CST AI RESEARCH GROUP

2021-22 WINTER SEMESTER PROGRAM

Mondays, usually from 4:15 to 5:45 pm CET (special times indicated below)

In presence: IZPH, Meeting Room (3d floor): Poppelsdorferallee 28 – D-53115 (Bonn)

Zoom link for the online or hybrid sessions:

https://uni-bonn.zoom.us/j/94884157465?pwd=VW1SdnVDeEdqa3VOeUJyUjkwZ25Idz09
ID: 948 8415 7465
Password: 731091

October 2021

Oct 4 – **Dr. Oliver Braganza** (Medical School) "Unethical optimization principle" (online)

Abstract: *Oliver Braganza will present and comment Beale, Battey & Mackay* (2020)*'s paper: "An unethical optimization principle"*

Beale N, Battey H, Davison AC, MacKay RS. "An unethical optimization principle," *R Soc Open Sci.* 2020 Jul 7(7): 200462.

The paper is essentially a mathematical proof based on the type of state space formalization pioneered in cybernetics, which is now the foundation most modern AI research/ theoretical neuroscience and more generally systems science. I'll first concentrate on introducing this type of reasoning and modeling in a hopefully interactive form (potentially highlighting some other seminal historic studies in this tradition). I'll then present the basic premise of the paper and the main finding, which I think can be done quite concisely on the basis of a single formula (T(s) = A(s) – C(s) + Q (s), where s is some single state of the world T, (s) is the total 'value' of that state, A(s) is an action of an AI leading to that state (or more precisely the reward value of the AI given that action), C(s) is an ethical cost diminishing the value of the state and Q is some random noise). The paper proves, that as long as there exist some actions with which the AI can get higher 'rewards' (according to its programmed reward function) but which entail an ethical cost (unbeknownst to the AI), then the optimization will tend towards those actions. However, then I'll highlight how curiously narrow the papers conception of 'ethics' is (it treats ethics as 'risks to future profitability') and argue that this seriously undermines the true scope of their findings.

Oct 11 – Dr. Uwe Peters (CST & LCFI Cambridge) "Algorithmic political bias: Cause for special concern" (online)

Abstract: *Some artificial intelligence systems can display algorithmic bias, i.e., they may produce outputs that unfairly discriminate against people based on their social identity. Much research on this topic focuses on algorithmic bias that disadvantages people based on their gender or race, and the related ethical problems are widely discussed. Is algorithmic bias against other aspects of people's social identity, for example, their political*

orientation equally problematic? This question is so far largely unexplored. Focusing on the epistemological issue of how we may detect algorithmic biases and recognize their harmfulness, I argue that algorithmic bias against people's political orientation differs from algorithmic gender and race biases in important ways. The reason is that there are strong social norms against gender and race biases, but this is not the case forpolitical biases. Political biases can thus more powerfully affect individuals' cognition and behaviour. This increases the chances that they become embedded in algorithms. It also makes it harder to detect and eradicate algorithmic political biases than gender and race biases even though they all can have similarly harmful consequences. Algorithmic political bias thus raises hitherto unnoticed and distinctive epistemological and ethical challenges.

Oct 18 – **Prof. Jens Schröter** (Digital Media) "How is Artificial Intelligence Changing Science? A Research Project" (online)

Link to the project: https://howisaichangingscience.eu/

November 2021

Nov 8, 4-6:30 pm – **AI & Mental Health I** (CST, Cassis, Medical School, TRA 4), **Prof. Tamar Sharon** (Radboud University) & **Dr. Saheli Burton** (University College London). More information on the CST website (hybrid)

Nov 29 – **Dr. Johannes Lierfeld** (Ethics/Technology, Centesimus Annus Pro Pontifice) "What it´s like to be another one: Philosophical zombies, data and the eternal question of the nature of qualia" (online)

Abstract: *In the age of artificial intelligence our world view is increasingly mechanistic. Reductive materialism seems to be able to answer everything, since most aspects of our lives appear to be representable in data. Cognition is thinking, thinking is brain activity, brain activity is either electrical or metabolic, and both forms of activity can be measured – hence, cognition can be* measured.

Moreover, the term "mind reading" suggests that artificial intelligence systems can also predict our minds. Recommender systems even anticipate the users next item of interest, and they do so with remarkable accuracy.

However, it is nothing but interpretations. Interpretations will surely reach more accurate levels of anticipation with more precise measurement methods and may access the said brain activities, for example through the advent of brain-machine interfaces. Yet, it is highly doubtful that we will ever be able to literally read the mind and decode thought itself. On the other hand, it is highly likely that – given the ever-increasing accuracy of interpretations – these methods will receive wide social acceptance. That, of course, might come with completely new ethical challenges. The first-person perspective appears as the ultimate custodian of qualia, and the subjectivity of the individual may never be objectifiable, regardless the means.

December 2021

Dec 6 – **Dr. Apolline Taillandier** (CST & LCFI Cambridge) “AI in a different voice: rethinking computers, learning, and gender difference at MIT in the 1980s” (hybrid)

Abstract: *This paper explores the “critical” AI projects developed around the Lego community at MIT in the mid-1980s. While a rich scholarship studies how programming and AI were made masculine, little has been said about those AI practitioners who drew on literary criticism and feminist epistemologies with the hope to overcome the “technocentric stage of computer discourse” and undo gender hierarchies underlying computer cultures and programming experimental standards. At MIT, AI researcher Seymour Papert and sociologist Sherry Turkle argued that cognitive theories of AI and intelligent behavior as flexible, intuitive, and object rather than task-oriented could help challenge the masculinist assumptions in formal AI and expert systems approaches, as well as the gendered labor division in computer science. Taking inspiration from “emergent AI” and “mind as society” models, more than from the earlier philosophical critique of AI as instrumental reason by Dreyfus, Weizenbaum, and Hofstadter, they tied computer programming to Piaget’s theory of intellectual development, Keller’s critique of objectivity and dominant scientific epistemology, and Gilligan’s moral psychology. At a critical moment of the history of AI projects, but also of debates about the social and moral responsibility of machine intelligence scientists, they sought to shift the social perception of computers for fear of backlash. Intersecting political history, feminist theory, and the history of science, this paper contributes to the “hidden history” of women and feminist activism in AI, to the material history of AI models and software, and to the history of AI as a human science located partly in the Harvard-MIT complex. This helps historicize recent discussions of the whiteness and masculinity of algorithms, but also to clarify the interweaving between discourses of gender difference and the sidelining of feminist agendas in computer professions from the 1980s onwards.*

Dec 13, 3-5:30 pm – **AI & Mental Health II** (CST, Cassis, Medical School, TRA 4), **Prof. Dr. Joanna Bryson** (Hertie School, Berlin) & **Prof. Dr. Elisabeth Hildt** (Illinois Institute of Technology). More information on the CST website (online)

January 2022

Jan 10, 10:15-11:45 am – **Dr. Sergio Genovesi** (CST) & **Dr. Julia Maria Mönig** (CST) “Evaluating Fairness in the Framework of a Trustworthiness Certification of AI Systems” (online)

Abstract: *Current publications on AI and fairness show that there is a need for a clear definition of fairness and that an ethical understanding of fairness exceeds the mere de-biasing of data and code. In this talk we make use of the interdisciplinary competence of our consortium and start from different definitions and understandings of "fairness". We are interested in those with regard to the certification of trustworthy AI. We will discuss which of the presented understandings of "fairness" can be operationalized in order to be able to certify what might be "fair". To illustrate, how a "fairness" certification can look like, we will discuss the use case of a credit loan algorithm, considering different fairness metrics from an ethical perspective.*

We stress that, while we agree that "de-biaising is not enough", making sure that removing bias is one way to look at fairness and to guarantee a certain equity.

Jan 17 – **Dr. Johannes Lierfeld** (Ethics/Technology, Centesimus Annus Pro Pontifice), "Robo Ethics" (online)

Jan 24 – **Prof. Wolfgang Koch** (Computer Science, FKIE) "Ethically Sensitive Applications of AI - Examples and Implications for Systems Engineering" (online)

Abstract: *"Intelligence" and "autonomy" are omnipresent in the biosphere. Before any scientific reflection or technical implementation, all living creatures fuse sensory impressions with learned and communicated information. In this way, they perceive aspects of their environment in order to act in accordance with their goals. In the complex technosphere, cognitive machines support human intelligence and autonomy via artificially intelligent automation, i.e. 'cognitive machines', by which they can increase their capabilities far beyond natural levels. Which requirements of systems engineering need to be fulfilled so that such machines take account of human beings using them as a responsible person?*

February 2022

Feb 7, 2:30-4pm **– Dr. Uwe Peters** (CST & LCFI Cambridge), "Negativity bias in research: Why comparisons between the transparency of artificial intelligence and human cognition are problematic" (online)

Abstract: *Artificial intelligence (AI) algorithms used in high-stakes decision-making contexts often lack transparency in that the internal factors that lead them to their decisions remain unknown. While this is commonly thought to be a problem with these systems, many AI researchers respond that we shouldn't be overly concerned because empirical evidence shows that human decision-making is equally opaque and isn't usually required to be more transparent. I argue that the empirical data on human cognition that are claimed to support this equal opacity view don't sufficiently support it. In fact, the equal opacity view rests on a narrow, selective, and uncritical survey of relevant psychological studies. Furthermore, there is reason to believe that many psychologists and AI researchers may have a negativity bias (a tendency to attend more to the shortcomings than the strengths of human cognition) that can contribute to systematic underestimations of the insights that people have into their decision-making. This issue and significant methodological limitations of existing studies on human cognition raise serious problems for reliable comparisons between AI systems and humans regarding their (lack of) transparency.*