AI, HUMAN VALUES AND MEANINGFUL HUMAN CONTROL

22-23 June | CST Bonn

ABSTRACTS BOOKLET

ZERTIFIZIERTE KI
Qualität sichern. Fortschritt gestalten.

CENTER FOR SCIENCE & THOUGHT

UNIVERSITÄT BONN

1. JOCELYN MACLURE

AI, Cognitive Scaffolding and Social Reasoning

Jocelyn Maclure is Full Professor of Philosophy at McGill University. He holds the Stephan A. Jarislowky Chair in Human Nature and Technology. He currently works on both speculative questions related to artificial general intelligence, artificial consciousness and moral status, and on practical issues in the ethics of AI.
His publications on AI include "The New AI Spring: A Deflationary View" (AI & Society); "AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind" (Minds & Machines); "AI For Humanity: The Global Challenges" (with Stuart Russell, in Towards Responsible AI); "Intelligence artificielle, automatisation et inégalités" (with D. Rocheleau-Houle in L'intelligence artificielle and les mondes du travail) and "AI's Fairness Problem: Understanding Wrongful Discrimination in the Context of Automated Decision-Making" (with H. Cossette-Lefebvre, under review).
As the President of the Quebec Ethics in Science and Technology Commission, he supervised the publication of several reports on the ethical and legal regulation of AI technologies. He was a member of the scientific committee of the Montreal Declaration for Responsible AI (2017) and of the Global Forum on AI for Humanity (Paris, 2019). He was part of the Canadian delegation at the UNESCO Intergovernmental Meeting of Experts related to a Draft Recommendation on the Ethics of Artificial Intelligence, 2021.
Before turning to artificial intelligence, he published extensively on value pluralism, public reason, secularism, multiculturalism and human rights. Secularism and Freedom of Conscience (Harvard University Press, 2011), co-authored with Charles Taylor, appeared in several languages.

2. THEORETICAL FRAMEWORK

Claudia Schon, Werner Moskopp

Towards Benchmarking Moral Decisions in AI Systems

In this paper we are going to point out some arguments about the task of measuring moral capabilities of artificial intelligence (AI) systems. In the field of AI, so-called benchmarks are commonly used to evaluate systems.
These benchmarks often take the form of multiple-choice tasks that the AI system is supposed to solve. Following the reasoning of the articles of LaCroix & Luccioni (2021, 2022) we want to analyze some problems of benchmarks in ethical AI that depend on metaethical presuppositions of authors, programmers and computer scientists. Thus, we want to make clear if ethics should be seen as some kind of toolbox whereof the fitting tool for a given problem can be chosen or if these crafting metaphors lead in the wrong direction. Should ethics specify best/worst case-scenarios or declare values? Does the moral claim of option-validity refer to the AI or to the people who work with the AI (in programming, evaluating and using the products)? We are going to discuss how these tasks relate to the current possibilities, technical implementations and limits of AI application. However, could it be that everything moral initially escapes the processes of machine-learning?
We will, first, paraphrase some reference to the articles of LaCroix & Luccioni and, second, we will show a variation of the dichotomies of realistic and non-realistic metaethics. Therefore, we are going to examine an argument of LaCroix & Luccioni in which the authors both criticize and transform the way ethics and AI should be combined. In order to close this gap and to match the moral capabilities, we are going to describe briefly how benchmarks have been further developed for mapping ethical factors in computational systems. Now, because the relevance of ethics is also clear as these systems are increasingly used and integrated into everyday life and because they seemingly act

"autonomously" we want to use the weight of the moral decisions concerned to differentiate between quantitative and qualitative conditions that could be taken into account for a combination of mere algorithms and moral decision-making. We conclude, that if such entanglements were the case, it would also be correct that for a certain range of outcomes, a compliance of moral standards in AI acting should be provided. But "where" would that have to happen? Is it possible to declare benchmarks for the application of ethical standards to models of AI, say, in machine learning or so-called autonomous systems? Or would that exaggerate the intentions of technical systems in use? Hence, we will argue that it is possible to simulate moral decisions if we use the architecture of human moral processing; but that doesn't mean that, at the end, we will have a "real" moral decision of AI itself.

Claudia Schon is a professor for Artificial Intelligence at Hochschule Trier (Trier Univerity of Applied Sciences), Germany. She received her PhD in Computer Science in 2016 at the University of Koblenz. Her main research interest is the application of automated reasoning methods in the area of common-sense reasoning, and she is particularly interested in combining symbolic and statistical methods in this area. She is the current co-speaker of the special interest group on deduction systems of the German Informatics Society (GI). She has co-organized 5 workshops on Bridging the Gap between Human and Automated reasoning at CADE (2015), the annual meeting of the cognitive science society (2017), at IJCAI (2016, 2019) and ECAI/IJCAI (2018) as well as the workshop on Practical Aspects of Automated Reasoning (PAAR 2022) at FLoC/IJCAR 2022.

Werner Moskopp is a researcher and senior lecturer for Philosophy at the University of Koblenz. Currently he is mainly concerned with topics of moral philosophy and of pragmatist methodology. His second book (Habilitationsschrift) with the title "Verbindlichkeit" was published in 2021 (Alber Verlag); in 2023 he edited an anthology concerning "Figurationen des Bösen". Also, he is the external editor of the Online Dossier "Bioethics" of the Federal Center for Political Education (Bundeszentrale für Politische Bildung).

José Antonio Pérez-Escobar, Deniz Sarikaya

Wittgensteinian precepts for AI safety

Consider the following. A "superhuman" AI is given a task—to reduce the number of people with cancer. One would expect that the AI would find new drugs, new treatments, and better means of diagnosis. We then notice that people start dying massively. It turns out that the AI poisoned running water in cities. This entails fewer people have cancer. Even worse—we try to stop it, but it tries to stop us from turning it off because it would be less capable of fulfilling its goal. What is more, even if we anticipated all this and tested the AI in a simulation first, it would realize that this was only in a simulation and would deceive us by "behaving properly".
We will explain recent concepts of AI safety, like mesa and base optimizers and the alignment problem (in its inner and outer version). We claim that at the core of these problems there are rule-following issues that can be understood from a Wittgensteinian perspective. Late Wittgensteinian philosophy about language and mathematics, heavily focused on rule-following, may yield insights on how to develop AI safely. The later Wittgenstein's philosophy of language and mathematics and further secondary literature characterize how there is no intrinsic, unequivocal meaning in the formal or material aspects of symbol arrays, how we humans use language irregularly, and what factors keep linguistic practices (including mathematics and logic) under control in social communities.
In principle, two strategies to address the alignment problem may be considered. The first is that we, humans, natural rule-benders, may stop bending rules by focusing on "clear" specifications of human goals. This may be achieved, for instance, by training in programming and computer science, in the formal aspects of "unbent" machine language. However, this training and linguistic clarifications are

patchworks; they do not fully address the underlying principle at the base of the problem: even master programmers make mistakes and cause bugs. In this context, mistakes like these can be fatal. Still, late Wittgensteinian philosophy can inform us on what the most effective patchworks would be.
The second is designing rule-bending AIs. "Proper rule-following" is underdefined by rules, and clarification may help but this issue is in principle unsolvable. According to Wittgenstein's notion of "meaning as use", the best way to understand a rule is to see how it is used in natural contexts. AIs can be supervised and trained in human-like training contexts, in the spirit of the notion of rule-bending. We argue that this is a more feasible option. In this talk, we will unpack the second strategy, identifying the factors that, for Wittgenstein, lead to mutual understanding among humans. These factors are of a psychological, social, and cultural character. We present a model integrating these factors according to their generality vs specificity. They must be carefully included in AI training to "humanize" AI, similarly to how the later Wittgenstein conceived logic and mathematics. Finally, we argue that this strategy is more reliable than the first regarding the development of safe AI.

José Antonio Pérez-Escobar is about to finish a two-year postdoc at ENS Paris on the epistemic effects of the mathematization of biology with a Swiss National Science Foundation fellowship. In 2024 he will be a postdoc at the University of Geneva with his project "Mathematical models and normativity in biology and psychology: descriptions, or rules of description?" from the Swiss National Science Foundation. He holds a PhD in Philosophy from ETH Zurich. In addition, he has a formal background in psychology (Bachelor and Master) and neuroscience (Master and PhD) and has been involved in topics which feature mathematical techniques, like psychometrics and computational neuroscience. José Antonio Pérez-Escobar's PhD in philosophy concerns the epistemology of the mathematization of biology. He has argued that mathematics is not epistemically neutral, that different mathematics can be used in epistemically productive or harmful ways, and that mathematics can promote or discourage a teleological worldview in biology.

Carsten Ochs

Negotiating Hypernormativity: A Situational Analysis of the AI Arena

Although algorithmic systems are still associated with „claims to objectivity, impartiality, and legitimacy" (Kitchin 2014: 9), they have a normative impact that oftentimes was not intended by those who developed these systems in the first place. Whereas this goes for algorithmic systems in general, the deployment of Machine Learning (ML) techniques obviously aggravates the problematic, as became clear, e.g., in the infamous case of Google's racist facial recognition classification (Kühl 2015). The fact that the only solution the company found was to shut off the classification system is clearly an indication for the relative unforeseeability of these systems when it comes to normative impact. As the system's output hinges on statistical probabilities it can be difficult of even impossible to gain a sematic understanding of their operations (Burrell 2016; Döbel et al. 2018). As a result, the social negotiability of norm-generation is further diminished.
I call the tendency of ML-based AI systems to produce norms in a way that does not consider general moral standards of society „hypernormativity", ultimately leading to hyper-nomy: the hypertrophic extension of rule triggered by AI systems' tendency to grow exuberantly norms that only operate in terms of impact, but not in terms of intent.
The analysis raises theoretical as well as empirical questions. The theory question is how to conceptualize ML-based AI systems' tendency to grow technonormative scripts and the social consequences this has; the empirical question concerns the societal negotiation of measures that help to contain hypernormativity.
As my presentation attempts to shed light on both issues, I will start with characterizing ML-based AI systems' impact as follows:
1. They regulate social processes: they execute rules whereas humans follow, i.e. interpret rules.

2. They generally have a normative impact: there is collective evaluation of their operations and the results they bring about.
3. Their scope of norm-variation is unknown: they have no internal moral limits that would narrow their potential operations or the consequences they bring about.
4. Their operations are generally opaque: they do not give observers a straightforward idea of how knowledge is encoded into decision-making processes, nor about the weight given to particular decision-making parameters.
Having said this, the social problematic of ML-based AI systems is currently not so much their becoming auto-nomous nor that they create heter-nomy or a-nomy, but the hyper-nomy triggered by these systems. This raises the question of how societies might set limits to ML-based AI systems' hyper-normativity in empirical practice. The second part of the presentation therefore presents insights gained from a research project that investigates the societal negotiation of AI; and draws ethical conclusions from a Situational Analysis conducted in this context.

Carsten Ochs is a postdoctoral researcher at University of Kassel, working for the research project Artificial Intelligence, Privacy & Democracy (BMBF). In 2022, he published his habilitation thesis Soziologie der Privatheit. Informationelle Teilhabebeschränkung vom Reputation Management bis zum Recht auf Unberechenbarkeit (https://doi.org/10.5771/9783748914877). Before coming to Kassel, he held Postdoc positions at TU Darmstadt's European Center for Security & Privacy by Design and Sociology Department. His PhD was located at Justus-Liebig-University Giessen's Graduate Centre for the Study of Culture (GCSC) where he did research on global digitization processes after having completed the Master Programme "Interactive Media: Critical Theory & Practice" at Goldsmiths College/Centre for Cultural Studies, London and undergraduate studies in Cultural Anthropology, Sociology and Philosophy at Goethe-University Frankfurt. Wherever being located, he dealt with digitization issues throughout his career.

3. DIGITAL NUDGING

Philippe Verreault-Julien

Ethical Nudging with Opaque Recommender Systems?

This paper examines ethical challenges that arise in the context of using deep learning models for nudging. In particular, I will show how opacity may make it difficult to assess whether the nudges 1) make people better off as judged by themselves and 2) do so without unduly interfering with their autonomy.
Recommender systems, viz. artificial intelligence systems whose aim is to select a subset of options from a larger set and present them to the user, are ubiquitous in digital environments. As such, they are an instance of digital nudging (Jesse and Jannach 2021; Weinmann, Schneider, and Brocke 2016): recommender systems steer people's behaviour in predictable directions. Deep learning models are increasingly used for designing these systems (Afsar, Crump, and Far 2022; Zhang et al. 2019). One potential benefit of using these models is that they allow for improved recommendation personalization. Using deep learning models for personalization, however, raises an ethical issue; these models are typically opaque. Indeed, their opacity implies that we often do not understand why they produce the outputs they do (see, e.g., Creel 2020; Watson 2022; Zednik 2021). Although personalization is prima facie positive for ethical nudging, opacity is not. Since opacity undermines our understanding of why the options were selected and what and why are the means of nudging (e.g. mechanisms), it seems personalization with deep learning models may make nudges ethically dodgy.
Illustrating using the case of Netflix's recommender system (Steck et al. 2021), I examine how opacity may undermine two conditions for ethical nudging, viz. that they 1) make people better off as judged

by themselves and 2) do so without unduly interfering with their autonomy. First, I argue that the opacity of deep learning models makes it difficult to assess whether they promote people's interest. One particular problem is due to a potential misalignment between the proxy metric and the real metric we care about. Then, I argue that opacity is also a problem for assessing whether the means of nudging interfere with people's autonomy. Knowing what are the means of nudging and why they were selected is crucial for making sure that the choice architecture respects people's autonomy (Grüne-Yanoff 2016).
I conclude by considering whether explainable artificial intelligence (XAI) techniques (see e.g. Adadi and Berrada 2018; Burkart and Huber 2021) may help solve the ethical conundrum by making the deep learning models more transparent. Whether current XAI methods can help understand the behaviour of deep learning models is controversial (e.g. Babic et al. 2021; Rudin 2019). However, if they can, this suggests that the challenges of nudging with opaque artificial intelligence systems are not categorically different from those facing more traditional behavioural interventions. In all cases, we need reliable evidence that the options would indeed promote people's interest and that the means do not unduly interfere with people's autonomy.

Philippe Verreault-Julien is a postdoctoral researcher at Eindhoven University of Technology, where he is part of a project on opacity in artificial intelligence. His work focuses on the epistemology and ethics of scientific modelling. He obtained his PhD from Erasmus University Rotterdam and was subsequently awarded a postdoctoral fellowship to conduct research at the Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science. You can learn more at https://pvjulien.net.

Marius Bartmann

Digital Nudging and Personal Autonomy

Navigating the online world is highly mediated by automated systems that select information and arrange options purportedly to support our decision-making and to improve our choices. On streaming portals, shopping sites, social media platforms, and online newspapers we are confronted with suggestions regarding what video to watch, what products to buy, what posts to like, and what articles to read. The automated systems behind these recommendations promise to prune back the digital jungle and guide us to the things we are actually looking for.
One way to conceptualize these types of decision support systems consists in characterizing them as a form of digital nudging: the employment of user-interface design elements to guide people's behaviour in digital choice environments. The concept of nudging was originally developed by Thaler and Sunstein as the central tool in their policymaking approach they call libertarian paternalism. This approach is designed to help people make decisions that are better for themselves but does not restrict their freedom of choice. Unlike conventional paternalism, libertarian paternalism tries to achieve this goal without bans and incentives, but simply by intervening in people's choice architecture, i.e. by arranging and presenting the options in ways that make it more likely for people to choose what is supposedly better for them anyways.
Digital Nudging is highly controversial. Selecting the possible options among which we can choose and presenting them in specific ways can profoundly affect our decision-making and hence our resulting choices. On the one hand, automated filtering may improve our choices by providing us with relevant options tailored to our preferences. On the other hand, these automated and often opaque mechanisms can be deceptive or even manipulative when they do not reflect our values, ends and preferences or when they exploit weaknesses of our decision-making and steer us towards options we do not actually want. What is thus at stake is our personal autonomy, i.e. our capacity to set our own ends and to achieve them by considering possible choices and weighing reasons to make decisions.

In my talk I want to examine how automated recommendations affect our decision-making and whether they pose a threat to autonomy, i.e. whether they compromise our capacity for making our own choices. I will argue that a basic requirement for integrating recommendations in autonomous decision-making consists in being able to identify the rationale behind recommendations. Only if we have a sufficient grasp of how we are provided with automated recommendations is it possible for them to be integrated into decision-making in a way that preserves autonomy.

Marius Bartmann is a Post-Doctoral Research Fellow at the German Reference Centre for Ethics in the Life Sciences (DRZE, University of Bonn). He completed his PhD in Theoretical Philosophy at the University of Bonn. His current research focuses on ethical questions revolving around autonomy in the digital age as well as the ethics of climate change. Marius Bartmann is also heading the ethics project of the interdisciplinary project "Practical Challenges of Climate Change" (funded by the Federal Ministry of Education and Research). Recent papers include "Reasoning with Recommender Systems? Practical Reasoning, Digital Nudging, and Autonomy" (forthcoming in Recommender Systems: Legal and Ethical Issues, ed. by S. Genovesi, K. Kaesling and S. Robbins) and "The Ethics of AI-Powered Climate Nudging – How much AI should we use to save the planet?" (Sustainability 2022, 14 (9), 5153).

Aaron Schultz

Distraction and Attentional Freedom

To better understand our relationship with machines that use AI and machine learning, it is important to consider the moral roles that attention and distraction have. When users interact with technology, they must give their attention over to it. Much of the technology we interact with on a day-to-day basis is designed specifically to maintain or increase user engagement. As a result of these design choices, we have developed systems that are highly effective at capturing our attention.
Some have pointed out that these technologies can be manipulative and deceptive. When AI technology is used to manipulate us by relying on deception, we have clear reasons for why this should be thought of as wrongful. It is wrong to deceive people because it overrides their ability to consent. For genuine consent to occur, one must be reasonably informed of what is happening and what they are consenting to. The obvious solution here is to find ways to remove the deception and thereby end the manipulation that occurs through deception.
However, even if we find ways to eliminate deceptive practices, I argue that our autonomy can still be threatened. This is because the technologies we use override our freedom not only through deception, but through distraction.
At the surface, distraction seems to be an innocuous charge. If a company makes a product that distracts you, even though we recognize the pull that the technology has, we tend to ultimately blame the user. The world is full of potentially distracting things, and it is up to each of us to decide what to attend to. Moreover, what might be a distraction for one is for another a valuable object of attention. What these trains of thought miss is the nature of attentional freedom. Our attention in any given moment is limited. We cannot attend to everything, and each moment of attention must be on one thing or another. When something attracts our attention, we may decide to place our attention on it or ignore it. Of course, many of us suffer from weakness of will and we often attend to things that we may, in hindsight, regret attending to.
In my paper, I argue that some forms of distraction are morally impermissible. It is wrong to distract someone when one overrides another's ability to withdraw from the interaction. This analysis will mirror some of the ways we typically consider rights violations to bodily autonomy. Similar to these kinds of rights violations, there will be borderline cases. However, there will also be clear examples where the distraction is wrongful, and the wrongfulness can be explained by the fact that the distraction overrides one's attentional freedom. I argue that those that design, implement, and improve technology that utilizes AI and machine learning to capture our attention have a moral

obligation to preserve users' attentional freedom. I conclude this paper by offering some recommendations about the kinds of changes that could be made to existing technologies to help fulfill this moral obligation.

Aaron Schultz is currently an Assistant Professor at Michigan State University. His past research has focused on Buddhist responses to wrongdoing and problems related to the justification of state punishment. Currently, his research is focused on the moral and political problems presented by artificial intelligence.

Charlie Kurth

AI-Driven Emotion Nudges: Inviting a Wicked Problem?

Cultivating one's emotions—learning to feel anger, say, at the right time and in the right way—has long been viewed as a central to moral education. Recently, a diverse group of educators, philosophers, and entrepreneurs has pointed to "emotion nudges" as a powerful, but under-utilized tool in our emotion cultivation efforts (Valor 2016, Engelen et al. 2018, Suh 2016). The core idea is straightforward. We know that nudges can bring better choices (Thaler & Sunstein 2008); this suggests they can also bring better feelings. In fact, adding artificial intelligence to the mix would only seem to make emotion nudges more powerful (Green 2019). Moreover, the initial results are intriguing. For instance, merely placing "watching-eye" icons in online chatrooms can prompt feelings of anxiety that help curb the proliferation of vicious posts (Park 2022), and virtual reality simulations can engage stereotype-challenging empathy (Bedrik 2017).

But while AI-driven nudges may be a good way to promote things like retirement savings, their appropriateness for moral education is much less obvious. In fact, I argue that in pursuing these emotion nudges, we are courting trouble: Given what emotion research tells us about emotions and our ability to shape them, the use of AI-driven emotion nudges to promote moral education brings a distinctive and vexing set of scientific and ethical challenges. To draw this out, I focus on three intertwined issues.

(1) We can start by asking which nudges work. But this seemingly straightforward question is actually extremely complicated. For instance, recent findings indicate that cultivation techniques that are effective in shaping one emotion are likely to set us back with regard to others (Hafenbrack 2022). Moreover, our ability to cultivate an emotion appears to vary depending on which emotion we're focusing on (Nussbaum 2004, Kurth 2019). So how can we design effective nudges given how little we know?

(2) This first set of issues invites a turn to AI, where big data and machine learning can help us optimize. But can it do so in a way that gives voice to the emotions of underrepresented groups? If not, then, AI-driven nudges threaten to further marginalize the already marginalized. Importantly, this second issue concerns more than just determining who gets to make these decisions, for there's also the deeper problem of making these decisions in the absence of any consensus on what the relevant feeling norms should be (witness recent debates about who can be angry and when (e.g., Pettigrove 2012, Srinivasan 2018, Cherry 2021, Flanagan 2021)). Relatedly, some the most "exciting" AI nudges—virtual reality simulations that engage empathy so that users can experience what it's like to be a cow going to slaughter or an undocumented worker being smuggled—are morally dubious insofar as they invite tokenism, deception, and voyeurism.

(3) Here one might hope that dialog and democracy can help us work through our competing values and curb our enthusiasm for ethically suspect applications. But recent controversies about social-emotional learning suggests such hopes are naïve—for, alas, talk of emotion and emotional education has become the latest hot-button issue of our culture wars.

Charlie Kurth is Professor of Philosophy at Western Michigan University and a Core Fellow at the Helsinki Collegium for Advanced Studies at the University of Helsinki. His research focuses on questions about the place of emotion in virtue and the good life. His recent work explores whether cultivating negative emotions like anxiety, disgust, and shame is central to our ability to become better, happier people. Kurth is the author of two books, The Anxious Mind (MIT 2018) and Emotion (Routledge 2022) and more than 20 scholarly articles. He has also written essays on emotions in mainstream publications like The Washington Post, Scientific American, and Aeon.

4. BEATE ROESSLER

Why Robots Can't Get Ill: On the Concept of the Digital Human Being

In recent literature on digital technologies, we increasingly come across the assumption that the new technologies, though making our economies, societies, and our lives essentially easier and more efficient, also risk to fundamentally change human nature. Already some years ago, Acquisti e.a. for instance, wrote that "technologies, interfaces, and market forces can all influence human behavior. But probably, and hopefully, they cannot alter human nature."
What interests me in this paper is how to spell out this idea more precisely: What does it mean that we hope that technologies will not change our human nature, what would that human nature be and why would it be bad to change it? I will argue for the thesis that the concept of human being can be defended against a number of critics: posthumanist theories and transhumanist theories. And I will place the human being in relation to social robots and discuss the extent to which social robots can replace humans in certain respects. Starting with an analysis of what it means to be ill, I'll try to demonstrate that being ill can be taken to be paradigmatically human – if we take up the first-person perspective, we see that we are embodied, vulnerable and finite beings, always living in social practices. Properties which exemplify the human being, or so I will argue.
In the end, I will claim that there's a categorical, not a gradual difference between AI or robots and human beings. But this is all rather tentatively, and I'll be predominantly interested in exploring these issues.

Beate Roessler is Professor of Philosophy at the University of Amsterdam. She formerly taught philosophy at Leiden University, the Free University, Berlin, Germany, and at the University of Bremen, Germany. She had fellowships and visiting professorships at the Institute for Advanced Study (Wissenschaftskolleg) in Berlin, at the Center for Agency, Value, and Ethics at Macquarie University, Sydney, at the University of Melbourne, Law School and at the New York University. She is a co-editor of the European Journal of Philosophy and a member of various advisory boards.
She has published widely on topics in ethics, social, and political philosophy; her most recent book is Autonomy. An essay on the life well lived, 2021 with Polity. Her current research focuses on the Being Human in the Digital World.

5. AI & VALUE ALIGNMENT

Leonard Dung

Current cases of AI misalignment and their implications for future risks

How can we build artificial systems which pursue objectives that correspond to human values? This is the AI Alignment Problem. Systems which are misaligned will optimize for goals which leave out or conflict with important values or ethical constraints such that harm might ensue. In addition, numerous authors have raised concerns that, as research advances and systems become more

powerful over time, misalignment might lead to catastrophic outcomes, perhaps even the extinction or permanent disempowerment of humanity (e.g., Bostrom, 2014; Ngo et al., 2022; Russell, 2019).
In this talk, I will illuminate the AI Alignment Problem by scrutinizing current instances of misaligned AI. The core question is: What do current cases of AI misalignment (and alignment) tell us about the prospects and risks of aligning more advanced systems with human values? I will focus on two examples: First, large language models (LLMs) like ChatGPT which recently gathered a lot of attention. Second, videogame AI like the reinforcement learning agent trained by OpenAI to play the boat racing game CoastRunners (OpenAI, 2016). In the case of LLMs, the alignment problem consists in teaching the AI to say things which are correct, helpful and harmless (e.g., no insults, expressions of bias or incitement to violence). In the case of videogames, the goal is to train the AI such that it plays the game in the way itis intended to be played.
Importantly, both cases are alignment problems because what prevents the systems from exhibiting the desired behavior is not (merely) a lack of capabilities. It is often not particularly difficult to provide true and harmless answers to questions or to aim to win a virtual boat race. The problem is that the system is not even trying, because it engages in a different task, i.e., text prediction or maximizing an imperfectly correlated reward signal. The distinction between lack of capabilities and of alignment can be brought out by noticing that, in these cases, simpler and less capable systems are frequently better in performing the intended task.
I will argue that these cases show that some degree of misalignment is the default outcome of training deep learning models, and that misalignment can be hard to predict, to detect and to remedy. On the positive side, there is a systematic connection between the extent of a system's usefulness and its degree of alignment.
What lessons for the prospects of aligning more powerful systems can we learn? First, while we can hope that more capable systems are aligned by default, this is not guaranteed and does not seem exceedingly likely. Second, the previous cases suggest that alignment will probably be a hard technical challenge. This is worrying as misalignment poses a higher risk in more capable systems. Third, I will sketch that, in more capable systems, new problems may arise which will compound the alignment challenges we currently encounter.

Leonard Dung is a Philosopher at the Centre for Philosophy and AI Research, located at the University Erlangen-Nürnberg. Previously, he earned a PhD from the Ruhr-University Bochum. Presently, he is working on the philosophy and ethics of artificial intelligence. His research especially focuses on AI sentience, AI moral status and risks, including existential risks, from advanced AI systems. Moreover, he investigates topics related to animal consciousness and welfare which were also the focus of his PhD.

Michael Cannon

Two Kinds of Control Problems

The aim of presentation is to propose a distinction between 1st and 2nd Order Control-Problems in order to identify requirements for meaningful control of AI technologies.
A challenge for developing meaningful control of AI is that we do not yet have an effective and comprehensive understanding of how AI technologies impact human choice-making and value-formation. We do not yet have a comprehensive understanding such that it is clear where and how to act, or where meaningful control is possible or desirable. One effect is that it is difficult then to see where existing work for meaningful control already exists, and where else it is still possible and required. In this regard, work on meaningful control remains uncoordinated. This presentation therefore aims to offer a distinction to organise the kinds of challenges to meaningful control of AI technologies.

The point of departure of the presentation is a brief description of the Cybernetic movement, the source of the distinction between 1st and 2nd order control problems. The cybernetic movement distinguishes two kinds of systems, "cybernetic" and "2nd order cybernetic". Cybernetic systems are systems in which feedback loops enable a regulation of behaviour in order to achieve a goal (Rosenbluth et al. 1943, Wiener 1948). Brains and AI systems are examples of a cybernetic system. A 2nd Order Cybernetic system is one that includes both a cybernetic system and the observer thereof (Mead 1968, von Foerster 2003). The founder of the theory, Heinz von Foerster, explained the difference by pointing out that "it requires a brain to produce a theory of a brain", and that ensemble of observing system and observed system is a 2nd order cybernetic system.
With this historical context, the presentation will describe 1st and 2nd order control-problems. A 1st Order control-problem concerns meaningful control of a cybernetic AI system. Examples of 1st order control problems can be split into two kinds. One kind is the "Value Alignment" kind of control problem (Bostrom 2014, Russell 2019, Christian 2021) that treats the cybernetic AI system as an agent and works to ensure its "values" are aligned with human values. Another kind of 1st Order Control problem concerns the effects of cybernetic AI/ML systems like Lethal Autonomous Weapons Systems, or recommender algorithms on entertainment and social media platforms, or classification systems that systematically harm marginalised populations (Birhane 2022). A 2nd Order Control-Problem concerns meaningful control of a 2nd order cybernetic system – the cybernetic AI system plus the people who built it, people who are themselves embedded in socioeconomic, technoscientific, and geopolitical systems. Here, work by (Crawford 2021) on the human, ecological, and geopolitical costs of chains of AI production, work by (Birhane and Guest 2021) on cultural change to decolonize the computational sciences, and work by (Cannon 2022) on metatheoretical questions about the affordances of our different models of minds for understanding humans and AI minds, are all examples of work pointing to challenges of 2nd order meaningful control.
Distinguishing kinds of control problems in this way can help identify the kind of work that "meaningful control" of AI technologies requires and thus enable greater coordination of collective efforts.

Michael Cannon is a PhD researcher at Eindhoven University of Technology where he is submitting a thesis on the question "Can AI become more ethical than humans?". His thesis research explores the cognitivist and post-cognitivist paradigms of mind research, comparing their perspectives on AI and its possibilities. His broader interests are situated in the context of the Anthropocene and concern making sense of AI in this time, and as a historical, enculturated story about humans and human activity.

6. JOERN LAMLA

Artificial Intelligence as a Hybrid Life Form. On the Critique of Cybernetic Expansion

Artificial intelligence (AI) challenges human intelligence and our humanistic self-conception. My contribution argues that this is happening for good reasons but is based on a mistaken opposition that falls short. Human beings and technology have always been intertwined in hybrid forms of life. Yet the exact nature of this hybridity is misunderstood when inadequate dichotomies of human subject and technical object are replaced by a totalizing conception of a cybernetic informational universe that reduces all that exists to this latter, single point of comparison. Representing the paradigm of digital society, AI is a bearer and expression of such a cybernetic expansion that both anchors digital analogism in society as a closed system of interpreting the world, or a cosmology, and renders it plausible at the level of knowledge. AI thus deepens and generalizes conventions and functional patterns of justification that have a long history in industrial society. The thesis proposed here is that, to counter this expansive dynamic effectively and critically, more needs to be done than evoke humanistic values. What we need is a better understanding of the ontological heterogeneity of the societal modes of existence that are assembled in hybrid forms of life.

Prof. Dr. Jörn Lamla, born in 1969, has headed the Chair of Sociological Theory since 2013 and has been Director at the Scientific Center for Information System Design (ITeG) at the University of Kassel since 2015. He received his PhD from Friedrich Schiller University in Jena in 2000 and habilitated there in 2012 with a thesis on "Consumer Democracy" (Suhrkamp 2013). In the summer of 2015, he held a visiting professorship at the Centre for Ethics at the University of Toronto. He was a member of the founding board of directors of the Hessian Centre for Responsible Digitalization (ZEVEDI) from 2019 to 2021 and has been a member of the coordinating body of the Federal Network for Consumer Research since 2015 and spokesperson since 2019. His work focuses on social theory, political sociology, and studies of digitality and consumption. Currently he is conducting research in interdisciplinary project networks on the transformation of privacy, democracy, and self-determination considering the increasing influence of AI and algorithmic valuation.

7. DEMOCRACY AND PARTICIPATION

Blair Peruniak

Artificial Intelligence and Workplace Democracy

Despite significant interest in the democratic potential of AI technologies and workplace democracy their relationship remains elusive. Drawing on recent work in the theory of artifactual design, I argue that the compatibility of AI technologies and workplace democracy can be understood as a function of a designer's substantive intentions to endow technologies with particular democratic or undemocratic properties and to realize these properties in the workplace. Part I addresses legitimate worries about the inherent incompatibility of AI and workplace democracy by distinguishing between conditions that are favorable or unfavorable to their fruitful alliance. Scepticism of the role of AI in fostering democratic forms of labour-capital relations often focuses on the threat of technological unemployment or the capacity of AI and machine learning technologies to enhance workplace oppression and managerial control, while failing to distinguish between workplaces that are inherently opposed to democracy from those that are more supportive of (or susceptible to) democratic reforms. As a result, the role of AI remains largely unaccounted for in a wide range of labour contexts. In Part II, I explain how the intentional properties of AI technologies can be used in evaluations of whether (or which) technologies are democratically empowering or exploitive of workers. I briefly discuss properties relevant to evaluations of the democratic potential of AI before moving to defend this account against the criticism that the creations of even well-intentioned AI designers will tend to support only superficial democratic labour reforms and, thus, that AI developments remain essentially incompatible with substantive forms of workplace democracy.

Blair Peruniak received his DPhil. from the Department of International Development at the University of Oxford (2019) as a Social Sciences and Humanities Research Council Doctoral Fellow. Blair is currently an Alex Trebek Postdoctoral Fellow in AI and Environmental Justice at the University of Ottawa's Centre for Law, Technology and Society where he focuses on identifying solutions to essential issues related to ethical AI and technology development with the support of UOttawa's AI + Society Initiative. Blair is also a lecturer at McGill University's Institute for the Study of International Development (ISID) where he teaches on related topics of international law and migration governance, forced displacement, and climate change.

Bjorn Fasterling, Gianclaudio Malgieri, Geert Demuijnck

Participating in – before contesting – data-driven activities? Addressing under-representation and power imbalances of vulnerable people

In this conceptual paper we focus on the question, to which extent, and in which manner, existing and upcoming legislation in the EU governing the use of artificial intelligence and advanced data analytics, enables, or contrarily, impedes the ex-ante participation of those people, whose lives may be affected by an algorithmic outcome or a data-driven project. We pay particular attention to the situation of vulnerable people or typically underrepresented social groups.

Privacy and data protection laws in the EU have generally relied on the "inform-consent-contest" (ICC) approach to guarantee autonomy, dignity, and privacy of individuals in data-driven environments. According to this ICC model, the best way to protect rights and participation of individuals in complex data processing activities would be to (a) inform them that their data are being automatically processed and (b) to ask their consent to such processing or, (c) eventually, to allow them to contest the automated decision ex-post and ask a new human-mediated decision after they have expressed their own view (cf. Cohen 2019; Barocas and Nissenbaum 2009). Here, the communication between different parties is generally based on an adversarial design model, generating "a condition of forever looping contestation" (cf. DiSalvo 2015).

The ICC model has, however, proven ineffective, and the autonomous authorisation model of a data subject capable of making a free and rational decision to consent may have been overemphasized. First, information notices might be fallacious, and people might be either uninterested or incapable to understand real risks and implications of data processing activities (Solove 2013; Schermer et al. 2014; Fortuna-Zanfir 2014). In addition, in many contexts there is no real freedom to give or deny consent (Austin 2014; Bergemann 2018). Moreover, many individuals are even unaware of being part of an especially vulnerable category (Wachter 2019) and being victim of discriminations (Barocas and Selbst 2016). Even well-informed individuals might be incapable to react to unfair data processing involving them. We argue that these disadvantages and dysfunctions become disproportionately harmful for social groups that are already vulnerable (Malgieri and Niklas 2020).

The above-mentioned problems of the ICC model are in some cases lessened if ICC is complemented by "due diligence" processes that aim at preventing harm. For example, the proposed European legislation on AI, the draft AI Act (COM/2021/206 final), bases the legality of the use of a high-risk AI system on the implementation of various risk management measures. However, neither existing data protection impact assessments under the GDPR nor the forthcoming AI risk management systems provide for systematic ex-ante consultation of potentially affected people.

In an outlook we raise attention to forms of ex-ante participation in data projects and design of algorithmic models. Key issues for further research include, then, how to consult the impacted people and how to guarantee a good level of representation of marginalized groups ex ante. With this study, we hope to demonstrate the benefits of ex-ante participatory models of decision-making by proposing processes that might overcome current institutional pitfalls.

Björn Fasterling is professor of business law and business ethics and EDHEC Business School (France). His research and teaching focus on compliance, digital ethics and business & human rights. At EDHEC he served as Head of Faculty for Law & Accounting and now directs the "Digital Ethics Officer" program of EDHEC's Augmented Law Institute. Prior to joining EDHEC he practiced law as a German lawyer in the Berlin office of Wilmer&Hale. He holds a phD degree (Dr. iur) from the University of Osnabrück and an LLM degree from the University of Stockholm.

8. POLICY & REGULATION

Juan-Pablo Bermúdez

Ethics in the Gray Zone: Guidelines for Autonomy-Supportive Automated Influence

Worries about the use of AI systems for automating manipulative influence at scale, leading to a loss of meaningful human control, are widespread. Legal systems are beginning to respond to this concern. For instance, the European Union's draft AI Act includes a prohibition of AI systems that use "subliminal techniques" to alter people's behavior in ways reasonably likely to cause harm (Article 5(1)(a)). But how can we distinguish legitimate from harmful instances of automated influence?
We propose an ethical framework for distinguishing autonomy-supportive and autonomy-undermining boundaries in human-AI interactions. Given the complexity and multi-dimensionality of terms like 'autonomy', 'manipulation', and 'control', a key challenge is dealing with grey-zone cases: cases of influence whose negative effects on autonomy are so minor that they do not seem to be impermissible. This is a central problem for the ethics of AI because many, if not most, cases of automated influence fall in the gray zone. Following recent work on online manipulation, we adopt a pragmatic approach to tackle this issue. Instead of building a conceptual account that draws a clear boundary between permissible and impermissible influence (which would end up mired in theoretical controversies), we propose a norm designed to deal with the ambiguity gray-zone cases.
Our starting point is a widely accepted notion of autonomy based on both deontological and consequentialist elements. On this basis we build a two-dimensional framework for ethical influence, according to which the ethical status of a particular influence attempt can be identified by answering two questions: To what extent is it transparent? And to what extent is it aligned with the values of the target agents? We use these dimensions to distinguish three zones of automated influence: the green zone (where influence is both transparent and value-aligned, and thus ethically permissible); the red zone (where influence respects neither dimension of autonomy and is thus ethically impermissible); and the gray zone, populated by influence forms that satisfy only one of the two dimensions of autonomy.
For gray-zone cases, we propose the following compensatory norm: to the extent that there is a deficit in one autonomy-supportive criterion, extra care should be taken that the other one is exhaustively met. Thus, if it cannot be ensured that the influence attempt is fully value-aligned, a high level of transparency should be ensured; and conversely, if full transparency cannot be ensured, the influence attempt should be designed to ensure a high level of alignment with the values of the influenced human agents. We propose the compensatory norm as a tool to preserve meaningful human control in human-AI interactions.
We finish by testing the two-dimensional framework and the compensatory norm against specific cases (subliminal techniques, digital nudges, and AI-powered chatbots) and relating the framework back to the draft AI Act's prohibition of manipulative subliminal AI systems. We argue that in all these cases the framework provides verdicts aligned with ethical intuition while also providing guidance for the responsible design of AI-based influence technologies.

Juan-Pablo Bermúdez is a philosopher of psychology and technology, and currently works as Research Associate in the Philosophy of Human-Computer Interaction at the Dyson School of Design Engineering, Imperial College London. Prior to joining Imperial College, he obtained his PhD in philosophy at the University of Toronto, and worked as postdoctoral researcher at the University of Neuchâtel and as assistant professor at Externado de Colombia University, where he is still affiliated.

Monique Munarini

Aligning soft compliance mechanisms with due diligence principles to promote human values and eliminate societal harms in the AI ecosystem

EU regulators have been working to strengthen legal frameworks involving emerging technologies, mostly based on machine or deep learning techniques such as artificial intelligence, highly used indecision-making processes. It is known that artificial intelligence inherits societal bias and has the potential to escalate harm that possibly leads to fundamental rights violations. While the GDPR is

already in force to protect citizens against unbalanced outcomes derived from a decision taken with the support of technology, there are ongoing legislative efforts focusing on artificial intelligence, commonly known as the alphabet soup of AI legislation. In fact, exploring ways to maintain meaningful human control over the possible influence of algorithmic outcome on human decision-making processes is a critical point in the governance of AI, which has quite few case law studies and which will soon be one of the major problems in the application of law. In this meantime, to bridge this regulatory gap, many soft compliance mechanisms were created focused on keeping mostly developers accountable to ethical principles such as accountability. Accountability is composed of a complex of instruments, such as auditing methodologies, codes of conduct and guidelines, which is paired with due diligence principles from the business and human rights field. At the same time, that it is desired to mitigate the inheritance of societal bias, it is proposed to include ethics by design of these technologies to foster responsible artificial intelligence. The aim of this paper is to understand how soft compliance mechanisms allied with due diligence principles can impact in shaping the filters between society and artificial intelligence to promote human values without reproducing societal bias. This critical desk research will be based on secondary sources with an explorative multiple case study approach to verify in which context in the so-called automated decision-making processes (ADM) soft ethics and due diligence principles were or could have been used to do this filter between society and technology or when the lack of ethics by design was critical to result in violation of fundamental rights. To conclude, and considering the fast-paced development of artificial intelligence, this research will shed some light to the ways in which due diligence principles can be used to verify how actors in the development and deployment of artificial intelligence can use selected ethics guidelines created by international organisations to keep them accountable to fundamental rights protection.

Monique Munarini is a PhD candidate in the Italian National PhD in AI at the University of Pisa. She is also a Brazilian qualified lawyer who received a Master in Human Rights and Multi-Level Governance from the University of Padova in Italy and a LLM in Law, Economics and Management from the University of Grenoble-Alpes in France. Her research involves gender mainstreaming strategies in AI Governance and accountability methodologies. Her research interests include AI ethics, business and human rights, AI due diligence, the impact of AI on women's rights and human rights.

Matteo Fabbri

An ethical perspective on the new transparency requirements for recommender systems set by the Digital Services Act

In the contemporary information age, recommender systems (RSs) play a critical role in influencing online behaviour: from social media to e-commerce, from music streaming to news aggregators, individuals are constantly targeted by personalized recommendations suggesting contents that may interest them. Despite such diffusion, the extent to which automated recommendations influence users' decisions is still underexplored, given that independent audits on the structure and functioning of RSs deployed on online platforms are usually prevented by proprietary constraints. The nudging potential of RSs can represent a risk for vulnerable people: indeed, judicial cases involving platforms' responsibility for displaying recommendations that may lead to political radicalization or endangerment of minors have recently caught public attention. The Digital Services Act of the European Union (DSA) is the first supranational regulation that sets specific transparency and auditing requirements for RSs implemented by online platforms with the aim of enhancing users' self-determination: in particular, apart from requiring platforms to explain clearly in their terms and conditions how their recommender systems work, it may allow users to modify the parameters on which automated recommendations rely so to let them choose autonomously the kind of content that they want to see. This paper focuses on how the enforcement of this regulation can mitigate the unfair consequences of the power imbalance between online platforms and users. To this aim, I firstly

outline the ethical and social implications of RSs starting from the consideration of automated recommendations as a multistakeholder phenomenon. Then I discuss the risks arising from digital nudging based on RSs and propose explanations as a tool that can reduce the impact of those risks by increasing users' awareness. Through a comparative analysis of relevant articles of the DSA, the General Data Protection Regulation (GDPR) and the AI Act, I outline how the provisions of the DSA fill some of the gaps left by other European regulations, while leaving the right to explanation not clearly determined. As a result of this analysis, I argue that, in order for the implementation of the DSA provisions on recommender systems to be effective, policy-makers should: 1) enhance users' awareness through clear and easily accessible explanations on how the recommendation process works and how they can be influenced by it; 2) grant users the possibility of intervening directly on the way in which RSs target them on the platform's interface.

I am a PhD candidate in Cybersecurity at IMT School for Advanced Studies and the University of Florence, Italy, currently carrying an industry internship in AI governance at BMW Group, Munich. My research, situated within the ethics of AI, concerns the impact of digital nudging through recommender systems on individuals' decision-making and self-determination. After obtaining a bachelor's degree in Philosophy from the University of Bologna (2020), I completed an MSc in Social Science of the Internet from the University of Oxford (2022), an MA in Sociology and Global Challenges from the University of Florence (2022) and a Postgraduate Diploma in Political and Social Sciences from Scuola Normale Superiore (2022). Moreover, I spent periods as visiting student at the University of Warwick and Ecole Normale Supérieure in Paris (ENS-PSL) and undertook a research traineeship at Imperial College London.

9. DILETTA HUYSKES

Enforce ethics & rights, control digital technologies: some empirical insights from civil society, auditing and research

We have long heard about the need for control and accountability in the development and use of artificial intelligence systems. Institutions, research centers and civil society are discussing it. The determinist and solutionist approach has always accompanied technological development, and recent developments in AI are making it worse. But how do we push this cultural change? Beyond theory, there are several ways to deal with this in practice. Working on the ethics and social impacts of technology may result in many different things, but it means bringing together research, advocacy, and empirical work to support and demand the development of technologies through participation, human control, and accountability.

Diletta Huyskes works on the social impact and ethics of technologies. She studied Philosophy and is a PhD Candidate in Sociology at the University of Milan, with a research on the use of algorithms and automated decisions by public agencies and the values that guide their design and governance. She has been working on artificial intelligence and anti-discrimination since 2019, when she also joined Privacy Network, an Italian digital rights association as the Head of Advocacy & Policy and coordinator of the Automated Administration Observatory, the first national mapping of impactful algorithms used by public administrations. In 2023 she co-founded Immanence, a benefit company that assesses digital technologies and offers solutions to ensure ethics and accountability. She is writing a book on the historical relationship between gender, feminism and technology to be released in 2024 and often addresses these topics with lectures and workshops also in institutional settings.

10. AUTONOMY & TRANSPARENCY

Jose Luis Guerrero Quiñones

AI opacity VS patients Autonomy in decision-making

The use of Artificial Intelligence (AI) in healthcare contexts is highly controversial for the (bio)ethical conundrums it creates (Floridi et al., 2021; Morley et al., 2020; Price II, 2015). One of the main problems arising from its implementation is the lack of transparency of Machine Learning (ML) algorithms, which are thought to impede the patient's autonomous choice regarding their medical decisions (Christian Bjerring & Busch, 2021). If the patient is unable to clearly understand why and how an AI algorithm reached certain medical decision, their autonomy is being hovered. However, there are alternatives to prevent the negative impact of AI opacity in shared (healthcare professional - patient) decision-making processes, and benefit from the high accuracy of such systems (di Nucci, 2019; London, 2019; Schönberger, 2019).

However intuitive, the thesis seems to miss the point of who is the object of trust, that is, in medical decision-making, patients trust their practitioners, not the tools they employ to execute their work. Thus, the relevant feature here is that trust occurs between two persons, patients need to trust their doctors, nurses, and other professionals, not the AI system, or their stethoscope, aiding them. If what truly matters is to trust the practitioner, the opaqueness of AI should not pose any novel problem as long as the medical professional is clear about what she does and does not know regarding all relevant aspects for the patient to make a rational and autonomous decision (Kerasidou et al., 2021). It would be enough to rely on the epistemic accuracy of AI systems when informing both healthcare practitioners and patients about diagnostic and treatment. Medical uncertainty caused by black-box medicine still enables practitioners to provide patients with useful information regarding their medical condition to encourage rational and autonomous deliberation (Braun et al., 2021).

To conclude, I would like to offer an alternative to increase trust in practitioners employed AI systems which also facilitates a shared decision-making process: the incorporation of value flexibility in medical AI systems (Savulescu & Maslen, 2015). A value-flexible design of AI could significantly facilitate and ultimately enhance medical decision-making. So far, most of the AI systems employed in healthcare settings are merely value sensitive, that is, they reflect general shared values within a determined society where they are used. A further development of AI, able to incorporate individual values (Meier et al., 2022), would improve medical decision-making, for the patient's values would be a primary parameter that the AI system would include on its algorithmic process to offer the best treatment ranked accordingly to the patient's own values and preferences. Similarly, patients seeing their values reflected on the AI process could increase trust on practitioners who use AI systems as medical tools to offer better services, as well as empower patients to use such technologies.

Jose Luis Guerrero Quiñones is a Visiting Research Fellow at the Institute for Ethical AI (Oxford Brookes University). After completing his PhD in Bioethics at the same institution, his current research focuses on the impact of AI in healthcare settings. He has also completed two masters, in Bioethics, and Education, and a degree in Philosophy at the University of Granada. He has published on the topics of the duty to die and biopolitics in various international journals.

Marten Kaas

A Perfect Technological Storm: Artificial Intelligence and Moral Complacency

Artificially intelligent machines are different in kind from all previous machines and tools. While many are used for relatively benign purposes, the types of artificially intelligent machines that we should care about, the ones that are worth focusing on, are the machines that purport to replace humans entirely and thereby engage in what Brain Cantwell Smith calls "judgment." As impressive as artificially intelligent machines are, their abilities are still derived from humans and as such lack the sort of normative commitments that humans have. So while artificially intelligent machines possess a great

capacity for “reckoning,” to use Smith’s terminology, i.e., a calculative prowess of extraordinary utility and importance, they still lack the kind of considered human judgment that accompanies the ethical commitment and responsible action we humans must ultimately aspire toward. But there is a perfect technological storm brewing. Artificially intelligent machines are analogous to a perfect storm in that such machines involve the convergence of a number of factors that threaten our ability to behave ethically and maintain meaningful human control over the outcomes of processes involving artificial intelligence. I argue that the storm in the context of artificially intelligent machines makes us vulnerable to moral complacency. That is, this perfect technological storm is capable of lulling people into a state in which they abdicate responsibility for decision-making and behaviour precipitated by artificially intelligent machines, a state that I am calling “moral complacency.” I focus on two salient problems that converge to make us especially vulnerable to becoming morally complacent and losing meaningful human control. The first problem is that of transparency/opacity. The second problem is that of overtrust in machines, often referred to as the automation bias. I examine each of these problems and how together they threaten to render us morally complacent before offering a potential solution that might allow us to resist moral complacency and retain meaningful control over artificially intelligent machines.

Marten Kaas is a philosopher working in the field of AI ethics with a research focus on the ethics and societal trust in artificially intelligent and autonomous systems. He has a Ph.D. in Philosophy from University College Cork in Cork, Ireland and is currently a Research Associate with the Assuring Autonomy International Programme at the University of York, UK. His current work is elucidating the concept of transparency in the context of artificially intelligent and autonomous systems, and in particular transparency's connection to safety assurance and its impact on enabling or impairing ethical principles.

11. MEANINGFUL HUMAN CONTROL

Alessio Tartaro

Control, Autonomy, and Responsibility: A Preliminary Exploration in the Context of Autonomous Systems

This paper presents an ongoing work-in-progress investigation into a complex question increasingly pressing in our era of fast-evolving autonomous systems: how can humans be responsible for actions and outcomes that are generated and influenced by autonomous systems and over which humans may have limited or no direct control?
This problem is raised by the growing use of autonomous systems that are increasingly performing tasks once exclusively handled by humans. Complex systems, such as Lethal Autonomous Weapons Systems (LAWS), Autonomous Vehicles (AV), and Autonomous Decision-Making Systems (ADMS), make it difficult to assign responsibilities to human actors embedded within these socio-technical systems, thus raising “responsibility gaps”.
In response to these issues, we have seen the emergence of concepts such as “meaningful human control” (MHC). Although numerous discussions on this subject have been enlightening, a more detailed and philosophically nuanced analysis of the core terms is needed. Without this in-depth scrutiny, the answers we provide might not fully capture the complexity of the question at hand.
This paper begins with a critique of certain inconsistencies characterizing the debate around meaningful human control. Specifically, the paper challenges the seemingly straightforward correlation between “having control over” and “bearing responsibility for” the outcomes of autonomous systems. This assumption underlies the discussion of meaningful human control but appears to be a logical fallacy. It is indeed reasonable to believe that without control, there cannot be responsibility for a particular outcome. However, this statement cannot be uncritically reversed to

imply that if there is control, then responsibility for the outcome is automatically assured. The reality, particularly within the context of autonomous systems, is substantially more nuanced.
This paper argues for a shift in perspective. Building on a foundational assumption of Western moral philosophy, the paper contends that autonomy, rather than control, serves as a fundamental requirement for responsibility. This opens up hitherto unexplored issues in the debate on human control over AI systems, namely the relationship between control, autonomy, and responsibility.
This paper aims to provide a conceptual clarification of the terms in question and their mutual relations. This examination will provide the basis for the formulation of a number of hypotheses aimed at addressing the discourse on human control over AI systems. These hypotheses should serve as a guide for a more coherent exploration of how artificial intelligence, by changing the dynamics of control over tasks previously handled exclusively by humans, affects human autonomy and responsibility.

Alessio Tartaro is a PhD student at the University of Sassari, Italy. Trained as a philosopher and a graduate of the Scuola Normale Superiore, Pisa, he is currently conducting an industrial doctorate, in which he explores the ethical, social, and legal implications of artificial intelligence.

Sabine Thürmel

First steps towards Meaningful Human Control of Generative AI

In the past years, the public perception of AI has received a major boost due to a variety of tools for generating text, code, digital images as well as audio. Generative AI enables the discovery of abstract patterns of co-occurrence in very large datasets. These patterns are then used to generate (statistically relevant) content. Such digital output appears to humans as if it were created in the worst case by a statistical parrot or some hallucinating being and in the best case by a fellow human. It is therefore not surprising how strong and diverse the reactions are to this technology: the general public is in awe and uses such apps for recreational purposes, whereas creative professionals fear for their livelihoods and see digital forgery. In the long run video or audio shots, presented on the Internet, mostly cannot be taken at face value nor will they be valued as an authentic artistic expression. Creation of texts will generally no longer be perceived as the output of an intellectual effort.
Production without understanding the subject matter, as realized by Generative AI, leads to many challenges for meaningful human control, that is AI systems where humans are ultimately in control. One exemplary control task is the prevention of the generation of toxic digital content (which in the case of some currently prominent tool is currently supported by third world workers earning a pittance). Automatic removal of unwanted content is some way off.
One central building block for meaningful human control, currently in its embryonic stage, is Explainable AI: The goal is to combine neural networks with reasoning and structured representations in order to generate digital output in a way which at least partly explains its reasons to humans. In the meantime, a broad education offensive on digital literacy, data literacy and AI literacy is necessary in order to empower humans to maintain a critical stance towards the output of such systems.
Production without understanding may diminish the appreciation of human efforts in these fields, too. Therefore, digital content created by humans or AI-assisted production of content must show some value add in order not to fall under the general suspicion to be only content solely created by machines based on vast amounts of data. Humans could make the most of conceptual thinking and their reasoning capabilities to explain digital works created by themselves or with support of Generative AI. Moreover, one could closely collaborate with an AI tool as some composers do even today. These artists train the AI with their compositions and create AI assisted oeuvres where the human composer takes up the themes generated by the AI. Thus, a piece is created where the parts composed by a human and the AI generated parts respond to each other. This is an example of meaningful human control, feasible even today.

I am an independent researcher and lecturer at the Technische Universität München (TUM), Munich Germany. My background is both in computer science (Ph.D. in Computer Science from TUM in 1989, being the fourth woman to receive this degree at TUM) and philosophy (Ph.D. in Philosophy of Science and Technology from TUM in 2013). From 1989 to 2009 I worked as a senior researcher and technical strategy advisor in industry. I possess a wide-ranging computer science experience. Since 2009 I solely focus on philosophy of technology. However, I have done interdisciplinary work on the foundations and effects of culture changing information technologies since the 1990s. For more information see http://www.sabinethuermel.de/ where computer science meets philosophy.

Martina Philippi

Talking to robots: The role of human value in AI-driven communication systems

In the topic of AI-driven systems, there is a special interlocking of ethical and epistemological questions. The ethical relevance of the topic is obvious: We are dealing here with a new kind of technology that intervenes deeply and presumably irreversibly in our lifeworld. It is to be expected that it will fundamentally shape our routines and our relationship to the world - in line with the dictum of the philosophy of technology "we form tools, then tools form us". This intervention consists not only in the fact that AI systems - especially those based on neural networks - can act and decide unpredictably. Most importantly, we cede some of our decision-making latitude to the systems when we rely on them. The human value of autonomy in this context is not only something to be protected from being altered unnoticed in undesirable ways by a technology. It is also the subject of an epistemological exploration of its conditions, which can be carried out phenomenologically: How are the diversity of world perception, the freedom in prioritizing options for action, and the liveliness of communication shaped and, under certain circumstances, impaired? My point is that the value of autonomy can become philosophically thematic in three ways: first, as a good to be protected; second, as a starting point for epistemological questions; and third, as an epistemologically grounded standard for judging technologies that is indispensable to retain meaningful human control.

This intertwining of ethical and epistemological issues in relation to AI is rooted in a peculiarity of technology that is particularly strong here. In the responsible design of technology there is a conflict of interests, namely between the smooth functioning on the one hand, which according to common positions in technology philosophy and media theory leads to a 'becoming invisible' of technology; and on the other hand, keeping open an understanding entry into the depth of technology, into its functioning and its effect on us. This is a particular problem with systems that we interact with directly. Communication systems like social chatbots (e. g. Replica) are a special case, but I think that the peculiarities I am concerned with can also be applied to decision support systems or other communicating systems. Although the communication situation between humans and machines as autonomous subjects is different than between humans, it is still an interaction. Communicating with a machine in this way is different from operating a machine. In this way, certain cognitive processes are set in motion that we are otherwise familiar with from interpersonal interaction. This can be situational adjustment, but in the long run it can also have a formative effect on our communication, our prioritization and our perception as we know it from socialization.

AI withdraws from our control, and that is intentional, since after all, it is essential to communication not to know in advance what the other person is saying. We must relinquish control, because we want to be surprised or at least - as in the case of systems that give us instructions, such as the talking supermarket checkout, the robot hotline, or the navigation system - we would like to be partly relieved of decision-making and forward-looking action. Ethically, therefore, it is necessary to build at least one door into the new technology that can be opened for regular reflection in the sense of technology-philosophical monitoring as needed. The key to this door is a well-reflected understanding of the relevant values that can be used to assess the technology in question and its impact on our lives and

our interactions. In my presentation, I highlight that the value of autonomy needs to be epistemologically refunded against the backdrop of novel technologies and their applications to be exactly this key to reflection. In my paper, I will use the example of AI-enabled communication systems to show how this can be done.

Martina Philippi studied philosophy, logic and philosophy of science as well as general and comparative literature in leipzig and wrote her doctoral thesis on the problem of self-evidence ("Selbstverständlichkeit") in Edmund Husserl's phenomenology. She is currently doing research in the BMBF project "UAV-Rescue" on UAV-borne remote sensing for AI-assisted support of search and rescue missions. Her current work focuses on the philosophy of technology and the ethics of AI.