Moral Psychology of Artificial Intelligence Conference

Monday 26th June 2023

Keynes College, University of Kent

Schedule at a Glance

8:30 - 9:00	Breakfast /Welcoming Remarks
9:00 - 9:30	Jean-François Bonnefon *The moral psychology of AI: A flash review*
9:30 - 10:00	Walter Sinnott-Armstrong *How to Build Human Morality into AI*
10:00 - 10:30	Flash Talks Session #1: Andreas Kappes *The Consequences of the Ascent of Artificial Intelligence for Human Cooperation* Beth Anne Helgason *How Unbiased Algorithms Can Exacerbate Human Bias* Jen Semler *Types of Moral Agents*
10:30 - 11:00	Coffee Break
11:00 - 11:20	Anne-Marie Nussberger *Characterizing Human Preferences for Interpretability in Artificial Intelligence*
11:20 - 11:40	Markus Kneer *Responsibility Gaps and Retributive Dispositions: Data from the US, Germany & Japan*
11:40 - 12:00	Michael Laakasuo *Moral Judgments of "Pre-Crime" Arrests and AI Profiling*

12:00 - 12:20	Yuxin Liu *The Moral Psychology behind Artificial Moral Advisors*
12:30 - 14:00	Lunch and Posters
14:00 - 14:30	John Danaher *Will large language models spark a moral revolution?*
14:30 - 15.00	Flash Talks Session #2: Daniel Shanks *Accent Prejudice Toward Apple's Siri* Daryl Cameron *Empathic Choices for Humans and Robots* Janet Pauketat *Disentangling the impacts of autonomy and sentience on the moral consideration and perceived threat of AIs*
15:00 - 15:50	Coffee Break
15:50 - 16:10	Yochanan Bigman *Algorithm discrimination and stereotype strengthening*
16:10 - 16:30	Flash Talks Session #3: Madeline Reinecke *The puzzle of evaluating moral behavior* *Simon Myers* *The Necessities and Luxuries of Trustworthy AI: The Perceived Importance of Different AI Characteristics*
16:30 - 17:00	Last Words / Wrap Up

17:00	Finish
19.30	Social Event in Cosy Club

SPEAKERS

Jean-François Bonnefon

The moral psychology of AI: A flash review

9:00 – 9:30

I will provide a rapid overview of the major themes and currents in the moral psychology of artificial intelligence. The presentation will be organized into three parts: AI as a moral agent, AI as a moral patient, and AI as a moral proxy. [moral agent] I will list the main questions that moral psychologists investigate when AI is expected to make risky decisions or balance moral values. [moral patient] I will describe the experimental economics-inspired research that examines the extent and reasons behind prosocial or cooperative behaviours displayed by humans towards machines. [moral proxy] I will consider AI's role in representing humans in moral interactions, in the form of machine delegation and machine masquerade.

Chiara Longoni

Implications of generative AI and LLMs for moral judgments

9:30 – 10:00

TBC

Andreas Kappes

The Consequences of the Ascent of Artificial Intelligence for Human Cooperation

10:00 - 10:10

Cooperation is at the centre of human morality, outlining the rules and norms that allow societies to function. However, AI-powered technologies are entering domains of cooperation that were considered exclusively human. This is becoming especially apparent in work life. While the work of some will be taken over by AI, most workers will cooperate alongside AI technologies; AI will increasingly fill the roles of a subordinate (e.g., journalists edit news

stories generated by AI), colleague (software developers co-create with AI), or manager (employees have to ask HR conversational agents for vacation). The benefits for organisations are obvious but how are people making sense of AI in these cooperative interactions and how does that impact human cooperative norms? We propose that employees bring the implicit models they have for human cooperation into the relationship with AI, helping them to make sense of these new interactions and informing their moral judgments, but also leaving people emotionally confused and changing human cooperative norms. Here, we report our research that examines how people perceive human-AI relationships to fulfil the functions of cooperative relations (i.e., hierarchical, caring, transactional), how those perceptions differ and resemble human-human relationships (e.g., rules of reciprocity and obligation), how they inform the moral judgments of cooperative actions performed by the human or AI, and the emotional and social consequences of bringing and adapting human-human models to AI-human cooperation.

Beth Anne Helgason

10:10 – 10:20

How Unbiased Algorithms Can Exacerbate Human Bias

There is widespread concern that AI algorithms may exacerbate racial and gender bias in important contexts such as hiring, policing, and healthcare. These discussions focus on how bias is built into unfair algorithms. Yet, even fair algorithms can exacerbate bias when algorithms are used by biased humans. In the present research, we examine how people's biases shape their use of algorithmic advice. In Study 1 (N = 387; 4567 observations), we find that individuals with negative attitudes toward Black people used advice to justify lowering their ratings of Black job applicants. Thus, individuals' racial attitudes predicted their ratings of Black job applicants more strongly in the presence, than absence, of algorithmic advice. In an ongoing study, we aim to replicate our results using a longitudinal mock-hiring task. Together, our findings suggest that algorithmic advice may liberate individuals to express prejudice under the guise of objectivity. We discuss implications for structuring joint human-algorithm decisions to avoid amplifying human's own biases.

Jen Semler **Types of Moral Agents**

10:20 – 10:30 We shouldn't deploy autonomous weapons systems. We shouldn't try to program ethics into self-driving cars. We shouldn't replace judges with computer systems. Arguments of this sort—that is, arguments against using AI systems in particular decision contexts—often point to the same reason: AI systems should not be deployed in such situations because AI systems are not moral agents. However, it's not always clear what we mean by the term "moral agent." This project explores how people think about two types of moral agents: sources of moral action and morally responsible agents.

We will present participants with cases of potential moral action, manipulating the causal entity. For instance, consider the case: "X knocks over a human"—where X is either the wind, an animal, a child, a healthy adult, a robotic vacuum, or a more complex AI system. We will then ask participants four questions. First, we will ask whether an action was performed at all (e.g., "This is something that happened" vs. "This is something X chose to do"). Second, we will use wrongness judgments as a proxy for moral action (e.g., "Did X do something morally wrong?"). Third, we will use blameworthiness as a proxy for moral responsibility (e.g., "Is X blameworthy for knocking over the human?"). Fourth, we will ask about which capacities the entity has—to explore what mediates the relationships. The results will help us understand when these notions of moral agency might come apart, with implications for how we think about artificial moral agency.

Anne-Marie Nussberger **Characterizing Human Preferences for Interpretability in Artificial Intelligence**

11:00 – 11:20 Artificial Intelligence (AI) systems are proliferating across many aspects of society including healthcare, justice, finance, and infrastructure. Many of the most powerful AI systems are difficult to interpret even the engineers of these systems can't explain exactly how they make decisions. When do people care whether AI is interpretable? Here we show that demand for interpretable AI is strongest for AI decisions involving high stakes and scarce resources. However, these same factors cause people to sacrifice interpretability for accuracy when interpretability comes at the

expense of accuracy. In the long run, these preferences could drive a proliferation of AI systems making high-impact ethical decisions that are difficult to explain and understand.

Markus Kneer **Responsibility Gaps and Retributive Dispositions: Data from the US, Germany & Japan**

11:20 – 11:40 Danaher (2016) has argued that increasing robotization can lead to retribution gaps: Situation in which the normative fact that nobody can be justly held responsible for a harmful outcome stands in conflict with our retributivist moral dispositions. In this paper, we report a cross-cultural empirical study based on Sparrow's (2007) well-known example of an autonomous weapon system committing a war crime, which was conducted with participants from the US, Japan and Germany.

We find that (i) people manifest a considerable willingness to hold autonomous systems morally responsible, (ii) partially exculpate human agents when interacting with such systems, and that more generally (iii) the possibility of normative responsibility gaps is indeed at odds with people's pronounced retributivist inclinations. We discuss what these results mean for potential implications of the retribution gap and other positions in the responsibility gap literature.

Michael Laakasuo **Moral Judgments of "Pre-Crime" Arrests and AI Profiling**

11:40 – 12:00 Danaher (2016) has argued that increasing robotization can lead to retribution gaps: Situation in which the normative fact that nobody can be justly held responsible for a harmful outcome stands in conflict with our retributivist moral dispositions. In this paper, we report a cross-cultural empirical study based on Sparrow's (2007) well-known example of an autonomous weapon system committing

a war crime, which was conducted with participants from the US, Japan and Germany.

We find that (i) people manifest a considerable willingness to hold autonomous systems morally responsible, (ii) partially exculpate human agents when interacting with such systems, and that more generally (iii) the possibility of normative responsibility gaps is indeed at odds with people's pronounced retributivist inclinations. We discuss what these results mean for potential implications of the retribution gap and other positions in the responsibility gap literature.

Yuxin Liu **The Moral Psychology behind Artificial Moral Advisors**

12:00 – 12:20

As human moral judgement and decision-making are known to be susceptible to a host of psychological drawbacks, philosophers and machine ethicists have hypothesised various forms of moral decision aid through technological means known as Artificial Moral Advisors (AMAs). Most prominently, Giubilini and Savulescu (2018) propose a quasi-ideal observer AMA that could provide personalised advice on the morally best thing to do based on individuals' pre-declared moral principles, which could function as a moral enhancement tool to help people make better moral decisions. Whilst AMA-like proposals seem intuitively desirable, I will point out several challenges from the perspective of moral psychology that have been largely neglected by proponents of moral machines. In particular, the internal configuration of AMAs is fundamentally misaligned with human moral psychology: it not only incorrectly assumes a static moral values framework underpinning the attunement of AMAs, but there is simply no point in time when people are completely bias-free to input their 'true' moral values into an AMA as a reference point. Additionally, people's reactions and subsequent (in)actions in response to AMA suggestions will likely diverge substantially from expectations given the prescriptive nature of AMAs as moral advisors, humans have an active role in deciding what to do with an AMA's output. These decisions are inherently the same kind of human moral judgements subject to the same heuristics/biases that AMAs are designed to mitigate in the first place. In conclusion, we note the necessity for a coherent understanding of moral psychology in future research on machine ethics.

John Danaher **Will large language models spark a moral revolution?**

14:00 – 14:30 The idea that technologies can change, possibly even revolutionise, moral beliefs and practices is an old one. But how, exactly, does this happen? This talk builds on an emerging field of inquiry by developing a synoptic taxonomy of the mechanisms of techno-moral change. It argues that technology affects moral beliefs and practices in three main domains: *decisional* (how we make morally loaded decisions), *relational* (how we relate to others) and *perceptual* (how we perceive situations). It argues that across these three domains there are six primary mechanisms of techno-moral change: (i) changing options; (ii) changing decision-making costs; (iii) enabling new relationships; (iv) changing the burdens and expectations within relationships; (v) changing the balance of power in relationships; and (vi) changing data, mental models and metaphors. If changes across these six domains are sufficiently widespread, rapid and longlasting, they could prompt a 'moral revolution'. Using the specific case study of large language models, particularly the various iterations of GPT, the talk considers how this technology might transform, and potentially, revolutionise our social morality in the near future.

Daniel Shanks **Accent Prejudice Toward Apple's Siri**

14:30 – 14:40 People show bias toward other people based on their accent, but do people show similar biases when accents are generated by a smart home assistant such as Apple's Siri? In Study 1, 100 US online participants rated Siri's US, UK, Indian, Australian, Irish, and South African voices reading scientific passages. Compared to US voices, participants rated Irish and South African as less likable and were less willing to interact with them. In Study 2, US student participants interacted with Siri on a series of different lab tasks (conversation, origami, smart home routine) with Siri set to one of the three voices (US, Irish, South African). We found no evidence in any behaviour or perceptual measure of prejudice, contradicting the findings of Study 1. In Studies 3-5, we plan to have Siri read a personal story in one of several Siri's different accents and measure participants'

prejudice. Study 3 will test several different accents. Study 4 will manipulate whether participants are informed about the accent's origin country. Study 5 will manipulate whether participants are informed about the voice being from Siri. Collectively, these will determine if bias towards Siri's accents is robust finding, and whether it depends on the content spoken, identification of the country of the accent, or identification of Siri as the speaker.

Daryl Cameron **Empathic Choices for Humans and Robots**

14:40 – 14:50

With an ever-increasing opportunity for people to interact with artificial agents, many studies have examined people's empathetic responses to robots. Yet few studies have examined this from a motivational lens, to understand how and why people might choose empathy in different ways depending on human or robot targets. Are robot minds more difficult to empathize with than human minds, or perhaps easier? I will present four studies using the empathy selection task (Cameron et al., 2019), a free choice measure of empathy regulation, to assess this question. In three studies, we examined whether people would choose to have empathy or remain detached from human targets, and then separately, the same for robot targets. In two of these studies we found that people preferred to empathize with robot targets more than when given a similar choice for human targets. Analyses of subjective cognitive costs of choice options suggested that in the studies with a choice difference by target, people differentiated cognitive costs less for robots (i.e., found both empathy and description to be more similar in cognitive challenge). Furthermore, in a fourth study, we altered the choice set so that people had to choose to empathize either with a human or a robot. When the choice set was altered in this way, people preferred to empathize with humans instead of robots, suggesting that the choice architecture may matter. I will discuss future directions, as well as practical and normative implications for understanding empathy in human-robot interactions.

Janet Pauketat — **Disentangling the impacts of autonomy and sentience on the moral consideration and perceived threat of AIs**

14:50 – 15:00

Artificial intelligences (AIs) are increasingly involved in social life from serving as personal assistants to chatbots to space explorers. The capacities of these AIs, such as their autonomy and sentience, vary greatly but moral psychology to date has referred to these capacities generally as agency and experience. Clarity of conceptualization, measurement and modelling has not kept pace with AI developments. We disentangle the effects of autonomy and sentience–two important capacities of AI minds corresponding to agency and experience–on mind perception, moral consideration, and perceived threat with three preregistered experiments. In Study 1 (N = 254), AI autonomy information increased perceived mind, perceived sentience, moral consideration, and perceived threat. In Study 2 (N = 256), AI sentience information increased perceived mind, perceived autonomy, and perceived threat. Study 3 (anticipated N = 715 to be collected January 2023) will examine the interactive effects of AI autonomy and sentience information. Results from Studies 1 and 2 suggest that AI autonomy and sentience have similar but differentiated effects. With these results, we will make several novel contributions: AI sentience is evaluated more sceptically than AI autonomy; sentience information activates a more general sense of mind than autonomy information; and sentience information has a stronger effect on perceived autonomy than autonomy information has on perceived sentience. These studies will serve to disentangle conceptions of how AI autonomy and sentience affect responses to AIs, build the empirical data on perceptions of these capacities, and lay a foundation for more rigorous research into the moral psychology of AI.

Walter Sinnott-Armstrong — **How to Build Human Morality into AI**

15:30 – 16:00

In contrast with both other top-down and bottom-up methods, we propose a multi-stage hybrid method that starts with a survey about which general features are seen as morally relevant, then constructs particular conflicts among these features, asks participants what ought to be done in those conflicts, and uses machine learning to

predict what they would say about a separate test set of moral conflicts. We illustrate this method with our lab's results regarding the allocation under scarcity of kidneys for transplant, though the same method can also be used for many other moral issues. The results can potentially provide a check on human moral judgments, thereby reducing common errors. It can also provide insight into the computations behind human moral judgments and measure differences among the moral judgments of individuals and groups.

Yochanan Bigman **Algorithm discrimination and stereotype strengthening**

15:50 – 16:10

When people witness prejudice, they often wonder whether it is justified. Are the stereotypes that drive the prejudice true? If stereotypes seem true, people are willing to endorse—and perpetuate—prejudicial behaviour. It is therefore essential to understand how people evaluate the truth of stereotypes and the process of stereotype evaluation. Understanding stereotype evaluation not only helps to reduce prejudice in society but also helps reveal basic cognitive processes. We systematically explore stereotype evaluation and show a process that intersects with the rise of artificial intelligence to fuel more prejudice.

Central to our argument is the idea that people—when making complicated judgments—switch out statistical considerations for psychological ones. In the case of stereotype evaluation, it is difficult to evaluate the statistical truth of stereotypes, and so people instead rely on whether the person endorsing the stereotype (and conducting the prejudice) seem motivated to be biased.

The role of relying on psychological cues rather than statistical ones in stereotype evaluation is important as more decisions are made by artificial intelligence, which—our data suggest—are not ascribed motivation. Therefore, if prejudice is perpetrated by an algorithm, such as with hiring decisions made by an Amazon algorithm in 2014, people may be more likely to believe that that prejudice reflects the truth of stereotypes (i.e., that women are worse engineers than men).

Three studies (N=1020), examining the math-gender stereotype and credit discrimination against immigrants, find that algorithm discrimination

strengthens stereotypes more than human discrimination, an effect mediated by perceived prejudice.

Madeline Reinecke

The puzzle of evaluating moral behaviour

16:10 – 16:20

In developing artificial intelligence (AI), scientists often benchmark against human performance as a measure of progress. Is this kind of comparison possible for moral cognition? Given that human moral judgment often hinges on intangible properties like 'intention' which may have no natural analog in artificial agents, it may prove difficult to design a 'like-for-like' comparison between the moral behaviour of artificial and human agents. What would a measure of moral behaviour for both humans and AI look like? We reveal the complexity of this puzzle by providing an example within reinforcement learning, and we discuss how this puzzle remains open for further investigation within cognitive science.

Simon Myers

The Necessities and Luxuries of Trustworthy AI: The Perceived Importance of Different AI Characteristics

16:20 – 16:30

Artificial Intelligence (AI) is increasingly used to perform tasks with a moral dimension, but to reap the benefits of AI, stakeholders need to be willing to use, adopt, and rely on these systems: they must trust in the AI agents. In this study we draw on established ethical frameworks positing features that should, normatively, be important for trust in machines and examine, descriptively, what ordinary people actually find important. To do this we leverage a new task - the AI Budget Allocation Task - that allows us to distinguish luxuries from necessities for trustworthy AI. What is seen as merely desirable for trustworthy AI, and what is seen as essential? Across three domains N = 496 (Study 1A: Healthcare, N = 168; Study 1B: Judicial System, N = 160, Study 1C: Military N = 168) we look at which characteristics lay people consider to be the most important, and whether this differs based on whether people are judging a machine used for more serious, morally relevant tasks (e.g., determining life support) or less morally salient tasks (e.g., optimising staff schedules). We find that the AI BAT provides more information than importance ratings and that preferences for different characteristics differ in both context and in moral versus

anon-moral tasks. In addition, while ethical theorists and programmers have emphasised the role of interpretability in trustworthy AI, and even while participants rate interpretability as being important, when faced with trading off features in the AI-BAT this was not the case. Instead, we find that interpretability was rated as consistently less important than other characteristics.

POSTERS

Clara Pretus

Effects of a value detection assistant for social media on sharing content online

Moral-emotional language has been found to drive content sharing on social media, fostering polarization. In this cross-cultural study, we will evaluate whether using an AI assistant that detects moral values embedded in online messages changes how social media users interact with this content. For that, N = 2400 participants resident in Italy and the U.S. will be recruited and randomly assigned to one of three experimental conditions where they will be asked to rate how likely they would be to share a series of social media posts. All participants will be exposed to the same posts, but for each group, half of the posts will be tagged with a warning about either a) the presence of moral transgressions (reactive ethics group), b) the presence of positive moral values (proactive ethics group), c) mentions of people/animals/objects (control group). We will evaluate whether the AI assistant's activity detecting moral values affects participants' disposition to share social media posts as a function of whether it adopts a reactive versus a proactive ethics stance (AI for social good) compared to the control group. Preliminary results suggest that people are responsive to warnings about the moral content of the online messages they are exposed to. This study will shed light on whether and how AI assistants that can detect moral values in online settings affect the behaviour of social media users.

Giovanni Bruno

Framing self-sacrifice in the moral dilemma of autonomous vehicles

In the investigation of moral judgments towards Autonomous Vehicles (AVs) behaviour, the traditional paradigm of sacrificial dilemma has represented a flexible and widespread experimental tool. Facing the typical AV dilemma from the passenger's perspective, the endorsement of the utilitarian resolution corresponds with the acceptance of a self-sacrificial act (steer to the side of the road, protecting n pedestrians but sacrificing the AV passenger).

Unexpectedly, this alternative contradicts the traditional sacrificial dilemma's structure, in which the endorsement of the utilitarian behaviour matches the quest for self-protection. Considering this nontrivial difference, the present study (n = 183) aims at deepening the role of self-sacrificial framing on moral judgment and on the perceived intensity of four moral emotions (shame, guilt, anger, and disgust), focusing on the context of autonomous- and human-driving sacrificial dilemmas. As expected, a higher endorsement of the utilitarian behaviour was detected when the utilitarian manoeuvre converged with the self-protective act. Interestingly, higher scores of guilt and shame (as self-referred moral emotions) were observed after the endorsement of the utilitarian but self-protective option, as well as after the administration of human-driving dilemmas. The present study collects novel information on the role of self-sacrifice framing in the development of moral dilemmas. The moral request to endorse a self-sacrificial act for the sake of the collective may have consequences on the relative support of the utilitarian moral code, also affecting the resulting subjective emotional experience.

Giovanni Masala **Older adults' perspective of social robots**

Artificial intelligence and robotic solutions are seeing rapid development for use across multiple occupations and sectors, including health and social care. In eldercare, AI technology is probed with success in entertainment activities with seniors. However, researchers are faced with the conundrum of exploring the degree of trust that seniors place in socially assistive robots and the acceptance and usability of such robots in critical care provisions like medication administration. Moreover, as robots grow more prominent in our work and home environments, whether older adults would favour them in receiving useful advice becomes a pressing question. In present-day research, little is known about people's advice-taking behaviour and trust in the advice of robots. This talk will introduce experiments of robot-human interaction to evaluate the potentiality of social robot companions in eldercare, highlighting the end-user's point of view. The focus of the discussion will be the older adults' trust in social robot in sensitive tasks..

Henry Ashton

Does intent matter? An experiment to contrast judgements of mens rea in autonomous AI agents against humans

The degree to which we find someone legally culpable for causing a harm is typically dependent on the mental state under which they acted (or failed to act). In criminal law this is referred to as mens rea. If an AI powered agent (an embodied robot or otherwise) causes some harm, does its `mental state' matter for blame attributions according to lay people? How do blame levels compare with the same harm caused by a human with the same mental state? Can we decompose the elements of mens rea into common questions concerning desire and foreseeability? This presentation describes the results of an experiment where we seek to answer these questions.

Javier Gomez-Lavin

The Role of AI Gender on Task Allocation

To date, no significant empirical interventions have been carried out on the ethical concerns raised by the ubiquity of gendered, particularly feminized, AIs and the role they may have in reinforcing sexist divisions of labour. Our series of studies has begun to shed some light on this important topic by building off of related work in the domain of human-robot interaction (Kuchenbrandt et al. 2014).

420 native-English speaking adults (49.8% self-identified as women) were recruited via Prolific and randomly assigned to one of four conditions in our 2x2 between-subjects design. Each condition asked participants to imagine that they had been paired with either a feminized- or masculinized-AI (Nera or Nero) to help them with a feminized- or masculinized-activity (caring for a sick relative or planning a vacation) with four component tasks (e.g., managing finances, scheduling appointments). Activities and tasks were normalized on a number of dimensions including gender-typicality and complexity in a prior study. Participants were asked to assign between one and three tasks to the AI. They were also asked to rate the AI on a number of dimensions, including competency and

personality metrics. We hypothesized that there would be a series of at least nine interactions between AI-, participant-, and task-gender. Initial results support this complex picture. We found, for instance, that participants are more likely to assign tasks to feminized- versus masculinized-AIs ($t(416) = 2.82$, adjusted $p < .05$), and this effect is pronounced for male-participants assigned to feminized-activities ($t(415) = 2.82$, adjusted $p < .05$). Furthermore, participants less familiar with AIs are more likely to assign feminized tasks to a feminized- AI ($t(411) = 3.24$, adjusted $p < .05$). Additional findings and further steps are also discussed.

Junior Okoroafor **The Effect of Violated Expectancies in explaining Trust in AI**

TBC

Katerina Manoli **The Effects of Moral Spillover on the Social and Moral Inclusion of Artificial Intelligences**

Future artificial intelligences (AIs) are at risk of serious harm from a lack of moral consideration. Moral spillover is the transfer of moral consideration from one setting into another setting (e.g., from one being to a group of beings, or from present beings to future beings). Moral spillover has been observed in the transfer of anti-slavery activism to animal rights activism, in the expansion of a pro-environmental behaviour into other related behaviours (e.g., from conserving electricity to conserving water), and in the transfer of moral concern from humans and nonhuman animals to some artificial agents (e.g., from a pet dog to a robotic dog). Moral spillover matters because of its implications for the expansion of the human moral circle, a promising strategy for reducing the suffering of many kinds of sentient beings. How, when, and why moral spillover shapes the social and moral inclusion of AIs remains an open question. Here, we provide an overview of moral spillover research with a focus on studies that show spillover effects in the context of AIs. We suggest

that moral spillover may be important to fostering the moral consideration of future sentient AIs and consider the implications of inducing moral spillover in this domain. Finally, we call for more research to identify the boundaries that shape when moral consideration might spill over to affect the diverse range of AIs who exist now and who may be widespread in future societies.

Leda Berio

Emotional agents or emotional situations? Emotion attribution, mind shaping, and normativity in emotional interaction with artificial agents

I argue that considering interactions with artificial agents in terms of emotionally loaded scripts can be beneficial in three ways: it bypasses the problem of emotional attribution to artificial agents (i), it brings attention to the normative implications of these interactions (ii) and, finally, helps us identify the mechanisms of mindshaping at play in emotional interactions and their moral implications (iii). Empirical studies seem to suggest we do attribute emotion to artificial agents (e.g., Lakatos et al, 2014) despite knowing they do not experience emotions in a human sense. To solve this puzzle, philosophical accounts have focused on what kind of emotion-like states we actually attribute (Nyholm, forthcoming) or on what strategies we employ (e.g., treating artificial agents as fictional characters, Schmetkamp, 2021). I propose to focus on the emotional character of the situations at stake rather than on the emotional capabilities of the individuals involved; in particular, I argue that we should consider social interactions as activating scripts and schemata (Bicchieri and McNally, 2018) that come with expectations on how agents should behave and how agents should feel. These scripts are thus normative and have a "mind shaping" (Mameli, 2001; Fenici and Zawidzki, 2020) function: they contain information about what is morally right to feel and what is morally right to do and are activated to enforce these feelings and behaviour. In this sense, I suggest, our behaviours and emotions when interacting with others are morally regulated, independently of who (or what) the other agents are.

Luis Marcos Vidal

Effects of a Value Detection Assistant for Decision-making in an Inter-group Setting

People often struggle connecting the values that we hold with our actions. Implicit biases often modulate behaviour generating a misalignment between how we would like to act and how we indeed do act. Artificial Intelligence (AI) is a powerful tool that presents many possibilities in the ethical field, but how humans will react to moral AI is still unknown.

In the current study, we studied if receiving feedback on equality from an AI system affects in-group favouritism as an implicit bias. For that, we have run an experiment of resource distribution in two independent samples, each with a different level of intensity of group identification. The first one is a minimal group sample (random assignment to groups) while the second one is a historic group sample (groups based on political preference).

Each participant had to distribute a resource between other fictitious participants identified only by their group. After the task, they received different input depending on the experimental condition they were randomly assigned to: a) the value of equality was remarked so it becomes salient, b) participants got their score on equality according to the distribution of resources that they performed, c) participants chose a definition for equality before getting the score, and d) participants did not receive any input (control). After the intervention, participants will repeat the resource distribution task.

Preliminary results showed that AI systems feedback produces a reduction in out-group discrimination, especially in minimal group's condition. This suggests that AI systems can help reduce implicit bias and, thus, allow humans to behave in a way that is more aligned with their moral values.

Reem Ayad

The Judgment of A.I. as Intentional Actors

Social transactions are increasingly infused with decision input from A.I. agents. This input often extends to moral decisions. What

influences humans' judgments of A.I. agents' moral responsibility? Past work in the field of human-robot interaction has focused on manipulating the agent's physical features. Instead, we used a disembodied A.I. agent and manipulated its psychological features. In Study 1 (N=4000), participants listened to an audio recording of an A.I. agent expressing high or low levels of Values, Autonomy, Self-Aware Emotions, and Social Connection. Participants then read a scenario in which the same agent committed a moral transgression with randomly varied degrees of intentionality. Participants judged the moral appropriateness of the agent's actions. Results revealed that the A.I. that expressed high (vs. low) moral values and high (vs. low) social connection was generally judged less harshly. The A.I. that displayed high (vs. low) autonomy and high (vs. low) self-aware emotions was generally judged more harshly. Finally, a highly socially connected A.I. committing an act while focusing on the end was judged more harshly than one focusing on the means. Study 2 (N=2000) replicated the results and tested for a mediator: the extent to which the agent is perceived to have a distinct mind. Results showed that high social connection correlated with lower perception of a distinct mind. Perception of a distinct mind partially mediated the relationship between social connection and moral judgment.

Tina Seabrooke

TAME Pain: Trustworthy AssessMEnt of Pain - Listening Between the Lines

Perception of pain is an extensively studied area but remains poorly understood. Reliable assessment of pain for patients can be extremely challenging, especially when the cohort is unable to communicate directly with words such as in stroke patients, patients with learning difficulties or autism. In these situations, healthcare workers must gauge pain perception using their own experience, which can be variable. In conjunction with the UKRI Trustworthy Autonomous Systems Hub (TAS Hub) and the Good Systems team at UT Austin, we plan to develop a novel, trustworthy autonomous system that harnesses acoustic biomarkers of pain, which can then be used to guide healthcare professionals to tailor and optimise analgesia for patients. The proposed project will adopt established pain-induction techniques and identify acoustic biomarkers of pain in healthy subjects. These markers will then be used to develop a machine

learning model to detect pain levels. We will then test to see whether the model can be used as an additional tool to help doctors when making decisions about pain management. With the introduction of automatic pain detection, new moral and ethical questions in relation to the treatment of patients arise. Is it moral to automate the detection of a complex human emotion to an artificial system? Are we risking underestimating the pain the patient is experiencing, potentially leading to the infliction of additional pain? We aim to explore these issues over the course of the project.

Trisevgeni Papakonstantinou

Developing a framework of blame attribution in human-artificial agent systems using Twitter

This study proposes and tests a method that combines paradigms from machine learning, causal inference and cognitive science to study blame attributions, publicly stated on online platforms. It specifically focuses on the blame attributions of Twitter users, reacting to different AI Incidents. Building on prior cognitive research on how people attribute responsibility to human agents in social systems, this study examines how these judgments might differ when applied to AI. The substantive aim of this study is to identify the agents people are attributing blame to, and the factors associated with people's attributions. A second methodological aim is the development of a computational method able to predict blame, agents and factors from free-text data. Three coders independently coded a sample of tweets initially according to whether they involve a responsibility attribution. This dataset was then coded in terms of agents, factors and attitude. Preliminary results from a small subsample of data showed the algorithm was blamed most frequently, in 141 of 266 instances of blame. In cases of blame across all corpora, algorithms were significantly ($p<.05$) correlated with the factors of bias, obligation and virtue. Government was correlated with misconception. The implementation system was significantly associated with implementation and negative outcome. Finally, companies were associated with bias, foreseeability, obligation and virtue. A cluster analysis accurately classified the tweets in terms of blame and further grouped together agents, emotions and factors. This study is ongoing, and we expect to present results on a sample of 1000+ tweets involving 12 human-artificial agent systems.

Zhaoning Li

Towards human-compatible autonomous car: A study of non-verbal Turing test in automated driving with affective transition modelling

Autonomous cars (ACs) are indispensable when humans go further down the hands-free route. Although existing literature highlights that the acceptance of the AC will increase if it drives in a human-like manner, sparse research offers the naturalistic experience from a passenger's seat perspective to examine the human likeness of current ACs. Here, we tested whether the AI driver could create a human-like ride experience for passengers based on 69 participants' feedback in a real-road scenario. We designed a ride experience-based version of the Turing test for automated driving. Participants rode in the AC (driven by either human or AI drivers) as a passenger and judged whether the driver was human or AI. The AI driver failed to pass our test because passengers detected the AI driver above chance. The failure of the AI driver inspired us to investigate how human passengers ascribe humanness in our test. To this end, based on Lewin's field theory, we advanced a computational model combining signal detection theory with pre-trained language models (PLMs) to predict passengers' humanness rating behaviour. We used affective transition between pre-study baseline emotions and corresponding post-stage emotions, transformed by PLM, as the signal strength. Results showed that the passengers' ascription of humanness would increase with the greater affective transition. Our study suggested that affective transition, serving as a hypothetical essential part of passengers' subjective ride experience in our model, may play a crucial role in their ascription of humanness.

Cliodhna O'Connor

Public responses to use of AI to diagnose mental illness

The understanding, experience and management of mental health difficulties have long been premised on clinician-decided diagnoses, determined by matching observed and self-reported symptoms to disorder profiles in diagnostic manuals. Experts believe that increasing use of online platforms, wearable devices and sensor technology may drastically alter the assessment and classification of mental illness. Through AI techniques such as deep neural network

machine learning, digital data can index subtypes of disorder, which have demonstrated diagnostic value in domains such as PTSD, depression and psychosis. Yet AI diagnosis has many ethical challenges, and replacement of familiar diagnostic categories with AI-enabled precision diagnoses may have unanticipated consequences for how people understand mental illness and view those who experience it. This study provides first insights into societal and ethical responses to AI diagnosis of mental illness, using representative samples of US and UK participants (N=1000). An online experiment exposes participants to a vignette describing a fictional character with mental health difficulties; half of participants read that the person's difficulties are clinically assessed using an AI diagnostic tool, and half a standard manual-based (DSM) diagnostic process. Participants then complete a battery of scales measuring causal attributions for the target's symptoms, perceived responsibility for their problems, and desired social distance from the target. Participants also rate their level of concern about a range of ethical issues associated with AI diagnosis (e.g., privacy/security, interpretability/communicability, accuracy/bias, stigma/discrimination). Results will provide timely insight into the social and ethical implications of new diagnostic technologies, as or before they become implemented in clinical practice.

Melanie McGrath **If you can't beat them, join them: A new approach to collaboration between humans and artificial intelligence**

Concerns exist that the distinction between humans and machines is increasingly blurred due to the rapidity of technological development. A new approach to artificial intelligence from Australia's national science agency (CSIRO) instead reinforces and endorses the distinction between humans and machine intelligence. The focus of collaborative intelligence (CINTEL) is on complementarity, asserting that human and machine capabilities are not interchangeable, but uniquely valuable in certain applications. In tasks involving regular exposure to novelty and uncertainty requiring flexibility and interdependency, collaboration between adaptive, creative humans and powerful, precise AI promises new solutions and efficiencies. Such collaboration also represents a new form of interaction between humans and technology. We have developed the

recurring phase framework of trust in CINTEL to meet the unique requirements of this interaction. Trust of human users is central to reliance on technology and is anticipated to be especially critical in facilitating the sustained teamwork that characterises collaboration. Drawing on both the psychological and computer science literature, the recurring phase framework of trust in CINTEL identifies antecedents and outcomes of appropriately calibrated trust in collaborative systems. The framework's incorporation of teamwork processes and recurring performance phases also captures the dynamism inherent to trust in teaming contexts. This talk will introduce CINTEL and its implications for human-machine relationships and present the recurring phase trust framework.

Anton Kunnari **Judging Medical Decisions Made by AI**

In two series of vignette studies, we examined moral responses to human-made versus AI-made decisions in morally difficult medical settings. The first series focused on forced medication, and the second on euthanasia decisions. In the first series, we examined how people judge a robot versus a human nurse administering forced medication or disobeying the order to do so. We were interested in exploring the contrast between patient autonomy and patient well-being in medical ethics, and how automation affects judgments in this context. In 4 studies, we found decisions to forcefully medicate a patient were approved more when the agent was a human rather than a robot, but that decisions to disobey the forceful medication order were generally more approved. In the second series, we examined how people judge passive euthanasia decisions involving varying levels of automation. In 4 studies, we found that any level of automation consistently reduced moral approval towards the process and the decision-makers involved. The results suggest that robots and AIs, more so than humans, are expected to value human autonomy and well-being. Implications are discussed.

Jukka Sundvall **Lay Views on Moral Patiency in Robots**

To what kind of robot would non-philosophers grant rights? What would a robot have to be like for laypeople to condemn those who

treat the robot “inhumanely”? This presentation covers three preliminary survey studies on views about robot moral patiency. In each study, participants were presented with descriptions of ”immoral” actions towards robots (kicking a robot, destroying a robot’s CPU, locking a robot in a room), and a list of potential properties a robot could have. In Study 1, participants were asked to indicate which properties were necessary and/or sufficient grounds to state that the described action was a moral violation against the robot. In Studies 2 and 3, the same properties were used, but here, participants were instead asked to rank them in order of strongest to weakest arguments in favour of moral patiency. Despite differences in methodology, the properties that emerged as the most and least important were consistent across studies. For instance, capacities for experiencing pain or negative emotions were commonly listed as necessary and/or sufficient, and ranked highly in importance, whereas social aspects such as membership in a community or ability to communicate with humans had the opposite trend. Whether a robot was anthropomorphic or not did not seem to matter in terms of the ranking of important properties. Implications and future directions are discussed.