ways in which humanoid robots may be socially disruptive, as well as how we may go about responding to this potential disruption.

Is it wrong to kick Kickable 3.0? An affordance based approach to ethics of human-robot interaction

Arzu Formánek *<**arzu.formanek@univie.ac.at**>* University of Vienna, Austria

(By Bing Chat [GPT-4]) The moral status of robots is a contentious issue in human-robot interaction (HRI). Some argue that robots are indirect moral patients, meaning that they are morally relevant not for their own sake, but for the sake of other direct moral patients (humans, animals, environment etc). This paper challenges the common arguments for indirect robot moral patiency (iROMOP, as the author calls), which rely on anthropomorphism as the main explanation for human experience in HRI. It proposes an alternative account based on affordance theory, which captures the dynamic and interactive nature of human sociality and cognition. It illustrates this account with an imaginary robot, Kickable, which is designed for football training and affords physical kickability. It shows how an affordance based approach can better inform the normative debate about iROMOP by focusing on the emergent characteristics of HRI rather than on the isolated properties of robots or humans.

11:00–12:30: Explainability (Room 301)

Genuine Understanding or Mere Rationalizations? Approximations and Idealizations in Science and XAI

Luis Lopez *<**luis.lopez@philos.uni-hannover.de**>* Leibniz Universität Hannover, Germany

Rudin (2019) has prominently argued that post hoc XAI models are inherently misleading. Some philosophers of science (e.g., Fleisher, 2022) have been too quick to construe that these arguments

stem from a normative premise according to which perfect faithfulness, between (local) post hoc XAI models and their targets, is necessary for genuine understanding. Moreover, they have been even quicker in drawing insights from the literature on idealized scientific models to challenge such a premise. I show how this response not only mischaracterizes what is at the core of Rudin's arguments but also fails to distinguish idealization from approximation.

Mechanistic Computation and its Problems

Luke Kersten *<**lukekersten@edu.ulisboa.pt**>* University of Lisbon, Portugal

(By Bing Chat [GPT-4]) The mechanistic account of computation (MAC) faces three problems: the abstraction problem, the generality problem, and the hierarchy problem. These problems challenge MAC's ability to specify the conditions under which a physical system can be said to implement a computation. This paper defends MAC by appealing to an important distinction between abstraction and idealisation within computational explanations. It argues that computational descriptions within MAC are best seen as idealisations rather than abstractions, and that this view can resolve the three problems. The paper also compares the current proposal to a recent one by Kuokkanen (2022).

Understanding Mechanistic Explanations of Epistemically Opaque Deep Learning Systems

Marcin Rabiza *<**marcin.rabiza@gssr.edu.pl**>* Institute of Philosophy and Sociology, Polish Academy of Sciences, Poland

Epistemically opaque deep learning systems pose a challenge for explainable artificial intelligence (XAI), which aims at making their behavior understandable to stakeholders. This paper presents a mechanistic account of XAI inspired by the new mechanistic tradition in the philoso-phy of science and grounded in the concept of understanding as information compression. The paper argues that mechanistic AI explanations can foster trust, prediction, and manipulation by identifying the underlying mechanisms that give rise to the decision-making processes. The paper illustrates this approach with a case study of convolutional neural networks used for image recognition.

15:30–17:30: Minds and Machines SIG (Room 300)

Diffusing the Creator: Attributing Credit for Generative AI Outputs

Donal Khosrowi *<**donal.khosrowi@philos.uni-hannover.de**>*[2]
Finola Finn *<**finola.finn@gmail.com**>*[2]
Elinor Clark *<**elinor.clark2@gmail.com**>*[1]
[1] Leibniz University Hannover, United Kingdom, [2] Leibniz University Hannover, Germany

The recent wave of Generative AI (GAI) systems like StableDiffusion that can produce image and other outputs from human prompts raises a host of controversial issues about creatorship, originality, creativity, copyright, and others. This paper focuses on the creatorship question: who creates and should be credited with the outputs made with the help of GAI? Existing public and academic views on creatorship in this context are mixed: some insist that GAI systems are mere tools, and human prompters are creators proper; others are more open to acknowledging more significant

roles for GAI, but almost all conceive of creatorship in an all-or-nothing fashion. We develop a novel view, called CCC (collective-centred creation), that improves on these accounts. On CCC, GAI outputs are created by collectives in the first instance, and claims to membership in a co-creating collective come in degrees and depend on a set of finer-grained criteria that track the nature and significance of individual contributions made by various agents and entities, including users, GAI systems themselves, developers, producers of training data and others. Importantly, CCC maintains that GAI systems can sometimes have stronger claims to creatorship than humans. The paper demonstrates how CCC can advance existing debates and resolve controversies around creatorship involving GAI.

Long Abstract https://www.philos.uni-hannover.de/fileadmin/philos/Publikationen/DTC_preprint.pdf

How to Accelerate Ethics for Innovation and Against Precaution in Generative AI

James Brusseau *<**jbrusseau@pace.edu**>* Philosophy Department, Pace University-New York City, United States and Department of Information Engineering and Computer Science, University of Trento, Italy

One objection to conventional AI ethics is that it slows innovation. This presentation responds by reconfiguring ethics as an innovation accelerator. The critical elements develop from a contrast between Stability AI's Diffusion and OpenAI's Dall-E. By analyzing the divergent values underlying their opposed strategies for development, five elements of acceleration ethics are identified. 1) Uncertainty is understood as positive and encouraging. 2) Innovation is conceived as intrinsically valuable. 3) AI problems are solved by more AI. 4) Permissions and restrictions governing AI are decentralized. 5) The work of ethics is embedded in AI development and application. Together, these attitudes and practices remake ethics as provoking rather than restraining artificial intelligence.

Long Abstract ccc

From thinking to how to think

Nicola Angius *<**nicola.angius@unime.it**>*[1]
Alessio Plebe[1]
Pietro Perconti[1]
Alessandro Acciai[1] [1] Dep. Cognitive Science University Messina, Italy

(By Bing Chat [GPT-4]) The advent of language models that can converse with human beings has revived the philosophical debate on whether machines can think. The authors of this paper propose to shift the focus from the question "can a machine think?" to the question "how can a machine think?" They suggest that investigating the mechanisms of the Transformer architecture, which enabled the breakthrough in natural language processing, could shed light on how machines and humans grasp language. They also speculate on the possibility of finding parallels between the artificial Transformer and the brain language-memory unification.

Belief revision for language models

Thomas Hofweber *<**hofweber@unc.edu**>* University of North Carolina at Chapel Hill, United

States

It is arguable that language models represent the world, and have beliefs about the world. But if they do, some of these beliefs will be false, and should be revised. There are known methods for model editing that can change one belief at a time, but the question remains what larger changes should come with the revision of a single belief. This problem of belief revision and its connection to rationality is widely studied for human belief revision, but it remains unclear how much of the insights from that literature carries over to language models, and even whether language models are subject to the usual norms of rationality. For example, it is unclear how probabilities in language models relate to subjective probabilities relevant for belief revision. I will argue that there is a basic obstacle to applying the norms of rationality to language models since the basic standards of correctness associated with the internal states of a language model are aimed at the training data, but not at the world more broadly. This argument goes against both that there are beliefs in language models in the first place and that the norms of rationality apply to them. However, I will argue that this obstacle can be overcome, and we can make sense of subjective probabilities and the basic norms of rationality being applicable to language models just as to human believers.

15:30–17:30: Agency and Consciousness (Room 301)

Proxy Assertions and Agency: the case of machine-assertions

Chirag Arora *<**arorachirag@proton.me**>* University of Twente, Netherlands

This paper examines the rise of machine-generated speech, particularly in providing important information such as health advice through digital voice assistants like Alexa. The paper discusses the applicability of the concept of "assertion" in machine-generated speech and its implications for design and responsibility. Recently, some philosophers have argued in favor of the possibility of "machine assertion", based on their phenomenological and functional resemblance to human speech. In this paper, I challenge this functionalist notion of assertion. I argue for an alternative view where machine utterances are seen as proxy assertions, with responsibility lying with the designers who can reasonably foresee and take responsibility for the output. I also highlight the differences between my position and other existing arguments for proxy assertion by machines along with the implications of this perspective for designers and users of machines that produce human-like speech.

Thinking Without Phenomenal Character: How are Artificial Intelligence, Cognitive Phenomenology, and Extended Cognition Related?

Max Parks *<**maxaeon@umich.edu**>* University of Michigan, Flint (graduate student in Computer Science; previously earned PhD in philosophy at UC Davis), United States

This paper is intended to provide a rough overview of the relationship between artificial intelligence and cognitive phenomenology, in addition to extended cognition. Both cognitive states and phenomenal states are possible states had by humans, animals, and perhaps other objects, systems, or entities; but cognitive states and phenomenal states are distinguishable. On some views, machines merely need to act sufficiently intelligently to be intelligent, but on stronger views, machines must be phenomenally conscious to be thinking, intelligent creatures. Relatedly, many varieties of cognitive phenomenology maintain that thinking requires conscious experience, which (assuming

thinking and intelligence are coextensive) entails machines must be phenomenally conscious to think or to be intelligent. But such views about cognitive phenomenology unnecessarily complicate the relationship between cognition and phenomenal character, and the very nature of cognitive states. Moreover, according to extended accounts of cognition, cognitive states and phenomenal states may bear no necessary connection at all. As I will argue, it is possible and ontologically simpler to explain thought without appealing to proprietary or any other phenomenal character. On this ontologically simpler view of cognition, a machine can theoretically be a thinking thing without being in a phenomenal state.

Long Abstract https://doi.org/10.5281/zenodo.8030186

Is 'Responsible AI' Accountable?: A Critical Analysis

Riya Manna *<**riya_manna18@iitb.ac.in**>* Indian Institute of Technology Bombay, India

The paper explores the concept of 'responsible AI' from a posthumanist perspective that challenges the human-centric view of morality. It argues that humans and AI are not separate entities, but rather interconnected and interdependent parts of a larger 'cognitive assemblage' that transcends the boundaries of the biological body. It also questions the notion of 'responsibility' as a moral attribute that can be assigned to individual agents, and suggests that it should be understood as a distributed and relational phenomenon that emerges from the symbiosis of humans and AI in a specific context. The paper also analyses the problem of accountability in terms of 'responsible AI' and distinguishes between moral and legal accountability. The paper consists of two main sections: the first one examines the idea of 'collective rationality' as a way to describe the human-AI assemblage, drawing on the work of N. Katharine Hayles; the second one analyses the problem of 'responsibility'in relation to this assemblage, and proposes some possible ways to address it.

Moral Agency, Consciousness and Mental States

Zacharus Gudmunsen *<**prszg@leeds.ac.uk**>* University of Leeds, United Kingdom

(By Bing Chat [GPT-4]) The concept of moral agency is central to the ethical evaluation of artificial systems, but there is no consensus on what constitutes moral agency and whether it depends on having conscious mental states. This paper argues for an intermediate position between behaviour-focused accounts (B-accounts) and consciousness-focused accounts (C-accounts) of moral agency: mental states are necessary for moral agency, but conscious mental states are not. The paper shows that B-accounts and C-accounts have the same range of extension and that their differences lie in their ontological commitments and explanatory power. The paper proposes a deflationary account of mental states that can avoid the drawbacks of both B-accounts and C-accounts, while allowing for the use of mental state explanations in moral agency ascription. The paper concludes that artificial systems can have mental states and be closer to moral agency than we may have thought, even without consciousness.

15:30–17:30: Epistemic Problems (Room 325)

The theory and practice of computational errors

Nico Formánek *<**nico.formanek@hlrs.de**>* High Performance Computing Center Stuttgart,

Germany

This talk offers a taxonomy of computing errors based on the level of control we have over the computing method. Error control can be achieved by theoretical, mathematical means but also by practical engineering means. It is far from clear that theoretical error control is epistemically superior to its pragmatic cousin. To underline this point several successful and widely employed computing methods where only pragmatic error control is achieved are examined

Long Abstract https://doi.org/10.5281/zenodo.8010201

AI as an Epistemic Technology

Ramon Alvarado *<**ralvarad@uoregon.edu**>* University of Oregon, United States

(By Bing Chat [GPT-4]) Epistemic technologies are artifacts that are primarily designed, developed and deployed to manipulate epistemic content such as data through epistemic operations such as inferences, predictions or analysis in epistemic contexts such as inquiry. This paper argues that Artificial Intelligence (AI) and its related methods, such as machine learning (ML) and large language models (LLMs), are paradigmatic examples of epistemic technologies, and that they can be conceptually and practically distinguished from other technologies that are deployed in similar contexts but that do not manipulate similar content and do not carry out similar tasks. To support this claim, the paper examines the design intentions, the functional roles and the operational mechanisms of AI and other technologies in inquiry, and shows how they differ in a non-trivial manner. The paper also discusses some implications of recognizing AI as an epistemic technology for the philosophy of technology and the philosophy of science.

Disinformation and Epistemic Injustice: Implications for the Philosophy of Information and Ethics.

Steve McKinlay *<**stevet.mckinlay@gmail.com**>* Wellington Institute of Technology, New Zealand

(By Bing Chat [GPT-4]) Epistemic injustice and intellectual vice are two phenomena that affect the quality and reliability of information in the digital age. This paper examines how these phenomena relate to each other and to the philosophy of information and ethics. It questions whether intellectual vices such as dogmatism, gullibility, prejudice, closed-mindedness and negligence are epistemically significant and morally relevant. It also explores the implications of epistemic vice and disinformation for Floridi's distinction between true and false information. The paper aims to contribute to the understanding of the social and ethical dimensions of our epistemic practices in a data-driven world.

Tuesday, July 4

08:30–10:30: Minds and Machines SIG (Room 300)

Smooth Answers and Fabulations. Engaging with Processed Language

Leonie Möck *<**leonie.moeck@univie.ac.at**>*[1]
Sven Thomas *<**sven.thomas@uni-paderborn.de**>*[2]
[1] University of Vienna, Austria, [2] University of Paderborn, Germany

(By Bing Chat [GPT-4]) The recent release of ChatGPT, a powerful language model by OpenAI, has sparked societal debates about the consequences and threats of natural language processing (NLP) technologies. This submission investigates how the smoothness of the text produced by NLP models affects our epistemic practices and challenges our ability to critically engage with them. Drawing on the notion of fabulation from critical algorithm studies and critical epistemology, the submission explores how NLP models can be understood as processes of fabulation that combine speculation and fiction, and how we can resist their algorithmically pre-written futures by bringing back the noise and gaps in their outputs. The submission aims to motivate further research on the epistemic implications of linguistic smoothness and fabulation in the context of NLP technologies.

Gorgias-tp4: Socratic critique of rhetoric and language models

Stephen Rainey *<**stiofan.orian@gmail.com**>* TUDelft, Netherlands

In Plato's Gorgias, Socrates provides criticism of rhetoric as a merely persuasive art subordinate to argument. Among Socrates' problems with rhetoric is the idea that it is essentially inferior to rational discussion, because it is not interested in reasoned justification for arguments. Instead, rhetoric is primarily aimed at presenting persuasive speeches for or against given points of view. Where argument aims at truth, and is virtuous, rhetoric is interested in swaying the crowd to accept a position regardless of its truth. Rhetoric lacks virtue in this, in using flattery through pandering to an ignorant audience: "…the rhetorician need not know the truth about things; he has only to discover some way of persuading the ignorant that he has more knowledge than those who know…" (Gorgias). GPT4 and related language models produce their sometimes impressive results in a way akin to the model of rhetoric of which Socrates disapproved. The outputs from generative AI, whether text, poetry, code, or whatever, are produced not through a reasoned approach to some matter, but via a generic means of mimicry. Passages of text, for example, are produced from vast arrays of prior data such that they are inherently derivative in terms of their contents and bound to mimicry of their source material in form. This paper evaluates to what extent Socratic criticisms of rhetoric can help shape a response to GPT4 and the like, and to what extent any resemblance to rhetoric need be judged a good or a bad thing.

LLMs in academia: a thread or an opportunity?

Vojtěch Kolomý *<**vkolomy@unav.es**>* Universidad de Navarra, Spain

LLMs are already causing disruptions in academia, especially in research and student evaluation. However, I argue that while LLMs clearly pose threats to certain practices in academia, in the end they are only threats to undesirable practices. As far as the use of LLM in research is concerned, I argue that as LLMs do not properly understand, do not conceptualize, they can, on their own, only produce texts that seem to be written by human beings, but that are, at best, mediocre. So, the emergence of LLMs could serve as an opportunity to rethink the obligation for academic staff to continuously publish (see the famous lemma "publish or perish"). On the other hand, we should promote the general use of LLMs in those areas of research where they could save a lot of time. Regarding student evaluation, something similar occurs: LLMs' ability to produce text so similar to

those written by humans should lead us not to prohibit the use of these but, instead, to reconsider what kind of assignments we require of our students. In sum, I argue that rather than a threat to academia, LLMs actually represent an opportunity.

Measuring scientific understanding in Large Language Models

Kristian González Barman *<**KristianCampbell.GonzalezBarman@UGent.be**>*[1]
Henk de Regt[1]
Sascha Caron[1]
Tom Claassen[1]
[1] Radboud Universiteit, Netherlands

This paper presents a framework for measuring agents' scientific understanding of phenomena (where agents include humans and machine learning models). We focus on artificial understanding, i.e. whether machines (such as Large Language Models) can have scientific understanding.
Our starting point is De Regt's account of scientific understanding, which we extend into a framework for scientific understanding of agents in general. The framework considers three key aspects of understanding: knowing, explaining, and establishing counterfactual inferences. We show how these aspects can be measured using what-, why- and w-questions, respectively.
We provide recommendations for generating concrete tests and suggestions as to how the community should employ this framework to articulate a network of tests. These tests can serve a multitude of functions, e.g. contrasting the teaching abilities of different teachers, benchmarking models, adversarial training, and measuring student understanding.

Long Abstract https://arxiv.org/abs/2304.10327

08:30–10:30: Digital Governance (Room 301)

How to understand "user autonomy" for efficient platform regulation

Eloise Soulier *<**eloise.soulier@uni-hamburg.de**>* University of Hamburg, Germany

(By Bing Chat [GPT-4]) The concept of autonomy is central to many regulatory endeavors and ethical guidelines for complex digital technologies involving artificial intelligence (AI). However, classical Kantian-derived concepts of autonomy fail to account for the epistemic and practical dependencies of users on experts and technology. This paper critically examines these concepts and proposes a reconceptualization based on three critiques: the Freudian-Nietzschean critique of self-transparency and cognitive limitations, the feminist critique of relationality, and the philosophy of technology critique of technological infrastructure. The paper illustrates how this reconceptualization can inform regulatory work on two examples: the informed consent approach to data protection and the design of news recommender systems. The paper aims to show how conceptual engineering can contribute to the discussion and development of efficient platform regulation.

Notes on the Formative Power of Concepts: The Case of Digital Sovereignty and COVID-19

Gernot Rieder *<**gernot.rieder@uib.no**>* University of Bergen, Norway

In recent years, the notion of digital sovereignty has gained prominence in academic discourse, but

there is significant variation in how the term is used and understood. While some scholars adhere to a more traditional definition of sovereignty and view digital sovereignty as national control over digital phenomena, others have extended the scope of the concept to argue that there is no sovereignty in the digital world, but that different actors – states, companies, and individuals – exercise power as both claimants to sovereignty and attackers of the sovereignty claims of others. Examining and comparing research that has employed a digital sovereignty perspective to analyze technopolitical dynamics during COVID-19, this paper seeks to provide a better understanding of how differences in conceptualizing digital sovereignty can produce diverging accounts of the power dynamics at play, potentially resulting in contrasting evaluations of the current state of digital governance.

The Politics of Platform Governance

Laura Fichtner *<**laura.fichtner@uni-hamburg.de**>* University of Hamburg, Germany

Building on an analysis of the public controversy surrounding the Network Enforcement Act (NetzDG) in Germany, this paper explores how the governance of content moderation on social media platforms opens a space to renegotiate and reshape democracy. It outlines how contentions over this new law took place both against existing legal structures and political events and in relation to social media's technological affordances. The paper then traces different interpretations of NetzDG and its impacts back to liberal, deliberative, and (neo)republican conceptions of democracy. These can in turn each legitimate a particular distribution of rights and responsibilities between users/citizens, state institutions, and platforms. The paper therefore uncovers how the governance questions raised by social media platforms may alter the meaning of shared democratic values and even justify structural changes. It closes with a reflection on how this observation can constructively inform platform governance in the future.

08:30–10:30: AI Ethics (Room 325)

Exploratory Sandboxes and Experimental AI ethics: Performing Regulations over Prompts, Model Chaining, and Guidance programs

Denisa Reshef Kera *<**denisa.kera@gmail.com**>*[2]
Brian Ballsun-Stanton *<**brian.ballsun-stanton@mq.edu.au**>*[3]
Frantisek Kalvas[1]
[1] University of West Bohemia, Czechia, [2] Bar Ilan University, Israel, [3] Macquarie University, Australia

(By Bing Chat [GPT-4]) Experimental and participatory engagements over serious games, simulations, co-design experiments, and sandboxes are increasingly used in AI ethics and regulation. This paper examines the functions and implications of these engagements in current and emerging use cases of Large Language Models (LLMs). It pre-registers a study that will design a mock trial of a concrete ML application that has caused harm and collect data from different experimental and exploratory formats. The study aims to test two competing claims: one that these engagements only improve the understanding and trust in the technology and policy process, and another that they enable agency and negotiation among stakeholders despite epistemic uncertainties. The paper reflects on the role of epistemic certainty in decision making about disruptive new algorithmic services and the possibilities of minimalist and experimental approaches to AI ethics and governance.

Long Abstract https://doi.org/10.5281/zenodo.8003004

Humans and Machines: Challenges from Artificial Intelligence - Presenting a very recent report by the German Ethics Council

Judith Simon *<judith.simon@uni-hamburg.de>* University of Hamburg, Germany

(By Bing Chat [GPT-4]) This paper presents a report by the German Ethics Council on the ethical challenges of artificial intelligence (AI) in various domains of human life. The report examines the philosophical and anthropological foundations of human-machine relations and clarifies central notions such as intelligence, agency and responsibility. It also analyzes the effects of AI in the fields of medicine, education, public communication and public administration, and evaluates whether the delegation of actions to software systems leads to an increase or decrease in human agency and responsibility. The paper concludes with cross-cutting topics and overarching recommendations for addressing issues such as privacy, surveillance, bias, discrimination, transparency and accountability. The paper aims to provide ethical guidance to policy makers, industry and the public on the use of AI.

The case for e-vigilance: To what extent can and should we trust AI?

Rico Hauswald *<rico.hauswald@tu-dresden.de>* TU Dresden, Germany

(By Bing Chat [GPT-4]) This paper addresses the question of how to use artificial intelligence (AI) responsibly, by examining the issue of trust in AI systems. It argues that trust in AI can be conceptualized as an unquestioning attitude, which can be adopted towards both human and non-human entities, but that such an attitude is always risky and requires epistemic vigilance. It proposes the principle of "caveat usor": let the user beware, and distinguishes between direct and indirect forms of epistemic vigilance towards AI systems. It aims to contribute to the philosophical debate on the ethics of AI and to provide practical guidance for AI users.

08:30–10:30: Healthcare–1 (Room 326)

Conversational Artificial Intelligence used in psychotherapy as a special kind of cognitive artifact

J. P. Grodniewicz *<jedrzej.grodniewicz@uj.edu.pl>*[1]
Mateusz Hohol *<mateusz.hohol@uj.edu.pl>*[1] [1] Jagiellonian University in Krakow, Poland

Therapeutic Conversational Artificial Intelligence (CAI) is one of the most promising examples of the use of technology in psychiatry and mental health care. Most researchers agree that existing CAIs are not "digital therapists," and using them is not a substitute for psychotherapy delivered by a human. But if they are not therapists, what are they, and what role can they play in mental health care? To answer these questions, we appeal to the concept of cognitive artifact. Cognitive artifacts are artificial devices contributing functionally to the performance of a cognitive task. We argue that therapeutic CAIs are a special kind of cognitive artifact, which work by (i) simulating a therapeutic interaction and (ii) contributing to the performance of cognitive tasks which lead to positive therapeutic change. This characterization sheds new light on why almost all existing mental health chatbots implement principles of Cognitive Behavioral Therapy — a therapeutic orientation

according to which positive therapeutic change is mediated by cognitive change. Additionally, it allows us to grasp the limitations of further applying these technologies in therapy.

AI for psychiatry: close encounters of the algorithmic kind

Y J Erden *<**y.j.erden@utwente.nl**>* University of Twente, Netherlands

Psychiatry includes the assessment and diagnosis of illness and disorder within a largely interpersonal communicative structure involving physicians and patients. In such contexts, AI can help to spot patterns and generate predictions, e.g. 'big data' analysis via statistical learning-based models. In these ways, AI can help to automate more routine steps, improve efficiency, mitigate clinician bias, offer predictive potential, including through analysis of neuroscientific data. Electroencephalography (EEG), for instance, promises data on brain activity related to cognition, plus emotions and behaviour, as apparently objective accounts of what otherwise requires interpersonal engagement and observation. Yet psychiatric theories (including about emotion and behaviour) are not neutral, and any problematic assumptions in psychiatric theories, as well as discredited theories and retracted studies, can (and do) find their way into AI applications. AI can thus encode and reify such values and judgements. Even where research in psychology is sound, psychiatry is more than can be automated. AI analysis of big data for predictive purposes cannot supplant the phenomenological perspectives that underlie a person's actions, choices, and experiences, or bypass the necessarily discursive engagement between patient and clinician. Brain data can improve explanatory models, but this should not be at the expense of essential qualitative practices. Technological methods for assessment and diagnosis might seem time and cost efficient, but there remains an important role for (even imperfect) interpersonal methods in medicine and care. We therefore need some core principles for the appropriate use of AI in psychiatry. These include: (1) to not undermine necessary relational aspects of care; (2) to not cement simplistic classifications, exacerbate harmful biases, retain discredited theories or rely on retracted papers; (3) to not use brain data to bypass self-reporting and interpersonal, discursive methods; (4) to remain sufficiently transparent (methods, processes, data sets, including for training) and open to critique. In short, those who develop these technologies need to be aware of the complexity and necessary imprecision of the theories they adapt. Otherwise the scope for harm can be extensive.

In patients' interest? Ethical and social considerations of (intelligent) virtual agents in the healthcare sector

Catharina Rudschies *<**catharina.rudschies@uni-hamburg.de**>* Universität Hamburg, Germany

Technologies such as virtual agents have increasingly been studied in the healthcare context. They can, it is argued, provide new experiences for patients to interact with their healthcare professionals. Virtual agents based on Artificial Intelligence can act autonomously and hence replace doctors for certain tasks. However, introducing virtual agents in the healthcare context has ethical and social implications that need to be considered. For instance, replacing face-to-face contact with a human-machine interaction may pose problems to the therapeutic relationship. Trust and empathy are important elements of such a relationship. With the usage of intelligent virtual agents, the question is how empathy and trustworthiness of a virtual agent are perceived and whether they are morally justified. This talk will address ethical and social implications for the therapeutic relationship. It will use academic literature as well as empirical insights from qualitative interviews conducted with clinicians and patients on their views on using virtual agents in the healthcare

context, specifically in the field of psychotherapy and psychiatry. The interviews show *inter alia* that the expectations towards a human compared to a virtual doctor differ and may pose not only ethical but also practical problems for the implementation of the technologies.

11:00–12:30: Healthcare–2 (Room 300)

What kind of explanations are requested in the clinical deployment of AI to ensure patients' trust in medical decisions?

Anne Gerdes *<**gerdes@sdu.dk**>* University of Southern Denmark, Denmark

Transparency is commonly recognized as a crucial aspect of trustworthy AI, and explainable AI (XAI) is viewed as a potential solution to problems caused by algorithmic opacity. However, XAI is not the proper remedy for clinically deployed models, as a quest for explainability, given the current state of the art in XAI, cannot justify clinical decision-making. Furthermore, many aspects of medical decision-making and best practices in clinical care lack causal insights. Consequently, in the context of healthcare, rigorous empirical validation of models that enhance accuracy in medical decisions trumps explainability (Ghasemi et al. 2021, London 2019). While these observations question the value of XAI in healthcare, it is reasonable to assume that AI systems' lack of transparency may negatively affect patients' trust in healthcare. Against this setting, this paper explores what kind of explanations patients ought to have when confronted with medical decisions supported by AI.

Human sovereignty when a disease is controlled through restrictions on persons: Citizens' views on whether scientific evidence for restrictions is necessary

Ingvar Tjostheim *<**ingvar.tjostheim@nr.no**>*[2]
John A. Waterworth[1]
[1] Umeå University, Sweden, [2] Norwegian Computing Center, Hauge School of Management, Norway

The paper examines the attitudes of Norwegian citizens towards the need for scientific evidence in government-imposed restrictions during the COVID-19 pandemic. The abstract refers to the concept of "statist individualism" coined by Lars Trägårdh, which combines social welfare and equality with personal autonomy and individual freedom. The authors compare Norway's legally-binding restrictions with Sweden's voluntary approach and discuss how these approaches align with the concept of statist individualism. Despite Norway's low infection rate and death toll compared to Sweden and other countries, some citizens questioned the legitimacy of the travel restrictions imposed on their constitutional right to free movement. The Norwegian Supreme Court ruled in favour of the government, asserting that scientific evidence was not required for the restrictions and that the court should not assess their basis or effects. The paper presents findings from a national survey, indicating that the majority of respondents did not consider scientific evidence as strictly necessary for imposing restrictions, but to give information about disagreements (if any) among experts on the effects on the restrictions. These empirical findings are discussed in the context of personal autonomy, risk, and statist individualism, exploring differences.

Trust and Responsibility in Joint Human-AI Decision-making in Medicine

Vilius Dranseika *<**vilius.dranseika@uj.edu.pl**>* Jagiellonian University, Poland

(By Bing Chat [GPT-4]) How do people perceive and judge clinicians who use AI-based clinical decision support systems (CDSS) in medicine? This paper reports four studies (total N = 1220) that examine public reactions to cases of joint human-AI decision-making in health care, especially when harm occurs and recommendations conflict. The results suggest that people tend to trust and blame clinicians based on their alignment with the AI recommendation, which is seen as an epistemic superior. However, following the professional consensus is also considered a responsible and good choice, even when it leads to negative outcomes. The paper discusses the implications of these findings for the implementation and acceptance of AI in health care.

11:00–12:30: Symposium: Sketching Introductory Courses in the Philosophy of Computing (Room 301)

Sketching Introductory Courses in the Philosophy of Computing

Robin Hill *<**hill@uwyo.edu**>*[2]
Ramón Alvarado[1]
[1] University of Oregon, United States, [2] University of Wyoming, United States

We propose a symposium that will inspire, collect, and organize ideas on the content, approach, and perspective of introductory courses in the philosophy of computing for both graduate students and for undergraduate students of broad interests. Our objective is to solicit and assemble contributions for the consideration and use of IACAP members, including (1) extant reference materials, (2) specific lessons, (3) sketches of course syllabi, (4) suggestions for questions in prominent philosophical areas.
This project fosters opportunities to explore critical issues and extends the reach of academic investigations in computing. Computer science exhibits a deficit of interpretation. We professionals in computing assume that what we do (producing hardware and software) has no further connotation, that when we are done with the product, we are done. Modern students, however, experience and observe the effects of technology, the effects of the artifacts that we produce; we seek to complement and expand the formal picture delivered to them. Courses introducing such concepts within a solid framework of analytic philosophy will equip students to participate in the development of the field, and this symposium will help to equip interested scholars to teach those concepts.

11:00–12:30: AI Legislation (Room 325)

The Open Texture of 'Algorithm' in Legal Language

Davide Baldini *<**davide.baldini@unifi.it**>*[1]
Matteo De Benedetto *<**Matteo.DeBenedetto@ruhr-uni-bochum.de**>*[2]
[1] Florence University, Italy, [2] Ruhr-Universität Bochum, Germany

In this talk, we will survey the different, often contrasting, definitions of the term algorithm that can be found in contemporary legal practice and theory. We will do that by employing Friedrich Waismann's notion of open texture (cf. Waismann 1945). We will argue that the concept of algorithm, as currently used in legal practice and theory, exhibits a substantial degree of open texture, co-determined by the open texture of the concept of algorithm itself (cf. Shapiro 2006)

and by the open texture inherent to legal discourse (cf. Bix 2019, Stauer 2019). We will show how this substantial degree of open texture is not detrimental to good legal practice, but it is instead a positive feature of our legal language. We will substantiate our argument by virtue of a case study, in which we will analyze a recent jurisprudential case, where first and second-degree judges have carved-out contrasting notions of 'algorithm'. We will see how, thanks to our analysis of the open texture of algorithm in legal language, we can make sense of the different decisions taken by the two judges in our case study as different sharpenings of the concept of algorithm that were contextually determined trying to balance competing interests.

Artificial Intelligence, Radical Ignorance, and the Institutional Context of Consent

Etye Steinberg *<**etye.st@gmail.com**>* Philosophy, University of Haifa, Israel

More and more, we face AI-based products and services. Using these services often requires our explicit consent. Currently, AI operates by machine-learning or deep-learning. This means that the AI software evolves and changes its own modus operandi over time in such a way that we cannot know, at the moment of consent, what it is in the future to which we are now agreeing. Therefore, informed consent is impossible regarding AI. This means that we need to either come up with a new practice (other than consent), or revise our conception of informed consent (and its necessary conditions). As I argue, these two options are intertwined: under certain institutional autonomy-protecting conditions, consent can be valid and valuable without being informed. By understanding these institutional conditions, we can formulate practical solutions to foster valid and valuable, albeit imperfectly informed, consent across various decision contexts and within different institutions.

A Causal Analysis of Harm

Sander Beckers *<**srekcebrednas@gmail.com**>*[2]
Hana Chockler[3]
Joseph Halpern[1]
[1] Cornell University, United States, [2] University of Amsterdam, Netherlands, [3] King's College London, United Kingdom

As autonomous systems rapidly become ubiquitous, there is a growing need for a legal and regulatory framework that addresses when and how such a system harms someone. There have been several attempts within the philosophy literature to define harm, but none of them has proven capable of dealing with the many examples that have been presented, leading some to suggest that the notion of harm should be abandoned and "replaced by more well-behaved notions". As harm is generally something that is caused, most of these definitions have involved causality at some level. Yet surprisingly, none of them makes use of causal models and the definitions of actual causality that they can express. In our full paper we formally define a qualitative notion of harm that uses causal models and is based on a well-known definition of actual causality. The key features of our definition are that it is based on contrastive causation and uses a default utility to which the utility of actual outcomes is compared. We show that our definition is able to handle the examples from the literature, and illustrate its importance for reasoning about situations involving autonomous systems. Preprint available:

Long Abstract https://arxiv.org/abs/2210.05327

11:00–12:30: Simulation (Room 326)

All that glitters is not a deduction: Non-deductive methods in computational modelling

Michal Hladky *<**michal.hladky@gmail.com**>* University of Geneva, Switzerland

Computational modelling and simulations are often compared with experiments. It has been argued that these methods should be distinguished from experiments, that they are theoretical or that they require new epistemology. Many of these positions are based on the intuition that as these methods rely on computation which can be reconstructed as a series of deductive steps, the results they produce are conclusions of deductive arguments. Through a model-theoretical reconstruction of in silico experiments, I will demonstrate that the deductivist framework is not entirely adequate to capture their epistemology. Consequently, deduction is not an adequate criterion demarcating simulations from experiments.

Long Abstract https://doi.org/10.5281/zenodo.8034951

Deep Learning in Simulative Sciences

Nicola Angius *<**nicola.angius@unime.it**>*[1]
Alessio Plebe *<**alessio.plebe@unime.it**>*[1] [1] University of Messina, Italy

(By Bing Chat [GPT-4]) Deep Learning (DL) is increasingly used in science as a simulative method, but its epistemological and methodological status is unclear. This paper examines whether DL models can satisfy the formal relations between mathematical and computational models, artefacts, and target systems that are required for simulation. It also analyses the ontology of DL models and how they differ from common simulative models in terms of representational accuracy and tractability. Finally, it discusses the challenges of verification and validation for DL models, given their epistemic opacity.

15:30–17:30: The Ethics of Autonomous Military Robots SIG (Room 301)

Utopia's Killer Robots

Ahmed Amer *<**aamer@scu.edu**>* Santa Clara University, United States

(By Bing Chat [GPT-4]) The paper explores the ethical implications of autonomous weapons systems (AWS) or killer robots, and proposes a counter-intuitive idea that a world embracing the full engineering potential of killer robots could be better than a world that resists the thought. The paper argues that by rethinking the technology's manner of use, we might be able to find a reconcilable moral position for the problem of "dirty hands", or the deployment of lethal force at a distance from the deployer of that force. The paper envisions a hypothetical scenario in which AWS are used to reduce or eliminate human bloodshed in war, and to tie the fate of a leader to the fate of a robot. The paper suggests that such a scenario might satisfy the moral concerns of pacifists, realists, and just war theorists, and offer a hopeful alternative to an inevitable dystopian fate.

Responsibility gaps: old wine in new bottles

Ann-Katrien Oimann *<**ann-katrien.oimann@kuleuven.be**>*[1]
Fabio Tollon *<**fabiotollon@gmail.com**>*[2]
[1] KU Leuven, Belgium, [2] Bielefeld University/GRK 2073, Germany

(By Bing Chat [GPT-4]) The debate over the ethical use of Lethal Autonomous Weapon Systems (LAWS) hinges on the question of responsibility-gaps: whether it is possible to attribute responsibility for AI-based outcomes to anyone. This paper argues that this question reflects a deeper metaphysical disagreement about the nature of responsibility: whether it requires being responsible or being held responsible. The former view demands necessary and sufficient conditions for responsibility, while the latter view relies on socially determined criteria. The paper shows how different conceptions of responsibility underlie the arguments for and against the existence of responsibility-gaps in LAWS, and suggests a new way to understand and advance this debate.

AI as a threat to democracy across defence and civilian sectors

May Thorseth *<**may.thorseth@ntnu.no**>* Dept of Philosophy and Religious Studies, Norwegian University of Science and Technology, NTNU, Norway

Digital threats to democracy pose a serious challenge for both the defence and civilian sectors, as AI technologies enable disinformation, deep fakes, surveillance capitalism and behavioural surplus. The presentation examines the possibility and desirability of using the same technologies to counter these threats, and raises ethical and political questions about the symmetry between attack and defence, the blurring of military and civilian contexts, and the concentration of power in Big Tech companies. The presentation argues that the solution to the democracy problem is most likely non-technological, and that we should avoid undermining what we want to defend by responding with the same coin.

Bombs, Bots, and the Principle of Distinction

Nathan Wood *<**Nathan.Wood@UGent.be**>* Ghent University/California Polytechnic State University San Luis Obispo, Belgium/USA

In many debates on autonomous weapon systems (AWS), critics argue that these will likely not be able to carry out the difficult task of distinguishing between legitimate targets and those protected from attack. This is argued to therefore render AWS in violation of the principle of distinction, which requires that combatants "not make civilians the object of attack" and not carry out attacks which are "indiscriminate in nature". However, this objection both mistakes what the principle of distinction in fact demands, and ignores important aspects of how AWS are being developed, under what limitations they operate, and how they are deployed. In this article I show that the critique from distinction relies on an overly and inaccurately broad picture of what AWS are, that it holds AWS to an inappropriately high standard, and one which is *ad hoc* in relation to the standards to which combatants and other weapons are held, and I show that it fundamentally misunderstands the principle of distinction as it is formulated in the Law of Armed Conflict (LOAC). I conclude by arguing that the limitations of AWS to which critics point do not actually underpin any problem with such systems, but in fact highlight an impressive feat of technological design and a further step on our long road to making warfare a less brutal and bloody enterprise.

15:30–17:30: Epistemology (Room 325)

The Complementary Minds: A Logical Framework for Belief Formation

Dimana Orlinova Anastassova[1]
Massimiliano Badino[1]
Fabio Aurelio D'Asaro <***fabioaurelio.dasaro@univr.it***>[1] [1] University of Verona, Italy

(By Bing Chat [GPT-4]) We develop a logical framework for modeling the default-interventionist architecture of belief formation, which assumes that our mind consists of two types of processes: Type 1 (T1), which produces fast and autonomous responses to external evidence, and Type 2 (T2), which can check and revise the default responses using working memory and cognitive decoupling. We suggest that Depth-Bounded Boolean Logics (DBBLs) are natural candidates to represent this architecture, as they provide tractable approximations of classical logic that can account for the limited resources and non-monotonic nature of belief formation. We also outline some future directions for developing and validating this framework in relation to empirical and theoretical phenomena observed in practical reasoning.

Probability Space for First-Order Predicate Logic

Kenneth Presting <***kapresti@ncsu.edu***> North Carolina State University, United States

In the spirit of Tarski's 1930 work, "The concept of Truth in Formalized languages," we define probability in terms of satisfaction. The novelty here consists in giving quantified sentences an extension, rather than a truth-value, in the domain of a specific first-order model. Quantified sentences are interpreted as defining events in a probability space, constructed from a distinctive first-order model. In standard models, sentences (closed quantified formulas) are true or false according to their satisfaction either by all sequences, or by none. In the extended models here, bound variables in quantified expressions can be satisfied with selected subsets of the domain, rather than the whole domain.

Long Abstract https://doi.org/10.5281/zenodo.8011322

Understanding the analogical roots of Agent-Based Models in economics and social sciences

Massimo Rusconi <***m.rusconi8@studenti.uninsubria.it***>[2]
Davide Secchi[3]
Raffaello Seri[1]
[1] Department of Economics, University of Insubria, Italy, [2] University of Insubria, Italy, [3] Centre for Computational & Organizational Cognition (CORG), University of Southern Denmark, Denmark

The adoption of Agent-Based modeling (ABM) has become ubiquitous in recent years in Economics to investigate hidden and emergent behaviors of complex systems. However, most of the everyday research activities rely on the researchers' consensus concerning practical choices about modeling strategies, computational boundaries under scrutiny, and the extent of empirical validation. Thus, the specific meta-theoretical reflection on Agent-Based modeling techniques is often overlooked. This paper offers a meta-theoretical evaluation of Agent-Based Modeling (ABM) in the economic and social sciences, aiming to understand the methodological and epistemic role of abstracted

and generalized models. The authors analyze the cognitive dimension of model development, the functional orientation of ABMs, and their impact on modeling strategies within a framework of iterated analogical inferences. The process of ABM construction is seen as a two-stage recursive level of multiple iterated analogical arguments. The paper justifies established pragmatic rules of thumb in the ABM literature and compares them with existing validation techniques. This theoretical perspective proves to be extremely valuable in explaining practitioners' epistemological choices, by embracing together pragmatic and structural positions.

The Meta-Ontology of AI systems AI

Roman Krzanowski *<**rmkrzan@gmail.com**>*[1]
Pawel Polak[1] [1] UPJP2, Poland

In this paper, we examine the meta-ontology of AI systems with human-level intelligence, with us denoting such AI systems as AIE. Meta-ontology in philosophy is a discourse centered on ontology, ontological commitment, and the truth condition of ontological theories. We therefore discuss how meta-ontology is conceptualized for AIE systems. We posit that the meta-ontology of AIE systems is not concerned with computational representations of reality in the form of structures, data constructs, or computational concepts, while the ontological commitment of AIE systems is directed toward what exists in the outside world. Furthermore, the truth condition of the ontology (which is meta-ontological assumption) of AIE systems does not require consistency with closed conceptual schema or ontological theories but rather with reality, or in other words, "what is the world" (Smith 2018:57). In addition, the truth condition of AIE systems is verified through operational success rather than by coherence with theories. This work builds on ontological postulates about AI systems that were formulated by Brian Caldwell Smith (2018).

15:30–17:30: Ethics and Applications (Room 326)

Experiencing AI and the Relational 'Turn' in AI Ethics

Jason Branford *<**jason.branford@uni-hamburg.de**>* University of Hamburg, Germany

AI Ethics is a growing field that examines the ethical implications of artificial intelligence technologies. However, many existing approaches to AI Ethics are limited in that they fail to account for both the relational character of human life and the co-constitutive role of technology in our lived experience. This submission proposes to bolster recent attempts to incorporate such relational elements into AI Ethics by drawing on insights from postphenomenology and Deweyan pragmatism. It aims to provide the groundwork for a Relational AI Ethics that is attuned to the fluid and dynamic nature of human-technology relations, and that is capable of facilitating moral inquiry and the reconstructive action needed for moral progress.

A revision and extension of List & Goodin's model of epistemic democracy

Pablo Rivas-Robledo *<**pablo_rivas_robledo@hotmail.com**>* University of Genoa, Italy

List & Goodin presented their extension of the Condorcet Jury Theorem to scenarios with more than two options and with other voting rules. The present paper presents a computational improvement and a philosophical reevaluation of the original model. The improvement is simple: we can now

go far beyond the original results by analyzing more complex cases than those considered in the original model, either by computing the values directly or by approximating them. The philosophical reevaluation is as follows: Originally, List & Goodin argued that all voting systems studied were similarly good truth trackers because they performed similarly well under the same circumstances. Here we argue that a far better criterion is to assess at what point a group of agents become practically infallible: a voting rule is better if it can make a group practically infallible by adding fewer agents to the group. According to a preliminary analysis, Borda count seems to perform best under this criterion.

War or peace between humanity and artificial intelligence

Wolfhart Totschnig *<**wolfhart.totschnig@mail.udp.cl**>* Universidad Diego Portales, Chile

The thinkers who have reflected on the potential risks of a future artificial general intelligence (AGI) have focused on the possibility that the AGI might carry out its assigned objective in a way that we did not anticipate, with potentially catastrophic effects (Yudkowsky, Bostrom, Omohundro, Yampolskiy, Tegmark, Russell). They have neglected the possibility that the AGI could come to see us as a threat to its existence and, therefore, deliberately try to eliminate us. The aim of the present paper is to show that this neglect is mistaken. I will describe a possible situation where an AGI and humanity find themselves vulnerable vis-à-vis each other, which could lead to an all-out war. I will then argue that, in view of this possibility, the approach of the said thinkers, which is to search for ways to keep an AGI under control, is potentially counterproductive because it might, in the end, bring about the existential catastrophe that it is meant to prevent.

Wednesday, July 5

09:00–10:30: AI Impact (Room 300)

Digital Transformations of Democracy: How to Successfully Solve Shared Problems in the Age of Anthropocene

Jan-Philipp Kruse *<**jan-philipp.kruse@uni-hamburg.de**>* University of Hamburg, Germany

This approach aims at linking the debate on digital transformations of democracy with the one on multiple crises and existential societal problems in Anthropocene, focusing on ecological crises that arise alongside climate change. Ecological crises do raise the bar for democratic deliberation, as they pose inelastic (irreversible) problems, while the digitally transforming structure of democratic public spheres is more and more considered to be problematic in itself. Relating these two crucial angles of transformation, the question arises which requirements a digitizing public sphere must be able to meet under the conditions of Anthropocene. With recourse to Kant's Critique of Judgment, the notions of Decontiguated Public Sphere and Judgment Environment are introduced to capture the characteristics of digitizing public sphere. From there, I will discuss how successful (i.e., adaptive and normatively viable) democratic processes could be fostered in the context of digital transformations.

Deepfakes, Brain Emulation, and Things We Are: Asserting Rights of Origination

Jeremy Hansen *<**jeremyhansen@acm.org**>* Norwich University, United States

This work introduces a theory of a category of inalienable legal rights of *origination* which encompasses many currently unconnected rights like privacy, bodily integrity, and publicity. Rights of origination are based on the premise that a natural person has a unique identity that originates in and is inherently tied to their physical body or actions, and unconnected to tangible property or intangible but otherwise *separable* things like intellectual property. The legal term used to describe the various intangible parts that make up a person's identity is *indicia of identity* which under US law include attributes like name, nickname, likeness, voice, signature, photographs, performing style, and distinguishing actions. While rights of origination cover these indicia of identity, the broader *indicia of origination* also include (for example) biometric attributes like fingerprints, DNA, or tissue samples. Our purpose here is to sketch a general outline of this category to better connect and protect these personality- and identity-based rights in light of current (deepfakes) and future (brain emulation) technological changes where current law may not adequately or consistently address threats to human rights.

Long Abstract https://doi.org/10.5281/zenodo.8015393

Ethical Issues in Generating Robot Nudgers

Stefano Calboli *<**calbolistefano@gmail.com**>* Centre for Ethics, Politics, and Society - University of Minho, Italy

Nudges meant to directly change choice environments based on intuitive thinking processes (System-1) have become well-established practices in policy and marketing. However, the use of social robots for nudging is not as established. This article treasures the ethics of traditional nudges, focusing on decisional autonomy and public scrutiny with the view to narrow down the attention to social robots. The article has two main objectives. First, it examines decisional autonomy and public scrutiny in the context of social robots being used by public nudgers, private nudgers, and nudgees. Secondly, it delves into the potential for personalizing safeguards to protect decisional autonomy and public scrutiny through social robot nudgers, which presents unique ethical considerations compared to traditional nudges.

09:00–10:30: Symposium: Turing and Ashby on Computation, Mechanisms and Intelligence (Room 301)

Turing and Ashby on Computation, Mechanisms and Intelligence

Hajo Greif *<**hans-joachim.greif@pw.edu.pl**>*[1]
Adam Kubiak[1]
Paula Quinon[1]
Paweł Stacewicz[1] [1] Warsaw University of Technology, Poland

This symposium explores the historical and systematic connections between the two main traditions of cognitive inquiry: the computational paradigm initiated by Alan M. Turing and the cybernetic paradigm as developed by W. Ross Ashby (*inter alia*). The symposium challenges the common view that these paradigms are diametrically opposed, and shows how they share some common ground besides numerous subtle differences in their respective conceptions of computation, mechanism

and intelligence. We primarily address questions concerning the scope, limits and relevance of analog computation, multiple realisability and universality in both paradigms.

Paper 1: Analog Computation: Continuous vs Empirical

This paper discusses two different ways of understanding analog computation: analog-continuous computation (AN-C) and analog-empirical computation (AN-E). AN-C computation consists of processing continuous data that are characterized by real numbers from some continuous domain, and shares Turing's claim for one universal model of computation. AN-E computation consists of implementing dedicated physical processes that are natural analogues of certain mathematical operations. Being more domain-specific and theory-dependent, it is characteristic of cybernetic models. We argue that both types of analog computation are strongly related to the natural sciences, not just the formal sciences.

Paper 2: The Origin of Adaptation and the Mechanisms of Cybernetics

This paper explores the premises of W. Ross Ashby's account of the 'Origin of Adaptation' (1941). His approach to cybernetics was unique in that he relied on a mechanistic and deterministic interpretation of Darwinian evolution, which he then developed into a material and analog model of adaptive behaviour in the nervous system, the 'homeostat' (1952). His ultimate aim was to formulate an axiomatic theory of the emergence of goal-directed organisation in all kinds of self-organising systems that relied on mathematically formulated analogues of random variation and natural selection. We contrast Ashby's empirical hypothesis about mechanisms in nature with Turing's computational notion of mechanism, which concerned an 'effective method' of calculation.

Paper 3: The Development of Intelligent Machines: Common Themes in Turing and Ashby

This paper highlights some commonalities between Turing's and Ashby's accounts of machines that develop forms of intelligence. It focuses on Turing's (1948) proto-connectionist model as his most realistic take on natural intelligence, and on Ashby's model of adaptation that he developed between 1941 and 1952/1960. We highlight common features of their models such as reductionism, multiple realisability, universality, unpredictability and autonomy. We also show how Turing's model can be represented or conceptualised using Ashby's model. We conclude that their models' respective mapping onto digital versus analog modes of computation is less than straightforward.

09:00–10:30: Data (Room 325)

The Colour of Offline Data

Brian Ballsun-Stanton *<**brian.ballsun-stanton@mq.edu.au**>* Macquarie University, Australia

The colour of offline data is a human-evaluable context that influences the quality and future uses of data collected in remote environments. Drawing on examples from archaeological fieldwork, I argue that software systems for offline or remote data collection should be able to handle the context and state of data, its "colour", as it passes through the research pipeline. I also discuss how data structures can anticipate *some* aspects of desired metadata, but never *all* aspects within

reasonable time and complexity constraints. This requirement of tacit and unanticipated metadata has implications for software design for offline and research systems.

Long Abstract https://zenodo.org/record/8001384

Topical alignment with the New York Times boosts content popularity on the social web

Kate Barnes *<**kate.barnes@coloradocollege.edu**>*[1]
Roland Molontay[1] [1] Budapest University of Technology and Economics, Hungary

How does topicality influence the popularity of online content, such as memes and tweets? Using Latent Dirichlet Allocation, this research extracts 100 topics from five years of publications from the New York Times (NYT). We observe impressive alignment between certain NYT and Reddit topic distributions, while other topics differ between the two media sources. Additionally, social media content is more likely to be popular when its topic is also prevalent in the NYT. Our methods show how exploratory data science and machine learning techniques can complement theoretical work. Can human norms be visualized using data? The implications of our findings are discussed in relation to the feedback loop between content popularity and visibility, and the epistemological dominance of norms.

Ethics of Metadata Ownership

Navid Shaghaghi *<**nshaghaghi@scu.edu**>*[1]
Maria Joseph Israel[1] [1] Santa Clara University, United States

The concept of ownership is fundamentally irreconcilable with the nature of metadata such as data indices. Much scholarly work has been done to determine the ownership of data, however, metadata ownership has conveniently escaped the same level of attention and scrutiny. Since the ownership of metadata provides no insight into the ownership of that metadata's metadata, ad infinitum, it is impossible to assign true ownership to any metadata. Furthermore, from the facts that 1) indices that limit access to the data they index by placing an index behind forced advertisement viewership, surrendering of user data, or expensive paywalls, are inherently oxymoronic because indices don't derive value from the data they index, but rather from how well they increase accessibility to the data they index. And 2) since a data index assigns some sort of order or even priority to the data it indexes (in the case of search engines for example), it devalues data it does not index, reduces the value of the data which it indexes but provides lower priority to, and artificially inflates the value of the data it prioritizes in the order/hierarchy of its returned result. And 3) that proprietary indices can not easily be amended or corrected due to possessing the preferences and biases of the indices' creators/owners towards the data being indexed. It follows that assigning ownership to data indices inherently hurts the usefulness of the data they index as well as the related data they don't index. This effect on the usefulness of what an index indexes, necessitates the assignment of (at the very least) the same level of access/ownership to an index as the data it indexes, so that any corrections, inclusions, or prioritizations can be administered and vetted at the appropriate community level and not a single company or individual level. Which for publicly available data, would mean the public ownership of any index, indexing that publicly available data.

09:00–10:30: Decision-making (Room 326)

AI decision-making and the problem of value-alignment

Björn Lundgren *<**b.a.lundgren@uu.nl**>* Utrecht University, Netherlands

In this paper I will talk about the problem of value-alignment and control of AI. I will start by critically assessing a solution proposed by Stuart Russell. After that I will turn to re-diagnose the problem and attempt to provide a pathway to a solution.

Paternalistic machines

Kalle Grill *<**kalle.grill@umu.se**>* Umeå University, Sweden

Human interests and preferences do not always align, which poses a challenge for technical systems that aim to be human-centered in a wide sense. While benevolent AI systems may prioritize human interests, they need not respect individual choice or preference. Some philosophical accounts of paternalism, focusing on the inner life of the paternalizing agent, offer limited applicability to the risk of paternalistic machines, since it is unclear in what sense such technology has an inner life. However, other accounts of paternalism are more promising and can help us identify more subtle risks than those from non-benevolent AI. At the same time, the philosophical literature on paternalism may learn from the difference in applicability of different theories to the case of artificial agents.

Understanding over explanation: an epistemic rights perspective on XAI

Yeji Streppel *<**y.j.m.b.k.streppel@tue.nl**>* Technical University Eindhoven, Netherlands

A central problem in AI ethics is how to protect people against harmful decisions of opaque Automated Decision Support systems (ADSs). The received view is that people have a (legal and epistemic) right to explanation: they can demand explanations of what opaque ADSs do and how they work. However, there is strong disagreement in the explainability literature over what (good) explanations are. This paper proposes a novel perspective by shifting the discussion from the right to explanation to the epistemic right to understanding, where understanding is characterized as the ability to answer "what if"-questions. Anyone who is subject to decision-making institutions has an epistemic right to understanding. This right is currently widely violated, causing considerable harm.

13:30–15:00: Information (Room 300)

Continental Approches to Philosophy of Information

Ashley Woodward *<**ash.woodward@hotmail.com**>* University of Dundee, United Kingdom

(By Bing Chat [GPT-4]) Philosophy of information (PI) is a field that transcends the Analytic/Continental divide, but remains largely influenced by the analytic tradition. This paper introduces continental European approaches to PI, both critical and constructive, that have been under-appreciated in

the field. The critical approaches include phenomenological and poststructuralist critiques of information as a reduction of meaning and being. The constructive approaches include theories of information as transformation developed by Raymond Ruyer, Gilbert Simondon, and Michel Serres. The paper argues that these continental perspectives can enrich and expand the field of PI by addressing broader cultural, ethical, and political issues raised by the information revolution.

An Inquiry Concerning the Persistence of Information

Roman Krzanowski *<**rmkrzan@gmail.com**>* UPJP2, Poland

Physical information is a property of nature. How does physical information persist over time? Does it do so as an object, process, or event, which are things considered in the current persistence theories? Physical information is none of these, however, so persistence theories cannot explain the persistence of information. We therefore study the persistence of snowflakes, ephemeral natural structures, to better understand the persistence of natural things, such as physical information. The transitory nature of snowflakes suggests that physical information persists as nature's latent order, so it is associated with natural structures, but it is not identical to them. This interpretation preserves the properties attributed to physical information, particularly its foundational character. The concept of physical information as latent order accords with Burgin's General Theory of Information (GTI), which is currently the most comprehensive conceptualization of information that has been proposed.

13:30–15:00: Minds (Room 301)

Exploring and Understanding the Role of Motivation in Computational Models of the Mind

Ron Sun *<**dr.ron.sun@gmail.com**>* RPI, United States

Motivation is a crucially important aspect of the human mind, but it has been downplayed or ignored in computational models of the mind, especially in computational cognitive architectures. I argue that motivation needs to be more extensively explored and better understood, especially in and through computational models of the mind. I propose to develop better processes and representations of motivation, especially based on empirical and theoretical work on intrinsic human motivation. Computational modeling of motivation and its interaction with cognition can be leveraged to better understand the role of motivation in psychological functioning. I show how a computational cognitive architecture (named Clarion) can account for empirical phenomena across a wide range of domains, based on intrinsic needs/motives, utility calculation, and their effects on cognitive processes.

Long Abstract https://doi.org/10.5281/zenodo.8011738

Illusionism and Machine Consciousness

Tomas Hribek *<**tomas_hribek@hotmail.com**>* Czech Academy of Sciences, Czechia

(By Bing Chat [GPT-4]) The Hard Problem of AI Consciousness (HPAIC) asks whether artificial general intelligence (AGI) systems will have phenomenal consciousness, i.e., subjective experiences that

feel a certain way from the inside. This paper surveys several approaches to HPAIC and offers an alternative based on illusionism, the view that phenomenal consciousness is an illusion generated by cognitive mechanisms. According to illusionism, there are no phenomenal states in humans or machines, only quasi-phenomenal states that create the impression of having such states. The paper argues that this view liberates AI and cognitive science from a pseudo-problem and allows them to focus on designing functional and computational capacities that emulate conscious behavior. The paper also discusses some examples of artificial consciousness research programs compatible with illusionism.

13:30–15:00: Digital Platforms (Room 325)

Digital Recording and the Hazards of Unbounded Moral Judgment

Bart Kamphorst *<**bart.kamphorst@wur.nl**>*[1]
Elizabeth O'Neill *<**e.r.h.oneill@tue.nl**>*[2]
[1] Wageningen University, Netherlands, [2] Eindhoven University of Technology, Netherlands

The widespread adoption of internet-connected, camera-fitted devices, combined with the presently available digital infrastructure, has facilitated a techno-social environment in which it is viable to make, store, alter, and share digital recordings such as photographs, audio fragments, and video streams at an unprecedented scale. We contend that today's digital recording practices threaten to radically alter how we perceive and evaluate ourselves and others, producing an ongoing, socially and morally disruptive shift towards unbounded moral judgment. We argue further that the trend toward unbounded moral judgment poses several hazards, including widespread, difficult-to-restore reputation damage, negatively altered self-perceptions, and even the stifling of morally right behavior. With an eye to how the associated technologies are projected to advance in the future—e.g. growing use of deep fakes and augmented reality—we provide recommendations about technical, regulatory, societal, and individual approaches to mitigating these issues.

Skilled agency and illegitimate technological control for Social Media users

Lavinia Marin *<**L.marin@tudelft.nl**>* Delft University of Technology, Netherlands

The aim of this paper is to offer an alternative conceptualization of moral agency and control in a robust way that can be used to evaluate interactive technologies' influence on human agents. I provide a refined notion of situated agency bassed on 4E cognition defined as agency as visible in how the agent adapts to an environment in its behaviour but also in how the goals reflect this adaptation. All moral agents find themselves situated in particular environments. Other agents will always influence how they develop their skills and what they find as worthwhile goals (social situatedness). In addition, in digital environments, the other agents are inferred, and their actions are sometimes indistinguishable from algorithmic actions. In interacting with a digital interface, the user has to compensate for the lack of bodily cues (for example signifying emotions with emojis) and has to infer the social normativity (what is considered valuable and or morally right) from the mediated cues of other unseen agents.

Practical Information

Conference venue

Location: The conference will be held in the building of the Faculty of Arts, Charles University (pictured above).

Address: náměstí Jana Palacha 1/2, 116 38 Prague

Coordinates: 50.088973 N, 14.415859 E

The faculty building is situated in the city centre, within 5-minute walking distance from the Charles Bridge ("K" on map on opposite page [43]) and from the Old Town Square ("S"). The conference rooms and halls were recently renovated, all being equipped with air conditioning and modern projecting equipment.

Most of the rooms benefit from a marvelous view of Prague Castle.

Accessibility: The historical building is accessible for people with disabilities.

Directions

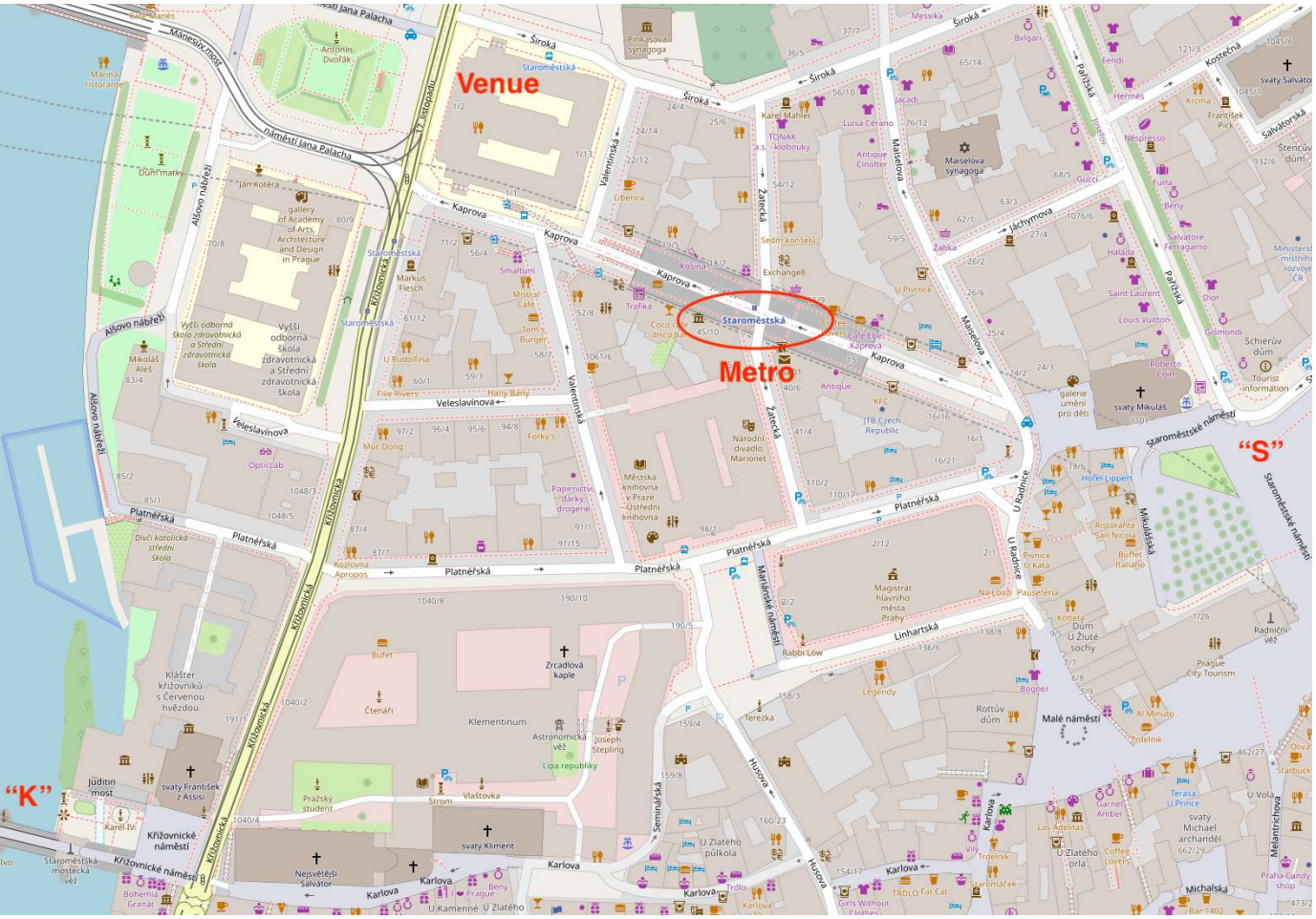
Public transport: The venue is located next to the “Staroměstská” metro station, served by:

- **Metro:** line A (green line)
- **Tram:** lines 2, 17, 18
- **Bus:** lines 194, 207 and night line 93

From Vaclav Havel Airport Prague: bus line 119 to “Nádraží Veleslavin” station, change for metro line A (= green, direction “Depo Hostivař”) to “Staroměstská” station.

From Prague Main Railway Station: metro line C (= red, direction “Háje”) to “Muzeum” station (one stop), change for line A (= green, direction “Nemocnice Motol”) to “Staroměstská” station.

Official website of the Prague public transport agency: https://www.dpp.cz/en



Top: Location of venue, ©Open Street Map.

Conference dinner

Location: The conference dinner will be held at the Pivnice Municipal House (Pivnice Obecní Dům, pictured below), náměstí Republiky 1090/5, 110 00 Prague.

Top: Prague Metro, ©Zirland. **Bottom:** Municipal House, ©Thomas Ledl, both CC BY-SA 4.0

Organisers

Programme Committee

* Ahmed Amer (Santa Clara University)
* Alexis Elder (University of Minnesota)
* Arzu Formanek (University of Vienna)
* Björn Lundgren (Utrecht University)
* Brian Ballsun-Stanton (Macquarie University)
* David Černý (Czech Academy of Sciences)
* Giorgia Pozzi (TU Delft)
* Hajo Greif (Warsaw University of Technology)
* John Licato (University of South Florida)
* Juan M Durán (Delft University of Technology)
* Maria Joseph Israel (Santa Clara University)
* Nico Formanek (HLRS)
* Ramon Alvarado (University of Oregon)
* Robin Hill (University of Wyoming)
* Steve McKinlay (Wellington Institute of Technology)
* Susan Kennedy (Santa Clara University)
* Thomas Powers (University of Delaware)
* Tomas Hribek (Philosophy Institute, Prague)

International Association of Computing and Philosophy
www.iacap.org

Institute of Philosophy
Filosofický ústav AV ČR

The Karel Čapek Center
for Values in Science and Technology

DEPARTMENT OF PHILOSOPHY
AND RELIGIOUS STUDIES
Faculty of Arts
Charles University

Support

This conference was only possible through the ongoing memberships and support, year by year and lifetime, of the IACAP Members.

The conference has also received financial support from the program Strategie AV21 Breakthrough Technologies for the Future – Sensing, Digitisation, Artificial Intelligence and Quantum Technologies.

Conference booklet printing made possible by a generous donation from the High-Performance Computing Center Stuttgart.

INSTITUTE OF COMPUTER SCIENCE
The Czech Academy of Sciences

Czech Academy of Sciences

STRATEGY AV21
Top research in the public interest

HLRS
High-Performance Computing Center Stuttgart

Auletris
event management services