UNIVERSITÄT BONN CENTER FOR SCIENCE & THOUGHT

CST AI RESEARCH GROUP

PRESENTATION

The guiding aim of the CST AI research group is to provide a forum in which researchers can gain new perspectives on their own work and regular input from across the university on topics surrounding **AI**, **digital transformation** and **technology studies**.

Our hope is that this will foster **transdisciplinary** work and facilitate interchange and cooperation between scholars focusing on overlapping areas of research, but who happen to be in different departments.

This AI research group currently includes researchers in philosophy/ethics, medicine, (neuro-)biology, political science, law, psychology, physics, mathematics, IT and is open to all researchers interested in AI.

We are now meeting on Mondays 10-11.45 am (CEST) via Zoom. If you or someone from your team wishes to join, please send an email to Dr. Charlotte Gauvry (**cgauvry@uni-bonn.de**), and she will add you to the mailing list. Should you wish to present or discuss a paper, a work in progress, or any question regarding AI, please let her know.

CST AI RESEARCH GROUP

2021 SUMMER SEMESTER PROGRAMME

Mondays, 10-11.45 am (CEST)

https://uni-bonn.zoom.us/j/93191701586?pwd=a2IrdVpTNjJCSzJReHVSVzBNNDFkUT09
ID: 931 9170 1586
Password: 267881

Contact: Dr. Charlotte Gauvry – cgauvry@uni-bonn.de

Mar 15

Dr. Oliver Braganza (Neurology) "Proxyeconomics and Goodhart's Law - the downside of optimization"

Abstract: Competitive societal systems by necessity rely on imperfect proxy measures. For instance, profit is used to measure economic value, the Journal Impact Factor to measure scientific value, and clicks to measure online engagement or entertainment value. However, any such proxy-measure becomes a target for the competing agents (e.g. companies, scientists or content providers). This suggests, that any competitive societal system is prone to Goodhart's Law, most pithily formulated as: 'When a measure becomes a target, it ceases to be a good measure'. Purported adverse consequences include environmental degradation, scientific irreproducibility and problematic social media content.

Despite the intriguing similarities and profound societal relevance of these phenomena, research on the underlying mechanisms is fragmented between disciplines, facing significant practical and philosophical challenges. Fortunately, interest in Goodhart's Law has recently surged in AI-safety research, which is beginning to systematically explore the mathematical underpinnings of the phenomenon. The talk will explore the notion that a systematic research program into Goodhart's Law, nascent in current AI-safety research, is urgently needed, and will have profound implications far beyond AI-safety.

Dr. Uwe Peters (Philosophy, CST Bonn & LCFI Cambridge) "Supersizing Confirmation Bias"

Abstract: *The hypothesis of extended cognition (HEC), i.e., the view that the realizers of mental states or cognition can include objects outside of the skull, has received much attention in philosophy. While many philosophers have argued that various cognitions might extend into the world, it has not yet been explored whether this also applies to cognitive biases. Focusing on confirmation bias, I argue that a modified version of the original thought experiment to support HEC helps motivate the view that this bias, too, might extend into the world. Indeed, if we endorse common conditions for extended cognition, then there is reason to believe that even in real life, confirmation bias often extends, namely into computers and websites that tailor online content to us. The view that confirmation bias extends into artefacts in this way is metaphysically interesting and has significant ethical and epistemic benefits. It helps us become more alert to, and better protect ourselves against, online manipulation than the alternative view that the bias does not literally extend into artefacts but is merely causally linked to them.*

Mar 22

Dr. Scott Robbins (Philosophy/Ethics) “What Machines Shouldn't Do: Meaningful Human Control over the Evaluative”

Abstract*: In an increasingly autonomous world, it is becoming clear that one thing we cannot delegate to machines is moral accountability. Machines cannot be held morally accountable or responsible for their actions. This becomes problematic when a machine has an output that has a significant impact on human beings. Examples of machines that have caused such impact are widespread and include machines evaluating loan applications, evaluating criminals for sentencing, targeting and killing people, driving cars, and whatever digital assistants are supposed to do. The question that governments, NGOs, academics, and the general public are (or should be) asking is: how do we keep meaningful human control (MHC) over these machines? I argue that keeping meaningful human control over machines cannot be achieved without restricting the decisions we delegate to machines to the descriptive. It must always be a human being deciding how to employ evaluative terms as these terms not only refer to specific states of affairs but also say something about how the world ought to be. The main reason for this is that machines delegated evaluative outputs cannot, in principle, be discussed in terms of efficacy. When we delegate the decision to decide who is the 'best' candidate to a machine, we are doing so with no way of verifying that the machine was correct. If we cannot say whether or not a machine was correct on any single output - we cannot say what its overall accuracy is. We might as well use a magic 8 ball. Delegating a task to a machine for which we can say nothing regarding its accuracy is a waste of time and resources at best, and horrifyingly unethical at worst. When we delegate evaluative outputs to machines we are also delegating the task of coming up with, and debating competing conceptions of what should ground the words "good" and "right" in a given context. This not only degrades meaningful human control over the most important concepts humans have, but could lead to a vicious cycle in which our considered ideas of what is good and right are changed by opaque machines that may be using unethical or - equally troubling - irrelevant considerations to generate their outputs.*

Dr. Charlotte Gauvry (Philosophy) “An optimistic view on the merging with AI? On Self-building technologies”

Abstract*: The recent literature has convincingly demonstrated certain perverse effects of what has been called the “merging with AI.” Susan Schneider (2019) even argued that different merging techniques, for example the implantation of silicon chips in our brains, could lead to a dissolution of our selfhood. In my talk, I propose to discuss this hypothesis based on a recent provocative article by François Kammerer, which considers, on the contrary, the possibility of "self-building technologies:"artificial technologies (paradigmatically some digital app) would actually allow us to “strengthen our selfhood.”*

Kammerer’s paper is based on two playful thought experiments (“Pr. Truffle the implicit racist and iDiversity®” and “Emma the inconstant wife and iFidelity®”). For the sake of the discussion, I will rely only on the first thought experiment in order to assess the validity of the criteria developed to define selfhood and self-building technologies.

Mar 29

Dr. Apolline Taillandier (Philosophy/Political Science, CST Bonn & LCFI Cambridge) "Taming Superintelligence: governing global change through AI safety"

Abstract: *This presentation focuses on the recent history of "AI safety," understood as a set of strategies for mitigating long-term AI risk. AI safety brings together computer science and policy expertise: while some AI safety researchers have redesigned algorithmic training methods for ensuring control over the goals or behavior of "human-level" AI, or systems with "artificial general intelligence," (AGI) AI safety policy experts have developed scenarios and metrics of AI progress in order to anticipate and prepare for uncertain technological breakthroughs through specific models of governance. I show how AI safety expertise intersects transhumanist debates about the singularity and the possibility to predict and tame the emergence of a posthuman kind of intelligence; economic models of rational behavior in computer science; and a realist model of global security >as a product of interest-based cooperation and alignment.*

Apr 12

Dr. Scott Robbins (Philosophy/Ethics) "Meaningful Human Control: A conceptual analysis"

Abstract: *Meaningful human control (MHC) is an oft-used phrase in the ethical literature on artificial intelligence (AI). But what is meaningful human control (MHC)? Let's break that phrase down because it is quite ambiguous. First, we have 'meaningful'. Meaningful could be modifying 'human'. This would mean that what we are searching for is control over how to be a meaningful human in the age of AI. Or is it modifying 'control'? Then MHC would be about having a human having control over a particular AI system in a meaningful way (whatever that amounts to). Meaningful could also simply be meant as a noun. In which MHC would be about retaining our control over what is considered to be meaningful (and why it is considered to be meaningful). In this case AI is interpreted as a threat to our ability to have such control.*

Second, we have the word 'human'. This could simply be a modifier for 'control' – as in the type of control we are looking for is human-like control. This opens up the possibility that a machine could be in control as long as the control it exerted was human-like. However, 'human' could also be referring to the 'who' that is supposed to be in control. The first paragraph of this paper is littered with 'we' and 'our' without specification of who these words refer to. This is because it is ambiguous. It could refer to the designer of a particular AI system. The designer could choose specific methodologies or training data sets that increase their ability to control the possible outcomes of a particular AI system. 'Human' could also refer to the operator of the AI system who has some kind of control over, for example, the specific outputs or consequences of that AI system. 'Human' may instead refer to the subject or subjects of a particular AI system. Those people that are being surveilled, categorized, and surveilled by AI may also have a right to some control – control of, for example, how they conceive of, and achieve, the good life. Finally, 'human' could refer to society at large having control over the contexts that AI can be used in. There may be contexts in which too much control is lost by having AI in them.

Third, and last, we have 'control'. What is it that we are having control over? In the most obvious sense, MHC refers to control over the outputs, or consequences, of a particular AI system. However, we could also be referring to society's control through laws and norms governing the acceptability of AI operating in certain contexts (as mentioned above). So it is not a particular AI system that we have control over, but the integration of AI into society that control refers to. Finally, 'control' could simply be had over 'the good life' broadly speaking. If it is the subjects that are the 'humans' in question, then they may be deserving of control over how they conceive of and achieve the good

life. AI which dictates who gets a job may be steering our conceptions of what a good job candidate is. Delegating to AI how 'good' and 'bad' are applied to people is a loss of control over what we conceive to be the good life.

Apr 19

Dr. Uwe Peters (Philosophy, CST Bonn & LCFI Cambridge) "When artificial intelligence outperforms human social cognition"

Abstract: *Suppose an artificial intelligence (AI) excels us in social cognition in that it ascribes psychological characteristics to people more accurately than us. It seems epistemically rational to then set aside our own judgments on those characteristics and defer to the AI. After all, being more accurate about people is better than being less accurate. I argue that this intuitive view is epistemically and ethically problematic. It overlooks three important points. (1) If the AI produced a verdict on our psychology that we disagree with, it would often remain epistemically rational for us to retain our own judgment on the matter because we can make that judgment true. The AI might predict this. But it can't factor it into the verdict it provides to us without becoming inaccurate. (2) An oversight of point (1) may result in a reduction of our autonomy to determine what persons we are and wish to become. (3) While adopting an AI that is highly accurate in social cognition may help us correct our social misconceptions, it is also likely to reduce the benefits that frequently arise when inaccurate ascriptions of adaptive features to people become self-fulfilling.*

Apr 26

Prof. Dr. Joachim Schultze & Stefanie Warnat-Herresthal (Medical School, Limes Institute & DZNE) "Swarm Learning as a fully decentralized and confidentiality-enabling machine learning approach for disease classification"

Abstract: *Fast and reliable detection of patients with severe and heterogenous illnesses is a major goal of precision medicine. We recently illustrated that leukemia patients are identified by machine learning (ML) based on their blood transcriptomes. However, there is an increasing divide between what is technically possible and what is allowed because of privacy legislation. To facilitate integration of any medical data from any data owner world-wide without violating privacy laws, we here introduce Swarm Learning (SL), a decentralized machine learning approach uniting edge computing, blockchain-based peer-to-peer networking and coordination as well as confidentiality protection without the need for a central coordinator thereby going beyond federated learning. To illustrate the feasibility of SL to develop disease classifiers based on distributed data we chose four use cases of heterogeneous diseases including COVID-19, tuberculosis, and leukemias. With more than 16,400 blood transcriptomes derived from 127 individual clinical studies with non-uniform distribution of cases and controls and significant study biases as well as over 95,000 chest X-ray images, we illustrate that SL classifiers outperform those developed at individual sites. Still, SL completely protects local confidentiality regulations by design. We propose this approach to noticeably accelerate the introduction of precision medicine.*

May 10

Dr. Tobias Keiling (Philosophy) “Controlling Evil. Norbert Wiener and cybernetic metaphysics”

Abstract: *Tobias Keiling will present and comment Peter Galison’s Paper on Norbert Wiener’s critique of the worldview of cybernetics.*

See Peter Galison. 1994. “The Ontology of the Enemy: Norbert Wiener and the Cybernetic Vision,” Critical Inquiry *Vol. 21/1.: 228-266.*

May 31

Prof. Dr. Wolfgang Koch (Computer Science, FKIE) “On Digital Ethics for Artificial Intelligence and Information Fusion in the Defense Domain”

Abstract: *For knowledge itself is power... Francis Bacon’s statement on achieving power as the meaning of all knowledge marks the beginning of the modern project. At the latest since the advent of AI in the defense domain, however, technology meant for the benefit of humanity may turn against humanity. This specific type of instrumental knowledge makes the modern crisis as visible as in spotlight. Ethical knowledge of man and his nature must complement Bacon’s knowledge. There is an 'Ecology of Man': he does not make himself; he is responsible for himself and others. How can the AI and information fusion community technically support responsible use of the power we are harvesting from AI and Fusion? To argue more specifically, we consider documents of the German Bundeswehr, founded in the 1950s when the term AI was coined. Since these Armed Forces have learnt lessons from 'total war' and tyranny, they seem conceptually prepared for mastering the digital challenge. A current example where these questions are pressing is the European Future Combat Air System FACS. There also exist parallels to on-going IEEE P7000 standardization activities on “Ethically-aligned Engineering.”*

June 14

Dr. Uwe Peters (Philosophy, CST Bonn & LCFI Cambridge) “Extended Mindreading and the Tracking of Digital Footprints”

Abstract**:** *There has been much philosophical and empirical research on mindreading, our ability to attribute mental states to people to explain, predict, or influence their actions. It is commonly assumed that mindreading is situated within an individual’s head. Here I challenge this assumption by relating it to Clark and Chalmers’ hypothesis of extended cognition, the view that the physical realizers of cognition can include artefacts outside the body. I argue that if we endorse common conditions for extended cognition, then there is reason to believe that people’s mindreading is often partly realized by (vs. merely causally coupled to) artefacts, specifically, computer algorithms on the Internet that track our digital footprints to infer our preferences, interests, etc. from them to personalize websites. The view that mindreading extends outside the body in this way has explanatory advantages that the alternative proposal that people use these artefacts merely as epistemic tools lacks. It offers a novel perspective on mindreading, extended cognition, and the tracking of digital footprints.*

June 28

Dr. phil. Johannes Lierfeld (Ethics/Technology, Centesimus Annus Pro Pontifice), speaker & Pater Dr. Dr. Justinus C. Pech (Institut für Führungsethik & Centesimus Annus Pro Pontifice), respondent: "Bridges between protein and silicon: Ethical challenges of merging biological and artificial cognition"

Abstract: *The notion of a bridge between protein and silicon has slowly evolved into a reality due to the development of brain-computer interfaces or brain-machine interfaces (BCIs / BMIs). Besides some alarmism there is much hope to achieve wonderful things with and through the use of BCIs. However, as fruitful the applications can be in the best case, as severe the ethical repercussions would be in a worst-case scenario. Thus, serious ethical challenges monger around the usage of these mind-machine-bridges, since their applications might interfere with personal integrity, identity, and accountability. Especially the field of affective BCIs poses cascades of ethical concerns regarding their operative field: human affections and emotions. Which kinds of applications are desirable, and which must be avoided at any costs? Furthermore, which of the ambivalent applications may not be forbidden, but bound to a voluntary decision of consent that has to be acted out by the individual?*

July 12

Prof. Dr. Marion Gymnich (British Literature) "AI in two contemporary British novels: Ian McEwan's *Machines Like Me* (2019) and Kazuo Ishiguro's *Klara and the Sun* (2021)"

Abstract: *It stands to reason that fictional representations of AI in general, and of androids in particular, which can be found across a wide range of science-fiction novels, movies and TV series, have had a considerable impact on how many people are likely to imagine AI today. In recent years, two of the most renowned contemporary British writers have published novels that feature androids very prominently: Ian McEwan's Machines Like Me and Nobel laureate Kazuo Ishiguro's Klara and the Sun. Machines Like Me is an alternate history, set in the 1980s in a world that has experienced enormous technological progress, largely due to the work of Alan Turing, who is still alive in this 'what if' narrative. Androids that can be purchased by a few, wealthy people are among the most striking technological innovations. In Ishiguro's Klara and the Sun, which is set in the United States at an unspecified point in time in the future, so-called Artificial Friends have become a commodity for the children of the well-to-do. In addition to examining what the androids are shown to be capable of in the two fictional universes, I will focus in particular on the psychological and ethical dimensions of the relationship between humans and androids that is portrayed in the novels by McEwan and Ishiguro.*