CEPE/IACAP Joint Conference 2021 - Abstract Submission and Registration

Report of Contributions

Contribution ID: **8** Type: **Individual Paper**

Should we Conceptualize Privacy as Value?

In this talk, I will argue for three things. First, I will argue that we must solve the question of whether privacy should be conceptualized as value –or value neutrally –before we proceed to analyze or define the concept of privacy. Second, I will argue that we should define privacy as a value. Third, I will argue that privacy should be conceptualized as a *pro tanto* good, meaning that having less privacy is always worse than having more (although other values may be more important). Moreover, contrary to the standard ways of analyzing privacy, privacy should not be defined as a state or condition, but as a something we can have less and more of; preferably, we should strive forward an analysis of privacy that allows us to measure the value.

The main argument, in the second part, will proceed by looking at some examples that may be used to defend the notion that we need a value-based conception of privacy, beyond the right to privacy. I will look at examples that are specific in the modern information society, such as challenges due to algorithmic or big data technology.

Keywords

privacy; pro tanto good; the right to privacy; conceptual analysis

Primary author: Dr LUNDGREN, Björn (Umeå University; Institute for Futures Studies; Stockholm University)

Session Classification: Privacy I

Contribution ID: **9** Type: **Individual Paper**

Informed Consent and Algorithmic Discrimination - Is giving away your data the new vulnerable?

With the development and implementation of algorithm driven decision-making procedures based on aggregated data collected from individual users, several moral challenges arise. In this paper, we explore the connection between voluntarily sharing or selling your data on the one hand, and the dangers of automated decision-making based on big data and artificial intelligence on the other. We call into question the voluntariness of this transaction especially for certain vulnerable groups and argue that the concept of voluntariness in some cases will reinforce historic discriminations and structural injustices.

While so-called statistical discrimination is not morally wrong per se, such methods are unjust if they disproportionately disadvantage members of certain social groups. We will show that vulnerable groups as already being disadvantaged are in certain respects more likely to "voluntarily" give away their data than more privileged groups and are thereby even more prone to additional forms of discrimination that also involve algorithmic decision mechanisms (ADM).

Keywords

Discrimination; Algorithm; Informed Consent; Privacy; Data Literacy

Primary authors: Dr BEHRENDT, Hauke (University of Stuttgart); Dr LOH, Wulf (University of Tübingen)

Presenters: Dr BEHRENDT, Hauke (University of Stuttgart); Dr LOH, Wulf (University of Tübingen)

Session Classification: Algorithmic Discrimination

Contribution ID: **10** Type: **Individual Paper**

No Algorithmization without Representation: Sandbox for Regulatory Experiments

We gathered feedback on intrusive algorithmic services and pervasive surveillance from a small sample of 59 participants before and during the first and second wave of the COVID-19 pandemic in Czech Republic (January, June 2020 and April 2021). In the pre-pandemic January workshops, participants provided feedback on functional prototype of a blockchain service that uses satellite data, while in June 2020 and April 2021 it was the Bluetooth-enabled contact tracing application. The comparison of the initial feedback about hypothetical blockchain application with the real contact tracing app during a crisis reveals no radical change in attitudes towards intrusive services. Almost 2/3 refused the "polite"surveillance in 2020 and expressed distrust in government oversight of emerging algorithmic services. Participants perceive blockchain or contact tracing applications as extraterritorial and lawless zones rather than infrastructures that allow government regulation. In the article, we will compare and discuss the regulatory expectations of both groups that accepted and rejected the services. They seem to agree on the support of experimental policy interventions and regulations based on independent rather than government oversight. Instead of reducing regulations to code or insisting on strong regulations over the code, participants were open to novel and exploratory interventions that we describe as hybrids of code and regulation or automation and oversight. The expectations of hybrid and experimental governance remind us of the famous slogan of the Independence war in the US demanding "no algorithmization without representation". Participants perceive the intrusive services as new algorithmic "territories"where "data" settlers have to redefine their sovereignty and agency on new grounds rather than rely upon the existing institutions.

Keywords

regulatory sandboxes, RegTech, governance by design

Primary author: Dr RESHEF KERA, Denisa (BISITE, University of Salamanca, Tel Aviv University)

Co-author: Dr KALVAS, František (University of West Bohemia)

Session Classification: Novel Practices in Design and Regulation of AI

Contribution ID: **11** Type: **Individual Paper**

Practical and Moral Challenges for the Project of Formal Ethics

This paper advocates caution against a formalization of ethics by attempting to show that it may pre-sent an obstacle to inclusive moral deliberation, a notion that borrows from Jürgen Habermas"ideal speech situation'. Formal ethics presents an obstacle to the attainment of this ideal, because it is likely to perpetuate unjustified power imbalances. Simply put, a formalism may disadvantage those without a proper command of it, which is particularly significant if it concerns moral matters. Consequently, the paper challenges the promises of proponents of formal ethics as way of increasing rigor in ethical evaluations and consequently facilitating moral progress. The paper also highlights dangers in em-ploying formal ethics for devising ethically aligned machines. For this purpose, practical limitations of formal ethics are first considered that stem from the challenges from formal languages, from a possi-ble ill-structuredness of moral problems, from biases incurred by simplifying assumptions and from moral disagreement. Such practical challenges contribute to further moral challenges for formal ethics. In light of the ideal of inclusive moral deliberation, power asymmetries represent such a moral chal-lenge–one that is rendered particularly significant, since a formalisms'limitations may also perpetuate limited moral conceptions. The paper presents a plea for taking seriously the mismatch between an ideal moral consensus and one that can be implemented given current frameworks. Transparently considering this mismatch may be a way forward without undue power asymmetries and in which responsible implementation of automat-ed ethical evaluations is facilitated. An outlook provides a brief description of using formalisms for ethics aimed at circumventing the dangers outlined in this paper.

Keywords

Artificial Intelligence; Machine Learning; Ethical and Societal Implications; Formal Ethics; Machine Ethics; Moral Deliberation; Inclusion

Primary author: Dr HERZOG, Christian (University of Lübeck, Ethical Innovation Hub)

Session Classification: Ethics of AI Ethics

Contribution ID: **13** Type: **Individual Paper**

Why AI is no threat to democracy. But to the rule of law.

I show that recent developments in AI technology (especially in Machine Learning in combination with Big Data) and its role in Surveillance Capitalism are not a direct, but only an indirect threat to democracy. Based on Lawrence Lessig's "Code is Law,"I draw a more elaborated picture of regulation and argue that AI is subsisting on an empty shell of democracy while threatening the rule of law. AI (or the companies who control it) even promotes and monetizes democracy while our societies have been shifting towards a rule of code. From this background, I argue that the attempt of legal regulation of current AI is a hopeless illusion (given the interdependence of Western governments and Big Tech companies) and that a free and open society urgently needs to regain control over their source code and the critical digital infrastructure to avoid being regulated, controlled, manipulated and oppressed by AI.

Keywords

regulation, surveillance capitalism, rule of law, democracy

Primary author: Dr ROSENGRÜN, Sebastian (CODE Berlin)

Presenter: Dr ROSENGRÜN, Sebastian (CODE Berlin)

Session Classification: AI and Regulation

Contribution ID: **14** Type: **Individual Paper**

Can high-stake decisions be delegated to a machine? (A rejoinder to Scott Robbins)

Worries abound out the use of (black box) machine learning (ML) algorithms. In EU circles, several regulations demand that high-risk applications of AI are to be sufficiently transparent to allow users to interpret the model output. Applications are classified as such if they pose high risk to the health or safety of natural persons or to their fundamental rights. Recently, Scott Robbins introduced an even sterner word of caution regarding the use of ML to generate 'evaluative outputs' (rules for deciding about aesthetic, ethical, or emotional issues). He argues that machines (read: ML) should not be allowed to generate the rules to decide about them –this task can only be entrusted to human reasoning.

The central question of this research is whether we really have to set restrictions on the type of ML involved for handling these 'critical'applications. Can ML in suitable form be useful in these scenarios, or does it have to be abandoned?

Four stances can be distinguished. The first stance argues that all ML is acceptable as long as its performance is accurate. The second stance pleads for a restriction to explainable ML; that is, outcomes as well the model as a whole can be explained to stakeholders concerned. Post-hoc, a model of the model in use is generated. Thirdly, an even more restrictive stance demands that only interpretable ML is used. Such models can be interpreted by their very design –think of decision lists and score cards. Finally, one may argue that only ruling out ML altogether can do justice to the complexities of 'critical'applications.

I formulate a preference for the third position. On the one hand, I prefer more restrictions than the second stance adopted by the EU, mainly because post-hoc explanations do not easily map onto the concerns of users involved; on the other hand, I prefer less restrictions than Robbins has advocated, mainly because interpretable models do allow to keep track of underlying assumptions that may be in dispute.

Keywords

explainability, high-stake decisions, machine learning

Primary author: DE LAAT, Paul

Presenter: DE LAAT, Paul

Session Classification: Human Machine Relations II

Contribution ID: **15** Type: **Individual Paper**

The Story Telling around the Use of AI: the Ethical Dimensions that are considered in the Governmental Official Discourses

The promptness with which new technologies have been deployed in the last few years raises particular ethical issues. In the case of the Artificial Intelligence (AI), various reports and guidelines have been developed by governmental and non-governmental institutions reflecting on ethical principles, requirements and best practices for when the technology is developed and used. As the European Parliament stated, "within the last five years, AI ethics has shifted from an academic concern to a matter for political as well as public debate"(European Parliament, 2020).

We consider that one of the pillars of an ethical approach to the use of AI relies on a transparent and accountable communication of both the opportunities, uncertainties and the risks induced by the use of AI. With this respect, this paper looks into AI ethics from a risk communication perspective by analysing the discourse of decision-makers and key actors. The purpose of this paper is to offer evidence on how risk and uncertainty around AI is communicated in practice, to offer guidance for researchers in the field, and to inform and give a risk communication perspective for those interested in the ethics of AI. This paper will analyse how the European Commission (EC) communicates about AI when referring to ethical AI. Using a European perspective in this paper offers us the possibility to look into messages that are conveyed to various types of stakeholders: individuals, organisations and society at large. There are multiple ways in which the topic of AI ethics can be addressed. This paper explores how the EU talks about the ethical AI principles in reports that represent the basis of the AI EU strategy, focusing on the way that risks of AI are communicated.

Keywords

artificial intelligence, story-telling, risk communication

Primary authors: RUSU, Anca Georgiana (Dauphine University); Dr MERAD, Myriam (Dauphine University (Paris 9))

Session Classification: Risks of AI

Contribution ID: **16** Type: **Individual Paper**

International Humanitarian Law and the Use of Algorithms for Military Decision-Making

Recently, much effort has gone into deciding what is the appropriate space for algorithmic decision making in domestic law. From discussions about the constitutionality of police officers'use of algorithms to justify probable cause, to discussions about use of recidivism algorithms in parole hearings, and the use of machine learning to aid judges in deciding on relevant precedents –legal scholars have been actively discussing the consequences of a changing world to our domestic norms. International legal scholars need to follow suit.

Conceptually the bedrock of international humanitarian law lies in a distinction between intentional and unintentional harm. After all, principles of distinction, humanity, proportionality, and other rules that underpin laws of armed conflict depend on intention. But what happens when military decision-making starts being augmented by algorithms? Where does responsibility for an action rest when a pilot must rely on algorithms in all stages of the so-called OODA loop (observe, orient, decide and act)? To illustrate, a pilot might rely on a machine learning algorithm that identifies civilians, she might rely on a different algorithm to adjudicate between alternative options for action, and she might even be aided (one day) in executing the action faster (shortening the time between 'decision'and act, by using for example skull caps that 'identify a decision'and perform desired act without the typical human lag-time).

As the effectiveness of actions in war (where effectiveness presumes lowering civilian casualties) due to reliance on more and more sophisticated -often machine learning algorithms- goes up, the space for human decision making, or even our understanding of why someone is assessed as 'civilian'by an algorithm, goes down. Law of armed conflict must keep up and it must do that in two ways. First, there need to be international legal norms that impose limits on the type of algorithms that are used and the way the machine algorithms are trained whenever they are used to augment military decision-making. Second, we need to reassess what it means to act autonomously and what it means to assign responsibility for actions in war. In this paper, I try to start on both of these tasks. Regarding the first task- I argue for principles of transparency, non-bias, requirement to impose meta-algorithms to oversee and provide human-like explanations for actions. I argue that the right balance between effectiveness and transparency should set the limit to the type of algorithms we can use and that beyond that limit no amount of increased effectiveness can justify a lowering in transparency and ability to explain decisions. Regarding the second task, I draw on rich literature on autonomy and moral responsibility to suggest upper and lower limits to a theory of autonomy under which we can operate to assign responsibility for actions in war.

In section I, I quickly survey recent debates regarding problems that arise in domestic settings as the juridical system increases its reliance on algorithms to augment or supplant decision-making. I do this, with two aims in mind, first to stress the undertheorizing regarding algorithmic decision-making in international humanitarian law and second, to draw some basic lessons about the issues and questions that arise that will translate into international legal realm. These include questions about transparency, accuracy, bias, potential for abuse and misuse, fairness, etc.

In section II, I argue for the centrality of the principle of discrimination in international humanitarian law and the role of intent in making sense of key principles of discrimination and proportionality. More importantly, in section II, I examine ways in which algorithms are affecting decision-making in areas of action that are governed by international humanitarian law- specifically ways we fight wars. I focus on a few key examples to illustrate the problems that arise regarding assigning responsibility, and interpreting what it means to act with proportionality and discrimination when those terms require a certain understanding of intent. Traditionally, to act discriminately and proportionately means that one does not intend to target civilians and that one makes sure that the harm to civilians that is unintentional, but foreseeable, is proportionate to the military objective. With increased reliance on algorithmic decision-making the line between

intentional killing and foreseeable killing becomes increasingly blurred. This can be for a number of reasons, but primarily it is because increasingly sophisticated machine learning algorithms have lower transparency and therefore lower explainability. Often, much like in domestic law, explainability is necessary to judge the legality of some action.

In section III, I argue that as the lines between intentional killing and foreseeable, but unintentional killing become increasingly blurred due to increased reliance on algorithmic decision-making, that so augments our decision-making process that it is also hard to asses responsibility for certain actions. So, in addition to suggesting that it is hard to assess whether an action in war is discriminate, it is also hard to establish responsibility for such actions. The U.S. military and others have made great strides in attempts to respond to this worry. For example, the Defense Innovation Board (DIB) has proposed a robust set of principles for using AI and algorithms in weaponry and they have in particular suggested that the military use traditional models of responsibility to identify those responsible in cases when the AI augmented weapons misfire or cause unjustified harms. I argue that DIBs proposal is a great start domestically.

In section IV, I finally turn to what all of this means for international humanitarian laws and jus in bello (justice in war) more broadly. First, I argue that some of our central legal instruments in international humanitarian law need to be expanded so as to capture the changes in the way we fight wars. Specifically, we need robust treaties that further govern AI weaponry in ways that are sensitive to the needs of international humanitarian law (those argued for in sections II and III). Second, I argue for a limit on weaponry that would ban any and all weapons whose actions cannot be explained in a way that allows a judge to weigh legally and morally salient features of an action: intent, context of action, etc. To illustrate, we might have a weapon that has significant success in avoiding civilian casualties, but when it executes an attack that kills some civilians, we need to be able to distinguish between two scenarios: i. machine wrongly identified a civilian as a combatant and therefore the human in the loop (if there is one) executed an action on flawed facts, and ii. machine rightly identified this person as a civilian, but the machine or the person aided by the machine decided rightly or wrongly that the risk of harm to this civilian is proportionate to some military objective. If we are faced with the first scenario, then key questions revolve around responsibility for a mistaken identification of a civilian, if we are faced with the second scenario, then the key questions revolve around intent and who we can assign the intent to, and whether it violates conditions of discrimination.

Keywords

algorithmic decision-making, international humanitarian law, AI weapons

Primary author: DAVIDOVIC, Jovana (University of Iowa)

Session Classification: Ethics and Automated Warfare

Contribution ID: **18** Type: **Individual Paper**

Will we exploit benevolent AI?

In hospitals, courts, banks, cars, algorithms decide over increasing aspects of our lives. Scientists as well as the public are concerned that, by negligent design or in unforeseen circumstances, these artificial intelligence systems may come to treat us unfairly. But how fairly are we inclined to treat them? Evidence from behavioral game theory shows that despite the risks of losing out or of being exploited by others, people often cooperate with one another to attain mutually beneficial results. Here we offer evidence indicating that, when interacting with AI, people show exploitative, less cooperative dispositions. In 9 experiments, human participants played one-shot economic games with either another person or an AI agent emulating typical human behavior. Humans cooperated less with AI agents than with other humans. More importantly, participants predicted that AI agents would be as cooperative as humans, but were ready to exploit their co-player's anticipated benevolence when it was an AI agent more than when it was a human. These results warn that vulnerability to exploitation may be an unexpected challenge when introducing AI into human society.

Keywords

human-AI cooperation, algorithm exploitation, game theory

Primary author: KARPUS, Jurgis (LMU Munich)

Co-authors: KRÜGER, Adrian; TOVAR VERBA, Julia; BAHRAMI, Bahador; DEROY, Ophelia

Presenter: KARPUS, Jurgis (LMU Munich)

Session Classification: Digital Life and Ethics

Contribution ID: **19** Type: **Individual Paper**

Interfaces that make us think. Conceptualising online critical thinking as designed interaction

This paper investigates the possibility of extending online user's critical thinking by re-designing the interfaces that users interact with when surfing the Web. The paper brings a novel contribution to the fields of computer ethics and Internet ethics by using the concept of ecological cognition to describe the necessary conditions for enacting critical thinking online, by reconceptualising critical engagement online as an interaction between humans and interfaces where human skills need to fit technological affordances. The paper uses the conceptual tools from the phenomenology of cognition, computer ethics and human-computer interaction studies to articulate a complex concept of online critical thinking which places the responsibility and agency for critical thinking online both with the user's skills and with the designed interface. By explaining online critical thinking as a dynamic interaction between our skills and the online interface, I bring to the fore certain affordances that need to be designed for purposefully if we are to foster more critical engagement with information online and, ultimately, to empower users to tackle online misinformation on their own.

Keywords

critical thinking, extended cognition, human-computer interaction, designed interaction, value sensitive design, affordances

Primary author: MARIN, Lavinia (TU Delft, Ethics and Philosophy of Technology Section)

Contribution ID: **21** Type: **Individual Paper**

Ethics of Adversarial Policies and Data Poisoning: Collateral Damage Considerations

This paper investigates the ethical implications of using adversarial policies (such that the purpose of the policy is to exclusively attack another policy). We suggest two significantly different understandings of "adversarial policies"in zero-sum and open-ended environments. While zero-sum scenarios find acceptable all adversarial applications, open-ended scenarios can accept only red and white adversarial attacks because open-ended scenarios do not allow "collateral damage"policies which are designed to directly cause damage to an opponent through making smaller sacrifices. Lastly, we find some counterexamples to the last conclusion where the use of black adversarial policies in open-ended scenarios might still be permissible.

Keywords

AI ethics, adversarial policies, data poisoning, privacy

Primary authors: Dr ADOMAITIS, Laurynas (Nord Security and CEA); Mr OAK, Rajvardhan (Microsoft)

Session Classification: Cybersecurity and Adversarial Attacks

Contribution ID: **22** Type: **Individual Paper**

Responsible Deployment of AI by a Smart Society

Already in 2008 the vision of a Smart City was introduced by industry leaders. Ten years later concrete implementations of subsystems of such a city were realized introducing Smart Mobility for the optimal regulation of traffic, Smart Energy for efficient energy management or Smart Health for ambient assisted living. The prerequisite of any smart system is a sensor rich and datafied environment. The data gained are used by predictive algorithms to predict future behavioral patterns and optimize the resources accordingly. In the next development stage, the transition from prediction to prescription takes place: future behavior is not only anticipated but formed. Both predictive smart systems and prescriptive smart systems rely on actionable data in order to "know ahead and act before"thus anticipating and shaping the future. Their focus is on process-oriented efficiency.
In contrast, a Smart Society needs to imagine ahead and act accordingly. Its perspective should be open and broad profiting from the insights provided by the humanities and the arts. The anticipatory stance in the Smart Society should not be technology driven. The art of anticipation –not the technology of anticipation - may open up new perspectives, and provide stimuli for smart and ethically sound innovations.

Keywords

Smart City, Smart Society, AI,microdirectives

Primary author: THÜRMEL, Sabine (Technische Universitaet Muenchen)

Session Classification: Data, AI, and Responsibility I

Contribution ID: **25** Type: **Individual Paper**

Looking for a Mark of Intelligence

In this paper I defend the explanatory and theoretical importance of the notion of intelligence in cognitive science and AI by setting it apart from the related notions of cognition and rationality, with which intelligence is often confused. I put forward a capacity-based mark of intelligence that helps reveal the distinctive explanatory role that the notion of intelligence plays in the relevant sciences. On this picture, intelligence involves the possession of an interdependent cluster of capacities to behave, i.e. the capacities to behave in general, flexible, adaptive, and goal-directed ways. I argue that this mark of intelligence sheds light on key aspects of the minimal functional architecture required for intelligence. I suggest that such architecture involves representational capacities of a specific kind, namely that of producing and using structural representations.

Keywords

Nature of Intelligence; Intelligent capacities; Functional Architecture; Internal Representation; Structural Representation

Primary author: COELHO MOLLO, Dimitri (Humboldt-Universitaet zu Berlin)

Session Classification: ”Intelligence” in Artificial Intelligence

Contribution ID: **30** Type: **Individual Paper**

Governing AI and Society: Worker, Expert, and Policy Engagements with Artificial Intelligence/Automation at the Port of Los Angeles

Wednesday, 7 July 2021 16:15 (30 minutes)

Industrial ports increasingly seek to leverage artificial intelligence (AI)/automation as a means of reducing labor costs and carbon emissions in a globally competitive marketplace. The Port of Los Angeles, the busiest port in the western United States and a major economic engine for Southern California, recently announced technological upgrades for the adoption of autonomous vehicles for the movement of cargo. The introduction of AI-based automation risks the displacement of human labor and severe community loss, and the announcement was met with worker backlash resulting in a series of public hearings over automation/AI and the future of work. Drawing on the public understanding of science framework (Wynne 1992), this paper examines these hearings to compare public engagement with AI across three relevant social groups: workers, policy makers, and experts. Qualitative analysis of hearing transcripts (Charmaz 2014) compared public perceptions of societal impact of AI as well as levels of engagement and claims-making across groups. We find that current public engagement with emerging technologies is uneven across social groups and risks reproducing existing inequalities. We call for greater engagement from AI experts and data scientists in strengthening AI governance in society, and a shared commitment of working together in a technology-driven world.

Keywords

ethics of AI; automation; public understanding of science; public engagement; port logistics; future of work

Primary authors: Dr CRUZ, Taylor (California State University, Fullerton); CHEN, Austin (California State University, Fullerton); MOORE, Emily (California State University, Fullerton); PARK, Jaewoo (California State University, Fullerton); GORDILLO, Andrea (California State University, Fullerton)

Session Classification: Governance of AI

Contribution ID: **32** Type: **Individual Paper**

A Falsificationist Account of Artificial Neural Networks

Machine learning operates at the intersection of statistics and computer science. This raises the question as to its underlying methodology. While much emphasis has been put on the close link between the process of learning from data and induction, the falsificationist component of machine learning has received minor attention. In this talk, I argue that the idea of falsification is central to the methodology of machine learning.

It is commonly thought that machine learning algorithms infer general prediction rules from training data. This is akin to a statistical procedure by which estimates are obtained from a sample of data. But machine learning algorithms can also be described as choosing one prediction rule from an entire class of functions. I will argue that this is in line with the idea of *falsification*. In particular, I formulate a falsificationist account of artificial neural networks. This opens the door to conceiving of artificial neural networks as operating actively within a falsificationist regime. I will discuss how this falsificationist account has consequences for understanding artificial neural networks, in particular regarding their *explainability*.

Keywords

Methodology of Machine Learning, Falsification, Artificial Neural Networks

Primary author: Mr BUCHHOLZ, Oliver (University of Tübingen, Cluster of Excellence "Machine Learning: New Perspectives for Science")

Session Classification: Philosophy of Computing and Machine Learning

Contribution ID: **34** Type: **Individual Paper**

Ethics State of AI Startups in Turkey: A Look at the Developers' Perspectives

In this paper, we analyzed the ethical and social aspects of AI systems from the developers' perspectives on data privacy, transparency, and accountability. To contextualize our inquiries, we wanted to understand the current ethics state of AI startups in Turkey from the developers' perspectives to deep dive into their technical oriented worlds. In order to get their points of view, we have designed our research as a digital ethnography. We conducted in-depth interviews with 24 interviewees selected according to simple random sampling on the business-focused social media platform, LinkedIn. From our observations and interviews, we can say that the conscious of engineers, data scientists, coders, computer scientists are low in the social aspects of AI systems. Still, their main concern and focus are to develop technology with high technical capability. In terms of privacy, transparency, and accountability, most of them think that these issues are essential, but they perceive that these are not the works that "technical" employees should be concerned about. In these issues, they often refer to their companies in regulations and actions. We saw that the gap between technical development and social & ethical aspects still exists in the developers' perspectives, mindsets, and workplaces.

Keywords

Artificial intelligence, AI ethics, digital ethnography, data privacy, transparency

Primary authors: Ms KILIC, Busra (Istanbul Bilgi University); Mr SAKA, Erkan (Istanbul Bilgi University Media Studies Department)

Presenters: Ms KILIC, Busra (Istanbul Bilgi University); Mr SAKA, Erkan (Istanbul Bilgi University Media Studies Department)

Session Classification: Business Ethics and AI Startups

Contribution ID: **35** Type: **Individual Paper**

Data Privacy: Stakeholders Conflicts in Medical IoT

MIoT (Medical Internet of Things), AI and Data Privacy are linked forever in a Gordian knot. This paper explores the conflicts of interests between the stakeholders regarding data privacy in the MIoT arena. While patients are at home healthcare hospitalization, MIoT can play a significant role in improving the health of large parts of the population by providing medical teams with tools for collecting data, monitoring patient's health parameters and even enabling remote treatment. While the amount of data handled by MIoT devices grows exponentially, different stakeholders have conflicting understandings and concerns regarding this data. MIoT technology is in its early phases and hence a mixed qualitative and quantitative research approach will be used, which will include case studies and questioners in order to explore this issue and provide alternative solutions.

Keywords

MIoT, Data privacy, Stakeholders, Home Healthcare, Information privacy, AI

Primary authors: Mr SAND, Benny (Ben-Gurion University, Israel); Prof. LURIE, Yotam (Ben-Gurion University of the Negev, Israel); Prof. MARK, Shlomo (SCE - Shamoon College of Engineering)

Session Classification: Privacy II

Contribution ID: **38** Type: **Individual Paper**

Rethinking Data Infrastructure and its Ethical Implications in the Face of Automated Content Generation

When it comes to the processes of scholarship and publication, the evolution of data storage and presentation technologies demand that we rethink the basic processes, in particular the nature of anonymity and the mechanisms of attribution. There are two approaches to address anonymity: The first overly emphasizes the anonymity of authors; the other is what we see in a traditional double-blind review process. The former provides the authors'pseudonymity to achieve 'academic freedom', while the latter ensures the anonymity of reviewers to promote 'fair reviews and academic quality/integrity.'Rather than arguing the merits of either of these perspectives, we propose to reconsider how their respective goals are achieved, and perhaps simultaneously reconciled, in light of emerging digital storage technologies that may better support the mechanisms of attribution (and fulfilling broader goals of accountability, transparency, and trust). We discuss the scholarship review and publication process in a revised context, specifically the availability of a digital storage infrastructure that can track data provenance while offering: immutability of stored data; accountability and attribution of authorship; and privacy-preserving authentication mechanisms. Our metaScribe system supports these features. We believe such features would allow us to reconsider the nature of identity and anonymity in this domain, and to broaden the ethical discussion surrounding new technology. Considering such options in an underlying storage infrastructure means that we could discuss the epistemological relevance of published media more generally.

Keywords

ethics of electronic publishing, immutable data storage, blockchain technology in scholarship, AI-automated content generation, authorship and authentication

Primary author: ISRAEL, Maria Joseph (Santa Clara University)

Co-author: Dr AMER, Ahmed (Santa Clara University)

Contribution ID: **39** Type: **Individual Paper**

Experimental Machine Ethics and the Problem of Entrenchment

The Moral Turing Test (MTT) has been suggested as a solution to the testing problem of artificial morality under widespread moral disagreement. The test requires a metaethical underpinning ofr its justification, which is provided by the pragmatist view on ethics as constraint by human perspective, iterative-experimental, and open to reasonable pluralism. However, the practical implications of this reconstructino of the rationale for the MTT confronts certain practical problems, in particular that of technological entrenchment. The problem of ethical experimentation at scale and the obstacles it faces is illustrated by the domain of autonomous individual transportation.

Keywords

Moral Turing Test, autonomous transportation, pragmatism, technological entrenchment

Primary author: MERDES, Christoph

Session Classification: Ethics of AI Ethics

Contribution ID: **40** Type: **Individual Paper**

It does not work but it still is dangerous: Evading the double bind of AI ethics

Short abstract (extended abstract attached)
Applications of AI in research and development often come with huge ambitions and the evocation of making the hitherto impossible possible. Thus, one important task of an ethical engagement with AI is debunking such stated possibilities, showing that such technologies do actually not work as stated or intended. A second, equally important task for ethicists is to show that such attempts are ethically problematic or even dangerous.
However, these two tasks come with a potentially contradictory logic. If the technology does not work (as the first task shows) then why is it so dangerous? Vice versa, in order to show the ethical problems of e.g. surveillance, the surveillant power of the technology is often presupposed.
Thus, ethicists trying to pursue both tasks find themselves in a kind of double bind. They need to say something like: "It does not work but it still is dangerous."
The contribution illustrates the emergence of this double bind in recent papers from AI ethics. A second step traces the emergence of the double bind to two common characteristics in analyzing AI from an ethical perspective: a representative logic and an implied neutrality of algorithms opposed to data. The contribution concludes in discussing ways of avoiding the double-bind by engaging with these two features of analysis.

Keywords

ethics of AI, representation, neutrality, double bind, artificial intelligence

Primary author: MATZNER, Tobias

Session Classification: Ethics of AI Ethics

Contribution ID: **42** Type: **Symposium/Panel Proposal**

Exploring the Possibility and Ethics of AI Paternalism in Health Apps

Health apps aim to promote their users'health by tracking health related data, and by influencing their users to act in a healthier manner. They might thus be considered to be "persuasive technologies", which raises concerns about how they might affect their users'autonomy. Some argue that such health apps in fact promote their users'autonomy, by better allowing them to pursue their (authentic and autonomously chosen) goals. Others suggest that these apps diminish autonomy, by infantilizing and unduly manipulating their users. In any case, it is clear that such apps do have an influence on their users'decision-making and behavior, and thus affect users'autonomy.

It can be assumed that these types of apps are not actors themselves, since they do not act autonomously—however weakly defined. In contrast, consider health apps which include AI technology and are capable of

1. analyzing their users'behavior in light of the individual user's tracked data and on the ba-sis of a more exhaustive database of numerous individual health profiles,
2. drawing conclusions concerning which behavior would benefit the individual user, and
3. influence the individual user's behavior accordingly, for example, by way of making "nudg-ing" suggestions on what to do, or by means of gamification.

Such AI based health apps might arguably be considered to be autonomous agents when it comes to influencing the users'behavior for their own good. If so, it seems that AI based health apps do not only raise common ethical questions about their influence on the users'autonomy, but also gain a paternalistic valance.

However, consider the following traditional philosophical notion of paternalism:

"X acts paternalistically towards Y by doing (omitting) Z:

1. Z (or its omission) interferes with the liberty or autonomy of Y.
2. X does so without the consent of Y.
3. X does so only because X believes Z will improve the welfare of Y (where this includes preventing his welfare from diminishing), or in some way promote the interests, val-ues, or good of Y."

(Dworkin, https://plato.stanford.edu/entries/paternalism/)

Following this definition, one might question whether the concept of paternalism can plausibly be applied to AI based health apps. Firstly, an AI does not have beliefs or intentions, which feature prominently in paternalistic action. Yet, as an instance of persuasive technology, such apps are clearly at least goal-oriented concerning their users'health. Secondly, one might suggest that the user makes a free decision to install the app and only then be subjected to its influence. Yet, the initial decision to install and the ongoing interaction need to be distinguished. Even when the user installs the app autonomously, over time, recommendations made by it may nevertheless qualify as paternalistic influence on concrete decision-making. Hence, assuming that the phenomenon of AI based health apps influencing their users'behavior for their own good is not too far-fetched, one might argue that it is rather the definition of paternalism that needs to be adapted to fit the phenomenon.

In addition to these conceptual questions, the symposium/panel is intended to address resulting ethical issues. Aside from questions of influencing users'autonomy, the question of which health-related goals specifically may count as good for the individual user needs to be discussed. Although an interest in maintaining one's health may reasonably be attributed to everyone, one might question whether the notion of health employed by AI based health apps captures all the nuances usually associated with this concept, and whether it might be applied to all users equally. Usually,

any externally defined notion of what is considered good for a person raises serious worries and fuels our liberal anti-paternalistic consensus. Furthermore, paternalistic influences may come in many different flavors, ranging from straightforward interferences in freedom of action to subtly influencing or 'nudging'one's decision-making process. Given the increase in linked 'smart'technologies, the potential of AI based health apps—and AI based technology in general—to influence our choices and liberties may significantly increase in the future.
The symposium/panel is intended to engage in critical discussion of all these points and thereby introduce and emphasize the possibility and ethics of AI paternalism in health apps.
The symposium/panel is planned for one slot of 90 minutes.

Keywords

AI, Paternalism, Health Apps, Autonomy

Primary authors: KÜHLER, Michael; Dr STOPPENBRINK, Katja (Universität Münster); Dr WHITE, Lucie (Universität Hannover)

Presenters: KÜHLER, Michael; Dr STOPPENBRINK, Katja (Universität Münster); Dr WHITE, Lucie (Universität Hannover)

Contribution ID: **43** Type: **Individual Paper**

Moral Responsibility and Explainable AI

Decisions made in high-stakes contexts such as finance, medicine, and policing are increasingly driven by algorithms. These decisions can have morally adverse consequences. Although software developers are generally assumed to bear at least partial responsibility for these consequences, the increasingly widespread use of machine learning challenges this assumption. Because many of the algorithms developed using machine learning are opaque, software developers might no longer possess the ability to predict, intervene, and justify that is characteristic of morally responsible agents. In this talk, we will evaluate the extent to which software developers can rely on Explainable Artificial Intelligence (XAI) to avoid this loss of moral responsibility. By considering several different kinds of XAI methods, we will show that although software developers can regain the ability to predict and justify, more effective methods are needed to control and intervene systematically. Thus, although XAI allows developers to retain some of the features of morally responsible agents, we argue that it does not yet allow them to retain them all.

Keywords

Machine Learning, Explainable AI, Moral Responsibility, AI Ethics

Primary authors: ZEDNIK, Carlos; Dr WIDDAU, Christoph Sebastian (OVGU Magdeburg)

Session Classification: Ethics of AI Decisions

Contribution ID: **44** Type: **Individual Paper**

A method for responsibility in innovation in AI firm start-ups

The integration of social and ethical responsibility aspects in artificial intelligence (AI) innovation processes is a considerable challenge for AI start-up firms. This paper presents a method (or modular assessment toolbox) for recognising, supporting and maintaining firm-level efforts to innovate by taking some possibly hidden or unwanted social implications into account. The tool was developed as part of a series of research projects for developing trust in AI in Austria, funded by the Austrian Research Promotion Agency (FFG). We describe the case example of developing the tool together with an AI start-up, and how it became possible to explore and operationalise four responsible innovation dimensions (anticipation, reflexivity, inclusion, and responsiveness) in a replicable method in an AI firm environment. Although the method cannot be a substitute for collective technology governance at the societal and policy levels, we argue that it makes an essential step towards opening up micro-level innovation processes to unpredictability in societal impacts and identifies solutions for responsiveness. From an academic perspective, the method helps link the technology governance field with organisational culture in software start-up firms. Further research will address the applicability of the method to the broader AI innovation community.

Keywords

responsible innovation, AI, start-ups, technology assessment, tool

Primary authors: Dr SINOZIC, Tanja (Institute for Technology Assessment (ITA), Austrian Academy of Sciences (ÖAW)); BETTIN, Steffen (Institute for Technology Assessment (ITA), Austrian Academy of Sciences (ÖAW)); Dr UDREA, Titus (Institute for Technology Assessment (ITA), Austrian Academy of Sciences (ÖAW)); Dr SCHÖNAUER, Annika (Forschungs- und Beratungsstelle Arbeitswelt (FORBA) (Working Life Research Centre)); Dr SMITH, Lisa (Prewave GmbH); MAMMES, Georgia (Prewave GmbH)

Presenter: Dr SINOZIC, Tanja (Institute for Technology Assessment (ITA), Austrian Academy of Sciences (ÖAW))

Session Classification: Business Ethics and AI Startups

Contribution ID: **45** Type: **Symposium/Panel Proposal**

Towards Gender Just Artificial Intelligence

Gender bias is frequently understood as an ethical deficit of humans whereas machines are considered neutral and unbiased. Today we know that technologies can be biased in at least three ways (Bath 2009; 2014; Michelfelder et al. 2017): they duplicate biased social norms, ignore physical-biological traits, or impose cognitive norms. AI has entered our lives with the promise of acting in a fair and neutral manner, thus being beneficial to the whole of society. However, AI algorithms, models and techniques are frequently biased and lead to discrimination of people in terms of their assumed gender, class or race (e.g., Ali et al. 2019; Eubanks 2017; Hoffmann 2019; Noble 2018; O' Neil 2016; Tannenbaum 2019; Wachter-Boettcher 2018; Yapo & Weiss 2018).

In this panel, we follow Peter-Paul Verbeek (2011) and regard technologies as moral agents. AI systems are a special case since they are more autonomous than most other technologies. They can even exhibit "relegation" on their users (Wellner 2020). But so far, they do not exercise justice of any kind. From the user's perspective, justice can become meaningful in concrete intra-actions (Barad 2007) between humans and machines, which is even more relevant as AI learns with every interaction (Bratteteig and Verne 2018). Consequently, the results produced by AI-based systems occasionally turn out to be different than intended by their designers. As these systems are trained on past data, responsible development of AI transcends the boundary between design and use (Suchman 2002) and call for new ethical approaches.

One of the major challenges in implementing ethics in technologies is that ethics –and especially gender justice issues –involve contradicting values rather than achieving a yes/no answer (Michelfelder et al. 2017; Puech 2016). In this context, our panel explores alternatives to the beaten path of applying vague ethical guidelines (originally developed for humans) to machines (Jobin et al. 2019). If we want to integrate the norm of gender-justice into AI-based systems, we need to deal with the challenge of translating this ethical value into technical features of AI systems. This is not a simple translation because gender-justice is a complex and context-sensitive concept. We will present a project proposal that aimed at developing methodologies to implement values in AI machines via philosophical, empirical and technical expertise, inspired by the value-sensitive design approach (Friedman & Hendry 2019) taken into the realm of Machine Learning.

There are different models and dimensions of gender justice, which philosophy, STS and gender studies offer, such as subjective, structural and symbolic dimensions of gender (Pimminger 2012, Harding 1986); equality of resources or capabilities (Nussbaum 2000); and ethics of care (Puig de la Bellacasa 2017). AI raises new questions for these theories and at the same time shed a new light on ethics of AI. For example, ethics of care can provide a new direction for the development of gender-just AI that transcends the mere analysis of power relations inherent to technologies by feminist studies.

Current AI-based systems do not allow for human feedback or adaptation, once deployed. Experiments where machines learnt in real-time from their users (recall Microsoft's chatbot Tay) have often ended with less-than-impressive results. The problem is to make a machine learn and adapt, but not learn everything it comes across. The challenge is not how to learn, but how to distinguish between the ethical and the unethical. The technical challenge is to ensure that the AI adapts to the social context, while observing some ethical guidelines. This is particularly challenging, since both the context as well as the ethical values change over time and cultures. We will analyze which socio-technical factors contribute to changing human-machine conversations towards gender-just dialogs, and thus enhance the use of AI for affirmative action processes, especially in light of the fact that users can learn from the interaction with the AI.

By initiating a truly interdisciplinary discussion that involves gender studies, philosophy of technology and computer science perspectives, our panel tackles two concerns: A) how to define and model gender bias and gender justice in AI systems; and B) how to integrate ethical values into AI algorithms. We thus aim at addressing the challenge of how to "translate"contingent context-sensitive social values and ethics into engineering development processes and technical systems.

Keywords

Gender, values, bias, value-sensitive design, interdisciplinary approach

Primary authors: WELLNER, Galit (Tel Aviv University); Prof. BATH, Corinna (TU Braunschweig); Dr NALLUR, Vivek (University College Dublin)

Presenter: WELLNER, Galit (Tel Aviv University)

Contribution ID: **48** Type: **Individual Paper**

Insights and considerations from a project to derive a tentative ethical guideline for the value-based development of AI systems in medicine

AI-based technologies are often ascribed a transformative and disruptive character. Therefore, ethical design of respective systems must be regarded as being imperative for a socially acceptable development and implementation that achieves a balance between the benefit for the specific ap-plication area and the benefit for society and thus the common good. Particularly in highly sensitive areas of society such as medicine, these considerations are crucial elements for a successful devel-opment and implementation of AI applications themselves. Users and other stakeholders could also regard them as key aspects for acceptance. Developing ethical guidelines for the application of AI-based technological solutions takes a broad approach. Ideally, this approach should take into account the diversity of theoretical discourses on the one hand, and the perspectives of stakehold-ers and considerations of their specific requirements, needs and concerns regarding AI applications on the other. The aim of the paper is to show a possible way to derive a preliminary ethical guide-line for an AI-based diagnostic medical application.

Keywords

artificial intelligence, ethical guideline, medicine

Primary authors: Mr SONAR, Arne (Universität zu Lübeck); Prof. WEBER, Karsten (OTH Regensburg)

Session Classification: AI and Medical Practices II

Contribution ID: **49** Type: **Individual Paper**

Developing moral machines: Hybrid implementation of ethical theories

Intelligent and especially autonomous machines, that increasingly interact with humans, must have some degree of independent moral decision-making capability. If humans are to trust and accept sophisticated machines as interaction partners, they must be convinced that the machines share their essential values and concerns, or at least act accordingly.
For moral machines to be accepted, it is important that they behave similarly to humans because this increases trust in the moral machine. In moral psychology, a distinction is made between models in which the basis of moral action is rationalistic or intuitionistic, or which are based on a combination of these two. Ethical theories can be divided into generalist, particularist and coherentist theories. In terms of implementation approaches in AI, there are logic-based top-down approaches, sub-symbolic bottom-up approaches, and the hybrid approach, which is a combination of these.
This circular bottom-up/top-down structure is thus found in all three areas considered here. Because of the isomorphism found in the circular structure, the hybrid approach in implementing a coherentist ethical theory into an AMA is a viable method for creating a convincing simulacrum of the human capacity for abstract moral reasoning.

Keywords

Maschinenethik, Hybride Implementierung, Generalismus, Partikularismus

Primary author: SOHRMANN, Beate

Session Classification: Moral Machines

Contribution ID: **50** Type: **Individual Paper**

A Value-Centered Exploration of Data Privacy and Personalized Privacy Assistants

While privacy notices are becoming ever more prevalent on our phones and online, it is well-accepted that they are not sufficient to manage our data privacy. In Part I, I briefly outline the current challenges data privacy notices face, ranging from dark patterns, information overload, and the privacy paradox. I then suggest that we move away from notice-and-consent and take a value-centered approach to data privacy decision-making. I argue that this value-focused understanding can be best conceptualized as a weighed expression of user values. Because this weighed expression can be viewed as a form of autonomous choice, I then draw upon Suzy Killmister' s four-dimensional theory of autonomy as a systematic way to incorporate an agent's values. In Part II, I utilize this conception of autonomy to suggest that user value expression can be optimized through Personalized Privacy Assistants (PPAs) - machine-learning systems that provide personalized privacy recommendations and automate privacy choices for a user. Applying Killmister to possible PPA use in the context of user smartphone application selection, I then suggest that PPAs could successfully realize the self-definition dimension of autonomy with remaining challenges to self-realization, self-unification, and self-constitution. With a few small changes, I conclude that PPAs have considerable potential to realize user expression of values when selecting smartphone applications and beyond.

Keywords

Digital privacy, privacy assistant, privacy notices, user autonomy, values

Primary author: CARTER, Sarah E. (Data Science Institute (DSI), National University of Ireland – Galway (NUIG))

Presenter: CARTER, Sarah E. (Data Science Institute (DSI), National University of Ireland –Galway (NUIG))

Session Classification: Privacy II

Contribution ID: **51** Type: **Individual Paper**

Before machine learning in medicine: What do we mean by medical expertise?

The development of AI algorithms is rapidly increasing in the field of medicine. Especially within image-based medicine, machine learning algorithms seem to be able to reach high levels of accuracy in performing medical tasks. This is often seen as an adequate indicator of these systems' potency to support or even replace medical experts in their professional tasks. However, the accuracy reached within those digital systems is fragile, since its information is based on the knowledge of experts and it is still up to discussion *who* can be called a medical expert. Normally, it is assumed that a pathologist or radiologist with the appropriate educational record can be called a medical expert. Nevertheless, for the high levels of accuracy demanded by these algorithms, you have to determine certain 'golden standards'of medical practice, i.e. standards for measuring the value of algorithmic data by means of the highest level of medical expertise. Moreover, a problematic aspect of the algorithm's dependence on medical expertise is, is that experience, intuition and implicit knowledge play a role in medical decision making. It is questionable to what extent those can and should be included in the development of machine learning algorithms. Within this article, we therefore argue that medical expertise should be scrutinized *before* developing machine learning algorithms in medicine. Specifically, we emphasize the necessity to establish new collaborative methods by which medical expertise can be applied to machine learning algorithms and to investigate what it means for medical expertise to be adapted to an algorithmic format. In this way, machine learning algorithms can be put to use in medical science and perform unique and innovative tasks in medical practice, without ignoring experience, intuition and all the other intrinsically human qualities of medical experts essential to medical decision making.

Keywords

Phil. of Technology, Epistemology, Machine Learning, Image-based Medicine, Medical Expertise

Primary author: DROGT, J.M.T.M. (UMC Utrecht/Utrecht University)

Co-authors: Prof. BREDENOORD, A.L.; Dr JONGSMA, K.R.; Dr MILOTA, M.M.; Dr VOS, S.; WYATT, Sally; LYSEN, Flora

Presenter: DROGT, J.M.T.M. (UMC Utrecht/Utrecht University)

Session Classification: AI and Medical Practices I

Contribution ID: **52** Type: **Symposium/Panel Proposal**

Ethical AI: between theory and practice

Tech companies spurred by new developments in the domains of AI and algorithmic decision-making have taken over the steering wheel of society. From facial recognition to securing the border to personalized health diagnosis and treatment, many central domains in human life are being transformed by data-driven innovations. Yet, as the manipulation of voters and the use of discriminatory facial recognition applications illustrate, ill-considered and poorly regulated technologies can induce grave harm, infringing upon key social and democratic values like privacy, dignity and fairness.

Next to implementing legal frameworks to mitigate these problems, a wide variety of data ethics solutions has seen the light: from issuing policy documents to publishing numerous ethical principles and guidelines. While principles and guidelines might partially fill the gap between lagging legal solutions and disruptive technological developments, they are not problem and risk-free. First, it is unclear how to understand and put them into practice. Second, there is the problematic tendency to use ethics to wash away concerns raised about a company's behavior, by picking ethical principles that limit one's actions as little as possible while simultaneously presenting oneself as doing 'good'(ethics washing).

The aim of this panel is twofold. It will critically evaluate the role of Data & AI Ethics and point towards directions to head towards when designing ethical technology. This panel will bring together empirical, normative, conceptual and applied research on this topic in order to come to a nuanced and in-depth discussion.

Our panel starts with a case-study on Facebooks AI suicide prevention program by **Tineke Broer**. Several private and public organizations have implemented programmes drawing on algorithms to better estimate the risk of committing suicide. Maybe the most disputed one is Facebook's programme. After several cases of livestreaming suicides on Facebook's platform, the company developed a 'proactive'algorithmic programme for detecting suicide risk. Based on a media-analysis, a mixed perception of this suicide prevention program becomes apparent. Some argue that it is part of Facebook's responsibility as one of the biggest social media platforms worldwide to pro-actively take action. Others, by contrast, claim that, Facebook's program is a form of ethics-washing, aiming to instill trust in a company that, simultaneously, monetizes data on mental health.

Esther Keymolen will continue this discussion on trust by tackling the question: what should it take for a tech company to be(come) genuinely trustworthy? The standard view in the philosophical debate is that trustworthy agents must: a) give assurances indicating their trustworthiness, b) be competent in some domain, c) and commit to putting their competences to work in the service of others. Because this literature takes a distinctly second-person approach, it cannot straightforwardly be applied to tech companies. This talk will bridge the gap between this second-person approach and the context of tech companies and present an account of trustworthiness tailored to tech companies. Keymolen will argue that tech companies can only be trustworthy if they signal trustworthiness through design, attract virtuous tech-employees, and invest in robust organizational commitments.

In the third presentation, **Gijs van Maanen** will take up questions of ethics washing and shopping from a political philosophical angle. Drawing from research on casuistry and philosopher Raymond Geuss, Van Maanen will make a case for a question, rather than theory or principle-based ethical data practice. The emphasis of this approach is placed on the acquisition of a thorough understanding of a social-political phenomenon like tech development. This approach should be replenished with one extra component to the picture of the repoliticized data ethics drawn so far: the importance of 'exemplars', or stories. Precisely the fact that one should acquire an in-depth understanding of the problem in practice will also allow one to look in the past, present or future for similar and comparable stories from which one can learn.

From a practice-oriented perspective, our panel ends with **Merel Noorman** discussing two experimental clinics that were set up to help project teams address their questions about how to ensure that public interests and values are safeguarded in the design and use of city crowd-management systems. A key aim of the clinics was to translate abstract ethical values and guidelines to the governance structures around existing development practices. Through these analyses the clinics provided insights into possible interventions that could be made in the shaping of the governance structures and the design of the technology. In her presentation, Noorman will reflect on the lessons that can be drawn from these clinics about what it means to go from ethical theory to practice and back again.

The AI-induced challenges we currently face are often messy and complex. For data ethics to have a meaningful impact, it is necessary to bring together different methods and units of analysis. By interlinking empirical, applied, conceptual and normative approaches, this panel wants to show how this can be done.

Keywords

Data ethics, ethics washing, trust, political philosophy, tech development

Primary author: VAN MAANEN, Gijs (Tilburg University)

Contribution ID: **53** Type: **Individual Paper**

Coping with Black Boxes. The Quest for Epistemic Transparency in Artificial Intelligence

On the background of two contrasting case studies from contemporary Artificial Intelligence, this paper mounts a qualified defence of epistemic opacity in computer modelling and simulation. If epistemic transparency is understood as analytic tractability or the 'ability to decompose the process between model inputs and outputs into modular steps'(Humphreys 2004), situations of analytical intractability do not necessarily give rise to conditions of opacity that would affect the system under consideration as a whole. A comparative discussion of Deep Neural Networks and Behaviour-based AI will help to demonstrate that, under certain circumstances, analytical intractability need not conflict with, or may even subserve, the representational properties of the epistemically relevant elements or of the model as a whole. The model might be 'transparent enough'for the purposes of inquiry or application. In consequence, there might be no unitary concept, criterion or norm of epistemic transparency for computer models to rely on.

Keywords

Artificial Intelligence; Models in Science; Computer Simulations; Black Box Problem; Behaviour-Based Artificial Intelligence; Deep Neural Networks

Primary author: GREIF, Hajo (Warsaw University of Technology)

Presenter: GREIF, Hajo (Warsaw University of Technology)

Session Classification: Transparency and Explainable AI

Contribution ID: **56** Type: **Individual Paper**

Statistics Do Not Disclose Personal Information or Violate Privacy

Despite continuing controversy regarding theories of informational privacy, at least one premise has seemed clear: Privacy is to be understood in terms of *disclosures of personal information.* A theory of privacy tells us, among other things, which cases of disclosure count as privacy violations. This article addresses an under-appreciated, although foundational, question: *Under what conditions does a true attribution of personal information count as a disclosure?* I argue that attributions of personal information purely on the basis of statistical inference do not count as disclosures. This conclusion is crucial for the relationship between data analytics and informational privacy. Many ethicists and commentators have been concerned about analysis of massive datasets yielding hitherto unacknowledged conclusions about individuals. Prominent theorists of data ethics have classified these as privacy violations, noting the risk that "big data"poses for privacy. This article supports an alternative view: When such cases are harmful, they are best conceived as a distinct type of harm, with as much in common with harms of defamation as with privacy violations.

Keywords

personal information, disclosure, statistical inference, privacy, ethics of belief

Primary author: KING, Owen

Session Classification: Privacy I

Contribution ID: **57** Type: **Individual Paper**

Synthetic Minds and Mental Disorders

Synthetic intelligence—in addition to the properties of reasoning, ethical thinking, judgment, and other characteristics of the human mind—will also exhibit a propensity for mental illness or mental disorder. Mental disorder can be understood here as a sort of malfunction in a synthetic mind or brain. It is obviously a natural feature of the human mind, so it is inherent in the brain's design. What this means is that the human brain cannot be recreated as a synthetic construct (i.e., an artifact) without this feature. In other words, mental disorder is, to some extent, a design feature of the natural brain, so any faithful synthetic brain must incorporate mental disorder by definition. This paper therefore discusses how we can conceptualize dysfunctions in synthetic minds when they exhibit pathological behavior.

Keywords

synthetic mind, psychotic disorders, mental illness of syntethic mind

Primary author: KRZANOWSKI, Roman (The Pontifical University of John Paul II)

Co-author: Dr KRZANOWSKI, Jacob (The South London and Maudsley (SLaM) NHS Foundation Trust)

Presenter: KRZANOWSKI, Roman (The Pontifical University of John Paul II)

Session Classification: Synthetic Minds and Consciousness

Contribution ID: **58** Type: **Individual Paper**

What in the World? The Site of the Ethical in Human-AI Relations

In this contribution, I argue that ethical assessments of AI require a pragmatic, yet philosophically robust reconsideration of the 'site'of inquiry. This requires an abandonment of translating issues of interpretability or fairness into decontextualized mathematical models in the technical field; and a more detailed engagement with actual technological deployments in the philosophical field without neglecting that ethics can only be realised in human contexts. The primary ethical issue I consider in this contribution is AI's technical opacity to those affected by AI-driven systems. Starting from a post-phenomenological theoretical stance, I will argue that this framework has the potential to address the 'hermetic'shortcomings of the technical fields (e.g., decontextualising AI technologies), but runs the risk of its own brand of hermeticism (e.g., separating AI and human into distinct ethical spheres). To counter this, I propose that both need to move beyond dialectical ethics, and suggest that the gap between human experience and AI technologies actually constitutes the appropriate site for ethical considerations. Specifically, I propose that empirical-analytical research into techniques that both constitute and transcend such sites, such as dimensionality reduction, will bring forth a deeper understanding of how human and AI are entangled.

Keywords

post-phenomenology, interpretability, ethics, discorrelation

Primary author: BENJAMIN, Jesse Josua (Universiteit Twente)

Session Classification: Human Machine Relations II

Contribution ID: **60** Type: **Individual Paper**

Looking for justice in fairness. Filling the normative gaps in the regulatory discussion of algorithmic decision-making systems

Wednesday, 7 July 2021 16:15 (30 minutes)

Regulating algorithmic decision-making (ADM) challenges policy makers as applications of ADM touch upon basic rights and values. A salient issue in recent risk-based regulatory proposals is fairness, mostly vaguely linked to normative claims of justice. In this paper, we focus on two flaws in the current algorithmic fairness discourse that impede to tackle larger justice claims.
The first flaw is the confused relationship between fairness and justice. We argue that fairness operates as procedural justice that can only judge the correctness of a process but cannot make justice claims concerning outcomes. Instead, justice claims must be introduced as reference points consciously beforehand.
The second flaw derives from the implicit premise of justice as anti-discrimination, mirroring the procedural take on justice as a technical problem of adjusting 'unfair'treatment between individuals. Anti-discrimination is an important concern but tends to narrow the focus on the individual realm, only. This leaves a blind spot of the structural problems that lead to injustice in the first place that cannot be tackled referring to protected classes. The confused relationship between fairness and justice complicates larger dimensions of justice, e.g. equality of opportunity, intersectionality or representation, which are needed to improve the situation of people affected.

Keywords

Justice, Fairness, Normative Dimension

Primary authors: BAREIS, Jascha; FOLBERTH, Anja (Karlsruher Institute for Technology)

Session Classification: Fairness and Justice in AI

Contribution ID: **61** Type: **Individual Paper**

All variables are not created equal.

The desire to systematise and quantify social science data for use in AI systems is growing. This article argues that social science data often cannot be converted into accurate categories on which algorithms work best. The literature raises this concern in a general way. Deeks (2018) notes that legal concepts, such as 'proportionality'cannot be easily converted into code. He notes that "The meaning and application of these concepts is hotly debated, even among lawyers who share common vocabularies and experiences,"(Deeks, 2018, pg. 1570) This article recognises that this is true, and provides a framework that explains why some concepts are difficult to codify, and allows practitioners to systematically assess the concepts they wish to use in quantitative analysis. Using the example of recidivism prediction, this paper demonstrates that AI systems often use very heterogeneous data sets. When this is so, the predictions yielded by these systems are likely to be unreliable, and potentially prejudicial. The paper finishes with recommendations for improving AI systems in the social sciences.

Keywords

Recidivism, Prediction, Variable construction, Nomadic concepts

Primary author: GREENE, Catherine (LSE)

Session Classification: Algorithmic Discrimination

Contribution ID: **62** Type: **Individual Paper**

Can 'Taking Responsibility' as a Normative Power close AI's Responsibility Gap?

- **Short Abstract**

Artificial intelligence (AI) increasingly executes tasks that previously only humans could do, such as driving a car, fighting a war, or performing a medical operation. However, as the very best AI systems tend to be the least controllable and the least transparent, some scholars argued that humans could no longer be morally responsible for some of the AI-caused outcomes, which would then result in a 'responsibility gap'. In this paper, I assume –for the sake of argument –that at least some of the most sophisticated AI systems do indeed create responsibility gaps and I ask whether we can bridge these gaps at will, viz. whether certain people could take responsibility for AI-caused harm simply by communicating the intention to do so, just as people can give permission for something (via consent) simply by communicating the intention to do so. So understood, taking responsibility would be a genuine 'normative power'. I first discuss and reject the view of Champagne and Tonkens, who advocate a view of taking prospective liability. According to this view, a military commander can and must, ahead of time, accept liability to blame and punishment for any harm caused by autonomous weapon systems under his command. I will then defend my own proposal of taking retrospective answerability, viz. the view that people can makes themselves morally answerable for the harm caused by AI systems, not only ahead of time but also when harm has already been caused.

Keywords

Autonomous Weapon Systems; AI; Responsibility Gap; Taking Responsibility

Primary author: KIENER, Maximilian (University of Oxford)

Session Classification: Data, AI, and Responsibility I

Contribution ID: **64** Type: **Individual Paper**

When AI Makes Rational Choice: A Challenge to the Possibility of a Universal Ethical Domain

The development of Artificial Intelligence systems raises important ethical questions not only because AI systems would participate in the human society in ways that invite moral judgments. It is also an interesting fact that they exhibit a kind of intelligence who can take agency in making moral choices. Philosophers and social scientists have been interested in characterizing and understanding the human faculty of instrumental rationality, our ability to choose an appropriate means in order to achieve the best, preferred end. In this paper, I examine the possibility of a universal, general AI, an ideally rational agent who is thought to make decisions by analyzing all the possible options and then computing the best means to an end. Such an AI could then form the foundation of an ethical AI with a universal moral domain. However, I present a challenge based on a puzzle about how humans make choices. I argue that the puzzle reveals a fundamental concreteness in our decision making that significantly reduces the possibility of a universal, ethical AI.

Keywords

Ethical AGI, Ethical AI, Rational Choice, Instrumental Rationality

Primary author: SONG, Ziming (Sun Yat-sen University)

Session Classification: Ethics of AI Decisions

Contribution ID: **65** Type: **Individual Paper**

Strictly Human: Limitations of System Autonomy and the Design of Autonomous Systems

Short abstract (See the attachment for the extended abstract)

Are there human activities that autonomous systems cannot perform? How does the answer to this question inform the design of autonomous systems? Evaluations of a technology or its features should be sensitive to the activities in which it is used. Each activity can be described by its overall goal, governing norms, and the intermediate steps taken to achieve the goal. This paper uses the activity realist approach to conceptualize autonomous systems in the context of human activities. By doing so, it first argues for epistemic and metaphysical conditions that demonstrate which activities autonomous systems can and cannot perform, and second, it highlights the ramifications of the limitations of system autonomy on the design of autonomous systems.

Keywords

Autonomous systems; Activity realism; Agency

Primary authors: SOLTANZADEH, Sadjad; Dr BOSHUIJZEN VAN BURKEN, Christine (University of New South Wales)

Session Classification: AI, Values, and Design

Contribution ID: **66** Type: **Individual Paper**

In-Between the Lines and Pixels: Cartography's Transition from Tool of the State to Humanitarian Mapping of Deprived Urban Areas

Cartography has been, in its pre-modern and modern production of maps, influential in determining how space and territory is experienced and defined. Advancements in telecommunications (e.g. the internet and geovisualization software) and geoinformation systems and geoinformation science (GIS), have transformed cartographic practice from a tool of predominantly state apparatus, to a scientific, commercial and humanitarian enterprise. This is exemplified in the use of remote sensing (RS) techniques to acquire, process and visualise images of the Earth. In the last decade, RS techniques have increasingly incorporated Artificial Intelligence (e.g. Convolutional Neural Networks, Random Forest and Support Vector Machines) to improve the speed and accuracy in the identification and classification of features in remotely sensed images. This paper will investigate the use of these techniques in the classification of deprived urban areas referred to as 'slums'and 'informal settlements'in the Global South. Using a spatial humanities and postphenomenological methodology, this paper shall analyse the role of classification and use of geo-information in shaping how deprived urban areas are defined in terms of the space and territory they occupy, and the socio-political as well as ethical impact of this classification on those living in these areas.

Keywords

postphenomenology remotesensing spatialhumanities AI ethics

Primary author: OLUOCH, Isaac (University of Twente)

Session Classification: Novel Practices in Design and Regulation of AI

Contribution ID: **68** Type: **Individual Paper**

Perception vs. reality of neutral AI

Current examples show that the use of Artificial Intelligence (AI) can lead to discrimination against (mostly already marginalized) groups in society (Krauß, 2018), contradicting the demands for human-centered AI use. There are various sources of discrimination that can be found in every stage of development and application. AI-biases are difficult to predict ex ante, as well as difficult to correct ex post (Beck et al., 2019). However, AI may also bring around positive changes for justice. While human decision-making is a partly unconscious process that, due to its unconscious nature, cannot be verbalized or reported, the decision-making process of an AI can (theoretically) be made transparent and explainable (XAI). Thus, AI can improve the detection of discrimination (Kleinberg et al., 2019). We will provide an overview of sources for AI bias as well as potential benefits of an AI application considering its decision-making processes. This discussion is not sufficient to fully elaborate how a human-centered AI use can be achieved. A user-centered perspective must be taken into account, to evaluate the fit between users'expectations towards as well as actual performance of an AI. We will present empirically informed results on users'opinions regarding fairness of AI decisions.

Keywords

AI, discrimination, perception, bias, decision-making

Primary authors: WEBER, Andrea (Hochschule für angewandte Wissenschaften Würzburg-Schweinfurt); HENKING, Tanja (Hochschule für angewandte Sozialwissenschaften Würzburg-Schweinfurt)

Session Classification: Algorithmic Discrimination

Contribution ID: **69** Type: **Individual Paper**

Subjectivity and intended gaps of information within digital client records in social work

There is an intensive debate on how we should deal with AI-based systems. But at the same time, there are many areas whose degree of digitalization is (still) low. In my contribution I would therefore like to focus on the process of documentation in digital client record systems in social work. Between January and June 2020, I conducted 20 guideline-based interviews with experts working either for service providers or service deliverers. One of the main findings is that there are intended gaps of information within digital client records. This is information which is not written down in digital documentation software, although it is important for further case processing. Findings are important for the current debate on AI because they underline that we should not only focus on the ethics of design, but also (and once again) on the ethics of usage.

Keywords

digital documentation, social work, privacy, subjectivity, IT

Primary author: SCHNEIDER, Diana

Session Classification: Digital Life and Ethics

Contribution ID: **70** Type: **Individual Paper**

ETHICAL REFLECTIONS ON HANDLING DIGITAL REMAINS: Computing professionals picking up bones

The essence of a digital life is digital information, often on different platforms and composed of different media. Once X digitally exists, data about X can be collected from external sources. X's digital transactions are recorded. Other entities might publish reviews of X.

It stands to reason that if there is digital life, there is the possibility of digital "death."A digital life may reflect a human life. For example, a collection of content from a single individual X posted on social media platforms could be interpreted as X being digitally alive. When X physically dies, this digital collection may live on, perhaps accidentally, or perhaps intentionally. Once such a presence exists, that existence can end; in this paper, we call that end "digital death."A digital death can lead to "digital remains."We will define digital death as: "the end of the digital presence of an entity."

Others have looked at ethical issues about X's digital death from the perspective of X, from the perspective of X's significant others, and from the perspective of those producing relevant laws and regulations. In this paper, we look at ethical issues about X's digital death from the perspective of computing professionals who design, develop, and deploy the digital reality within which X "lives" or has lived. We use three scenarios to analyze these issues.

Keywords

Digital death, ethics, digital remains

Primary authors: Dr GRODZINSKY, Frances (Sacred Heart University); MILLER, Keith (University of Missouri - St. Louis); Dr WOLF, Marty (Bemidji State University)

Session Classification: Human Life and Trust

Contribution ID: **71** Type: **Individual Paper**

Others'information and my privacy: a new ethical dimension of conceptualization

Abstract

Privacy has been understood as about one's own information; other people's information is not typically considered with regards to one's own privacy. This paper aims to raise this issue of conceptualizing privacy in relation of oneself and other people's information. To motivate the topic, we use current application of forensic genealogy as an example to illustrate how others'information may breach one's privacy or identity. Forensic genealogy may appear like an unusual case for most people's privacy concern. However, we will see that one's privacy and others'information has become an increasingly prevalent issue, especially with the rapid development of genetic informatics and recommender systems. Both offer abundant cases where information that falls outside of one's own comes back to have an impact on oneself, a situation to which the emerging conceptualizations of group privacy is responding. The fact that others'information comes to challenge our conceptualization of privacy presents a new ethical dimension that is yet to be recognized in our continued understanding of privacy.

Keywords

privacy, forensic genealogy, information ethics

Primary author: MA, Yuanye

Session Classification: Privacy I

Contribution ID: **72** Type: **Individual Paper**

Ethical Principles for Cybersecurity AI Applications: some controversial issues

To harness the "disruptive"potentials of new AI applications researchers as well as government organizations all around the world developed a whole body of ethical guidelines or principles to which technology developers should adhere to as far as possible.

In a recent study (Hangerdorff, 2020) it has been shown that the effectiveness of these guidelines or ethical codes is almost zero and they have no influence on the behavior of professionals from the tech community.

To further investigate this issue we contextualize some of the ethical requirements related to AI applications to the cybersecurity field. Starting from generic ethic requirements as described in literature, we assumed their implementation in the cybersecurity context and envisaged the correlated benefits and potential risks which could derive from such an implementation. The requirements we considered have been chosen among those mostly reported in literature namely: privacy protection, transparency, human oversight and cybersecurity.

The outcomes obtained are reported in this paper.

Keywords

Cybersecurity, AI Applications, Ethical principles

Primary authors: Prof. BRUSCHI, Danilo (Università degli Studi di Milano); Dr DIOMEDE, Nicla (Università degli Studi di Milano)

Session Classification: Cybersecurity and Adversarial Attacks

Contribution ID: **73** Type: **Individual Paper**

Community-in-the-loop: towards pluralistic value creation in AI, or—why AI needs business ethics

Today, due to growing computing power and the increasing availability of comprehensive, high-quality datasets, artificial intelligence (AI) technologies are increasingly entering many areas of our everyday life. Thereby, however, a number of significant ethical concerns arise, including issues of fairness, privacy and human autonomy, which so far have been addressed mainly through the formulation of principles and the development of general guidelines. With this article we take up recently raised concerns about a too principled approach (cf. B. Mittelstadt 2019; Greene, Hoffmann, and Stark 2019; Whittlestone et al. 2019) arguing that besides difficulties in putting principles to practice AI ethics is currently often driven by a focus on technical systems and their direct ethical implications. However, this results in neglecting important ethical questions at the organisational level of firms that introduce AI into the market as part of their corporate strategy. We therefore advocate to further incorporate business ethics in AI ethics in order to better take into account the business context of AI. To this end we present a contractualist approach to business ethics as normative theory conceptualising ethical conflicts between different values and interests as incomplete contracts. Building on this we argue that the concept of responsible innovation provides a suitable framework for dealing with AI from a business ethics perspective. We show how responsible innovation offers a procedural normative framework for stakeholder engagement to enable mutual understanding, promote trust and eventually even create shared value. By introducing a business ethics-based approach to responsible innovation, we are able to contribute both on a theoretical level to the current AI ethics debate as well as on practical level. We discuss implications of our findings for policy and firms that can facilitate responsible AI at both technological as well as business levels.

Keywords

AI ethics; business ethics; contractualism; deliberation; responsible innovation

Primary authors: HÄUSSERMANN, Johann Jakob; Prof. LÜTGE, Christoph (TU Munich)

Session Classification: Business Ethics and AI Startups

Contribution ID: **76** Type: **Individual Paper**

Stoic Philosophy and Technology: Who is in Control?

Abstract
This paper explores the control problem of Technology and how Stoic Philosophy addresses that problem, as well as offers a response to its solution. It comprises three parts: 1. The Core Problem of Stoic Philosophy and the Information and Artificial Intelligent (AI) Technologies to which I will refer collectively for convenience as Technology. 2. The Core Principles of Stoic Philosophy and How they Apply to Technology, and 3. Why Stoic Philosophy is of relevance and importance to Technology. Let me begin by clarifying, that the problem of technology is not technology as such, but the use of technology by the Big Tech companies for their financial gain to the detriment of the users'wellbeing and society generally. Ultimately the core problem of technology is not technical to be solved by engineers but a moral problem to be solved collectively by society and the global community. The overarching rationale for the application of Stoic Philosophy to evaluate the impact of technology on society is that stoic philosophy as a way of life with its focus on the attainment of wellbeing is singularly relevant and important to the eudaimonic assessment of technology.

Keywords

Stoic Philosophy; Technology; Control; Big Tech; Facebook; Google

Primary author: SPENCE, Edward Howlett

Session Classification: Human Machine Relations I

Contribution ID: **77** Type: **Individual Paper**

A Hybrid Model for Moral AIs

There are two paradigms of AI development, one being the symbolic approaches requiring explicit programming, and the latter based on machine learning. The advantage of machine-learning AIs is that it can be trained from large datasets to excel in features that are hard to program. However, one important feature of machine learning approaches is its inscrutability. We argue for a hybrid model for moral AIs: First, corresponding to the level of conscious processing, programable AIs may be developed to deal with complex moral situations and solve the inscrutability problem. Second, corresponding to unconscious moral sensibilities, deep learning may be used to train AI systems (with training data from ordinary people's moral judgement and behaviour) to develop moral sensibilities reasonably close to ordinary human beings. These two systems are to be further integrated to generate a comprehensive system that can mimic a typical human moral agent.

Keywords

moral AI, machine learning, codifiability thesis, anti-codifiability thesis, utilitarian AI, hybrid model for moral AIs

Primary authors: Dr SONG , FEI (Nazarbayev University); Mr YEUNG , Shing Hay Felix (The University of Hong Kong)

Session Classification: Moral Machines

Contribution ID: **78** Type: **Individual Paper**

AI In Justice

Wednesday, 7 July 2021 16:45 (30 minutes)

Application of AI to the justice system demands that we carefully consider whether the algorithms are just. While this statement seems simple, it highlights a deep and fundamental oversight in discussions of the ethical use of AI in judicial matters. It is an understandable oversight. The focus regarding the ethics of AI use in this context often instead deal with particular algorithms or problems regarding misapplication due to lack of training by users.
However, this oversight is particularly exasperated (and particularly dangerous) when AI is used to aid and augment a human judge's judgment. We show how this structural problem is ubiquitous in AI systems currently used in sentencing and incarceration decisions within the US and how it is easily remedied through a simple adjustment to the consultation and documentation processes surrounding their use. We offer an analysis of this use case in the risk assessment algorithms domain and further justify our proposed remedy's value in this context with arguments drawn from classical game theory.

Keywords

AI Justice Ethics

Primary authors: ALTMAN, Ryvenna; AMER, Ahmed (Santa Clara University)

Session Classification: Fairness and Justice in AI

Contribution ID: **79** Type: **Individual Paper**

Ethical concerns in India's AI policy

Wednesday, 7 July 2021 16:45 (30 minutes)

Artificial Intelligence (AI) is the ability of machines to take decisions as intelligent agents. AI now controls parts of our lives through AI-based services, gadgets and software. Unlike other technologies which merely comply to human instructions, AI has the ability to actively make decisions like a human. With this unprecedent ability of agency, AI faces the same ethical questions as humans, e.g. what is the right conduct? Ethical guideline provides a code of conduct. There are two possible ethical guidelines for AI: (i) how AI should act, and (ii) how organizations should develop and employ AI. They are treated as a single issue in this write-up because AI is oriented to serve specific objectives for an organization (*weak* AI).

Since the interaction between humans and AI will increase in future, with proportionate consequences, now is the time to establish the ethical guidelines regarding development and implementation of AI. The guidelines shall be universal in spirit. However, their application at the will need to be sensitive to the unique situations of application. This is why not only many governments, but also non-governmental and corporate bodies have been developing their own AI policies. For a judicious development of AI, it is necessary that these policies themselves be cross-examined for ethical concerns.

In this short write up, first we explore the questions that India needs to answer for the ethical guidelines for the use of AI. Then, using the social justice framework, we motivate some ethical concerns for India to consider.

Keywords

AI ethics; AI policy of India; Responsible AI

Primary authors: Dr BHARGAVA, Pranesh (BITS Pilani Hyderabad); Dr SATPATHY, Suchismita (BITS Pilani Hyderabad); Dr DASH, Biswanath (BITS Pilani Hyderabad Campus)

Presenters: Dr BHARGAVA, Pranesh (BITS Pilani Hyderabad); Dr SATPATHY, Suchismita (BITS Pilani Hyderabad)

Session Classification: Governance of AI

Contribution ID: **80** Type: **Individual Paper**

Beyond informativeness in the epistemology and ethics of XAI in healthcare

In this presentation, we aim at building upon the framework of what we define the *informativeness account* of the relation between epistemology and ethics of machine learning (ML). In particular, we refer to the methodology adopted by Mittelstadt et al. (2016). Our first goal is to show that, according to these authors, epistemological issues are *instrumental* to and *independent* from ethical considerations. This means, to our mind, that they take the epistemology as being uninfluenced and unregulated by non-epistemic normative elements. Our second goal is to argue that this approach neglects that ethical aspects also have a substantial influence on the epistemological assessment. Such influence is not to be understood as merely informative but rather regulatory of the epistemology. We refer to this as the *blended account*. We substantiate our claims through the analysis of explanatory AI in healthcare referring to a concrete example. Drawing on the latter, we claim that the informativeness account fosters epistemic practices that render patients vulnerable to cases of epistemic injustice. Finally, we claim that the blended approach can be functional to limiting cases of epistemic injustice arising in AI-mediated practices in healthcare.

Keywords

Epistemology and ethics of XAI, informativeness, explanatory AI in healthcare, epistemic injustice

Primary authors: POZZI, Giorgia (TU Delft); DURÁN, Juan Manuel (TU Delft)

Presenter: POZZI, Giorgia (TU Delft)

Session Classification: Transparency and Explainable AI

Contribution ID: **81** Type: **Individual Paper**

Are Autonomy Algorithms Acceptable?: Advancing A New Debate in Medical AI Ethics

Autonomy algorithms are a technology, first proposed in 2018, to determine what a patient incapable of autonomous choice would decide for his or her medical treatment, if he or she were capable. The algorithm would mine electronic health records, demographic information, and social media content, to predict a patient's preference. The prediction could inform clinicians'decision making in scenarios such as end-of-life treatment. Some bioethicists have embraced autonomy algorithms, arguing that they could be better even than human surrogates at predicting what a patient would want. Other commentators object that autonomy algorithms cannot live up to the success rates of "morally intimate"surrogates, and furthermore, fail to reduce the actual anxiety of doctors and loved ones when making choices about death. This extended abstract proposes two empirical studies about autonomy algorithms, one cross-cultural, and the other about surrogate intimacy. The cross-cultural study would be borne from data showing that different cultures, such as Japan and the United States, have different norms about who participates in end-of-life decisions. The surrogate intimacy study would examine more closely if morally intimate surrogates are indeed better predictors than surrogates or statistics. Together, these studies could advance, on multiple fronts, the debate on autonomy algorithms.

Keywords

autonomy algorithms, end-of-life, medical ethics, artificial intelligence

Primary author: CAMPANO, Erik (Umeå University, Sweden)

Presenter: CAMPANO, Erik (Umeå University, Sweden)

Session Classification: AI and Medical Practices II

Contribution ID: **82** Type: **Individual Paper**

Ethical Consequences of the Computer System's Intentionality in Machine Learning

The computer systems' intentionality recently becomes a debatable topic throughout the field of contemporary Information and Computer Ethics. However, this notion still lacks profound philosophical and ethical elaboration. This study is premised on the idea that the best illustration of the computer system's intentionality may be found in machine learning algorithms. In terms of this paper, the latter represents a powerful knowledge magnification tool, which plays a vital role in today's moral decision-making. In this interpretation, technological intentionality stands close to the so-called "algorithmic bias" and "algorithmic black-box" problems. Simultaneously comprehension of the nature of technological intentionality may serve as one of the possible solutions to these sorts of issues. Although the computer system's intentionality sounds like a theoretical term, it still has profound ethical potentialities. The crucial moral impact of machine learning algorithms lies in changing the decision-maker's general frame of information. As soon as every moral agent needs information for accomplishing specific moral tasks, algorithmic change in a current frame of information may lead to significant moral consequences.

Keywords

computer system's intentionality, epistemological component of moral action, artificial agency, machine learning algorithms, artificial autonomy

Primary author: MYKHAILOV, Dmytro (School of Public Administration, philosophy department, Nanjing Normal University)

Session Classification: Ethics of AI Decisions

Contribution ID: **83** Type: **Individual Paper**

AI Alignment and the Complex Structure of Human Values

Stuart Russell has proposed that we should achieve AI alignment by building AI agents whichlearn our values or preferences from our behaviour, then select actions in accordance with them.One challenge to this approach is that different people value different things, but I set this challenge aside in order to focus on a second, which is that there is considerable complexity in the structure of individuals'values. When thought of as an agent engaged in reinforcement learning, each of us has a reward function; we also represent phenomena as valuable in various ways; and these representations giverise to our dispositions to make choices. Different versions of Russell's approach could take any of these, or some combination, as the target for alignment. I examine some of the advantages and disadvantages of each of these options.

Keywords

AI alignment; value learning; human values

Primary author: BUTLIN, Patrick (King's College London)

Session Classification: AI, Values, and Design

Contribution ID: **84** Type: **Individual Paper**

War in the Age of Algorithms: Morality, Law, and Autonomous Weapons Systems

This paper challenges the widespread belief that the current law of armed conflict provides an adequate criterion for assessing the permissibility of using autonomous weapons systems - 'killer robots'- in armed conflict. I defend two claims. First, AI's capacity to comply with the current law is neither a necessary nor a sufficient condition for the moral permissibility of its employment in war. The widespread claim that autonomous weapons systems should be banned on account of the fact that they would not be able to comply with the law is therefore both dangerously misguided and misleading. Second, AI systems'actions should instead be governed by norms that are fundamentally different from, and more restrictive than, the current law. A more adequate framework for assessing the permissibility of using AI in war will, unlike the law itself, be sensitive to individual moral rights, and more closely resemble certain aspects of rights-based, 'revisionist' views of the ethics of killing in war, as well as the international human rights framework. Ultimately, the paper reveals that the legal norms that should apply to AI systems in armed conflict must differ significantly from the principles that govern human combatants'conduct in the same circumstances. To advance debates about whether 'killer robots'should be banned, the paper thus clarifies the moral principles that should play a role in this debate in the first place.

Keywords

ethics, war, killer robots, law, AI

Primary author: EGGERT, Linda (Harvard University)

Session Classification: Ethics and Automated Warfare

Contribution ID: **85** Type: **Individual Paper**

On the Moral Treatment of Social Robots: Towards a Hybrid Approach of Moral Consideration

This paper concerns the question of human moral duties towards social robots—robotic companions, caregivers, pets, and other machines that are built to interact with human beings on a social level. The issue is a pressing one, because social robots will become considerably more sophisticated in the near future while their moral and legal standing remains a contested point. In the present paper, I argue that neither the idea of "moral status", which is commonly associated with mental properties like consciousness, sentience or rationality, nor the social-relational approach to moral consideration, which finds moral significance solely in the fact that humans tend to socially bond with these interactive machines, are sufficient to fully grasp the ethical texture of the situation. This calls for what we may call a hybrid view of moral consideration, which takes into account both the intrinsic properties of the machine that are constitutive of moral status, as well as extrinsic ones, such as the capacity for social interaction and anthropomorphic appearance.

Keywords

Moral consideration; social robots; moral status; hybrid approach

Primary author: MOSAKAS, Kęstutis (Vytauto Magnus University)

Session Classification: Human Machine Relations II

Contribution ID: **86** Type: **Individual Paper**

Should We Consider a Robot to be a (Moral) Agent?

There are different views on the morally relevant capacities of robots. While some regard robots to be basically advanced forms of computers, others consider them rather to be agent to which we can ascribe responsibility, that we should treat with respect and ascribe rights to them. In this paper we investigate the moral status of robots with two methodological tools. In a first step, we will analyze robot behavior and possibilities of trust towards them on the basis of promise theory. In a second step, we will ask how their reflexive and judgmental capacities can be evaluated within the framework of philosophical anthropology. In a third part we will bring those two lines of investigation together and strive for an integrated framework regarding the moral status of robots. We expect that such a framework will as well provide us with some insights in the nature of morality and likewise will it form an important step towards a philosophically acceptable robot ethics. In a fourth and final step, we will outline the consequences of our considerations.

Keywords

Robot ethics, Promises, Philosophical Anthropology

Primary authors: Dr DÜWELL, Marcus (Minstroom Research BV); Prof. BERGSTRA, Jan (Universiteit van Amsterdam)

Presenter: Dr DÜWELL, Marcus (Minstroom Research BV)

Session Classification: Artificial Moral Agents

Contribution ID: **88** Type: **Individual Paper**

Ethical Behaviourism: yet another property based approach for machine moral status?

There are two main types of arguments in the debate concerning the moral status of machines: the properties approach and the relational approach. My aim is to investigate the most influential 'property'based approaches to moral status of machines and show that they all have a generic problem with soundness of the arguments. Then, I focus on the criterion of 'behavioral performance', John Danaher's 'ethical behaviorism'-which seems to be falling under the properties approach- to investigate to reveal whether this position can be understood as a bridge between the properties approach and the relationist approach, and find out if this position can escape the generic mistakes of properties approach. For this, I'll bring in the discussion a very recent term coined by Johanna Seibt: sociomorphing, which is claimed to be capturing more accurately what happens in human-machine social interactions as it defines our direct picking up of real social cues and behaviors of machines even if this sociality is asymmetrical. While Seibt herself doesn't conclude anything about the moral status of machines from the idea of sociomorphing, I'll take both Danaher's and Seibt's accounts into consideration and discuss the implications of behavioral criterion for the moral status question.

Keywords

moral status of machines, properties approach, ethical behaviorism, sociomorphing

Primary author: GÖKMEN, Arzu (PhD Candidate, Project Member)

Session Classification: Moral Machines

Contribution ID: **89** Type: **Individual Paper**

Is the problem with AI intelligence rather than artificial? The ethical significance of perspectives in evolutionary psychology and cultural evolution

In considering issues related to AI, ethicists and the public have tended to focus on the artificial nature of AI, differences between AI and humans. However, scientists have made progress in the development of AI by making programs more human, conceiving of intelligence in terms of the capacity to learn. As a result, ethicists should reconsider issues in terms of similarities between humans and AI, from the perspective of intelligence rather than artificial. This approach would dispel some concerns while raising others. To demonstrate it, my presentation sketches ethical implications of work in psychology and evolution for AI, since they illuminate non-obvious features of cognition. To do so, my presentation is divided into two parts. I begin by considering questions of autonomy, responsibility, and AI, showing how work in psychology can reframe the gravity of these concerns –it is unclear that humans are generally autonomous or capable of responsibility in a manner that would be undermined by AI. I move on to consider work in evolution, explaining how the nature of learning raises questions about the environments in which AI's should be deployed –uniquely human intelligence is based on selective social learning, although this is not always fitness enhancing.

Keywords

artificial intelligence, evolutionary psychology, cultural evolution

Primary author: CLANCY III, Rocky (TU Delft)

Presenter: CLANCY III, Rocky (TU Delft)

Session Classification: "Intelligence" in Artificial Intelligence

Contribution ID: **90** Type: **Individual Paper**

There Is Still At Least One Reason for Making Artificial Moral Agents

I argue that there is at least one reason for making artificial moral agents, contrary to Aimee van Wynsberghe's and Scott Robbins'argument (2019). Van Wynsberghe and Robbins argue that for any machine, whether they are autonomous or not, only its safety feature is enough. I argue, on the contrary, that when a machine is complex enough and is autonomous, the "safety"quality of the machine becomes so qualitatively different from a normal safety feature of an ordinary machine that it merits another concept. That is, when a machine becomes autonomous, it's increasingly difficult to separate between the normal safety feature and the ethical feature. This is so because, I argue, the ethical part of the machine becomes integral to its excellence, i.e., its ability to perform its function well. Thus, if a machine is to be fully functional, it cannot fail to be ethical. This idea has an ancient root that is found both in the eastern and western traditions. In the talk I will answer some of the more important objections against the idea, such as the charges that the argument conflates ethics with functioning well, and that ethics requires intention.

Keywords

artificial moral agents, ethics, reasons for making

Primary author: HONGLADAROM, Soraj (Chulalongkorn University)

Session Classification: Artificial Moral Agents

Contribution ID: **92** Type: **Individual Paper**

Ethically aligned Deep Learning: Unbiased Facial Aesthetic Prediction

Facial beauty prediction (FBP) aims to develop a machine that automatically makes facial attractiveness assessment. In the past those results were highly correlated with human ratings, therefore also with their bias in annotating.
As artificial intelligence can have racist and discriminatory tendencies, the cause of skews in the data must be identified. Development of training data and AI algorithms that are robust against biased information is a new challenge for scientists.
As aesthetic judgement usually is biased, we want to take it one step further and propose an Unbiased Convolutional Neural Network for FBP. While it is possible to create network models that can rate attractiveness of faces on a high level, from an ethical point of view, it is equally important to make sure the model is unbiased.
In this work, we introduce AestheticNet, a state-of-the-art attractiveness prediction network, which significantly outperforms competitors with a Pearson Correlation of 0.9601. Additionally, we propose a new approach for generating a bias-free CNN to improve fairness in machine learning.

Keywords

Fairness in Machine Learning; Responsible Artificial Intelligence; Discrimination Prevention; Facial Aesthetics; Unconscious Bias

Primary author: DANNER, Michael (Reutlingen University)

Co-authors: Mr WEBER, Thomas (Reutlingen University); Ms PENG, Le Ping (Hunan University of Science and Technology); Mr KAZMAIER, Markus (Reutlingen University); Mr OETTINGER, Markus (Reutlingen University); Ms FEI, Wu (Xi'an Polytechnic University); Mr RÄTSCH, Matthias (Reutlingen University)

Presenter: DANNER, Michael (Reutlingen University)

Session Classification: Issues of Facial and Emotion Analysis

Contribution ID: **93** Type: **Individual Paper**

A philosophical perspective on outcome and procedural fairness criteria for Machine Learning and the need for dynamic modelling

Wednesday, 7 July 2021 17:15 (30 minutes)

Fair Machine Learning (ML) research aims to provide and improve criteria for the fairness of ML algorithms. We review the proposed metrics which usually evaluate either the fairness of the distribution of goods, opportunities etc. as produced by the algorithm's decision (*outcome-based criteria*) or the fairness of the process itself which is used to arrive at a decision (*procedural criteria*). By choosing a criterion, a decision maker subscribes to a goal as well as a view of justice implicit in the metric. We propose explicating the underlying moral values in order to clarify the relation between different fairness metrics and enable decision makers to align their choice with their moral goals. After analyzing existing approaches, we conclude that a *temporal* perspective is necessary in order to model the goals of many real world fairness interventions such as affirmative action and urge ML researchers to develop dynamic fairness criteria that can capture an intervention's influence over time.

Keywords

Machine Learning, Fairness, Moral Goals, Ethics of Algorithms

Primary authors: Dr REMMERS, Peter (TU Berlin); Mrs SCHWÖBEL, Pola (Technical University of Denmark (DTU))

Session Classification: Fairness and Justice in AI

Contribution ID: **94** Type: **Individual Paper**

On the Limits of Safe Moral Enhancement through AI Mentors

Much attention has been paid to exploring how AI techniques could improve our moral behavior. Even though it might be in principle impossible to suppose that any AI system could do our moral thinking for us, there is still a substantial role that AI mentors could play in our moral education and training. In particular, an AI Socratic interlocutor might offer considerable improvements to education and training techniques in wisdom traditions such as Stoicism, even if the AI is manifestly not a competent moral agent. Such a technology is a more modest and safer gesture towards moral enhancement than what has been proposed by transhumanist scholars and their ilk. Instead of imagining some morally perfect artificial agent telling us what to do, a wiser course is to train up a competent Socratic interlocutor in concrete wisdom tradition to serve as a discussion partner and coach. It would be no more dangerous to agency than an interactive book: an AI Socratic Interlocutor could offer new insights and perspectives and exercise our intellect without directing it. We will illustrate the idea by describing how it might enhance the practice of Stoicism, with its emphasis on disciplined and ongoing intellectual and philosophical training.

Keywords

Artificial Intelligence; Moral Enhancement; Artificial Moral Agent; AI Socratic Interlocutor

Primary authors: Dr GABRIELS, Katleen (Maastricht University); Prof. VOLKMAN, Richard (Southern Connecticut State University)

Presenter: Dr GABRIELS, Katleen (Maastricht University)

Session Classification: Human Machine Relations I

Contribution ID: **95** Type: **Individual Paper**

Autonomous Driving and Public Reason: A Rawlsian Approach

With the prospect of fully autonomous driving (AD) on the horizon, adequate political answers to normative challenges of AD become increasingly pressing. Some of these challenges concern the AI based automated decision-making processes essential for the function of AD: It is as of yet unclear which set of normative criteria should be used to guide the decision-making processes of an autonomous vehicle –a problem not only for potential real-world trolley cases but for the distributions of risk in mundane driving situations as well. At the heart of these challenges lies the problem of reasonable pluralism: The fact that there exists a plurality of reasonable yet incompatible comprehensive moral doctrines (religions, philosophies, worldviews) within liberal democracies, so that a politically acceptable answer cannot refer to only one of these. Following the political philosophy of John Rawls, there is a solution to the problem of reasonable pluralism, according to which a politically adequate answer does not have to come at the expense of an ethical answer: The idea of public reason. We argue that a Rawlsian justificatory framework is adequate for answering the normative challenges of AD and elaborate on the way it might be employed for this purpose.

Keywords

autonomous driving; reasonable pluralism; public reason; trolley cases; risk distribution

Primary authors: Mr SCHMIDT, Michael (Karlsruhe Institute of Technology); Mrs BRÄNDLE, Claudia (Karlsruhe Institute of Technology)

Presenters: Mr SCHMIDT, Michael (Karlsruhe Institute of Technology); Mrs BRÄNDLE, Claudia (Karlsruhe Institute of Technology)

Session Classification: AI and Regulation

Contribution ID: **96** Type: **Individual Paper**

On the Epistemic Soundness of AI Personality Inferences based on Visual Data

What are the epistemic justifications that warrant computer vision artificial intelligence (AI) to make inferences about \textit{personality} based on portrait images? Are such justifications supported by reasons that refer to the epistemic soundness of the inferences themselves or the practical advantages that such inferences may have for the parties involved (or both)? In this paper, we investigate the rationality of the fundamental assumptions underlying the technological affordances of computer vision AI to make inferences about personality. Our theoretical account draws on empirical results from research on automatic personality inferences based on visual data. We propose that the underlying assumptions of personality inferences from images exhibit circularity when they rely on epistemic grounds that they seek to overcome: drawing semantically meaningful concepts from any form of visual data necessarily depends on human data labelling.

Keywords

computer vision AI; rational inferences, visual data

Primary authors: ENGELMANN, Severin (Technische Universitaet Muenchen); Prof. GROSSKLAGS, Jens (Technical University of Munich)

Presenter: ENGELMANN, Severin (Technische Universitaet Muenchen)

Session Classification: Issues of Facial and Emotion Analysis

Contribution ID: **97** Type: **Symposium/Panel Proposal**

Beyond Ethics: Proposing a Political Economic Framework to Interrogate and Supplement Ethics in AI Policymaking

Artificial Intelligence (AI) refers to a collection of computational technologies which make decisions semi-autonomously by learning from patterns obtained from pre-existing data. Since a major part of human actions which produces economic value are tied to constrained intelligent decision making, the economic consequences of AI are non-trivial. Over the last decade this potential of AI for productivity, logistics, as well as for surveillance, forms of automated control over workers like digital Taylorism etc. have not been unnoticed by states and the industry. The increasingly ubiquitous use of AI and the sheer speed and scale of AI operations has deep implications on society, human rights, and economics with often AI artefacts outpacing the decision making of policymakers.

In this environment one approach to policymaking is of normative ethical frameworks that guide how AI artefacts and systems should be designed, developed, and deployed. These frameworks are used by various stakeholders like governments, intergovernmental agencies, and the private industry to indicate their intent and in some cases, also explicitly spell out use cases that will not be pursued. For example, the United Kingdom has stated it wants to become the world leader in ethical AI. The European Union has drawn ethics initiatives, including Ethical Principles for Trustworthy AI published by the EU High Level Expert Group on AI. In May 2019, forty-two nation-states signed the OECD's Principles for Trustworthy AI, etc.

The private sector also has enthusiastically adopted the ethics framework. In June 2018, Google published its AI Principles announcing their intention to build socially beneficial AI systems that would not create or reinforce biases and would be safe and accountable. It released a list of applications it would not pursue as well. Microsoft published ethical principles under the umbrella of Responsible AI and made a committee for giving recommendations on what AI to deploy, the AI and Ethics in Engineering and Research Committee (AETHER). Facebook funded a research institute at the University of Munich. In response to such developments, technical institutions, academia, and civil society has also enthusiastically engaged with this framework, adding to it, responding to consultations, participating in multi-stakeholder spaces.

Recently, skepticism of the ethics as a framework has started permeating academic and technical discussions. Political scientist Benjamin Wagner has coined 'ethics washing'as the phenomenon of these frameworks being posed as an alternative or preamble to regulation. Existing principles are vaguely worded with no grounding in law and thus there is no accountability or redressal mechanisms. Even ethical standards for technical communities do not materially affect the design or development of AI –research has shown that the ACM code of ethics had no observed effect on the work of software engineers who were explicitly asked to consider it.

Why is this so? We propose that a major factor which is being ignored is the political economy of and around AI which deeply impact AI design, development, and deployment, and thus renders ethics as a toothless ineffective framework unable to understand the reality of AI development. Ethical frameworks in their focus on looking at "impact"of AI as artefacts do not look into the long pipeline of how AI is produced, which is important because one of the biggest impacts of AI is on labour power, both directly due to production of surveillance mechanics and indirectly by incentivising a different mode of organising work, namely platforms. Even if one only looks at AI as commercial products the ethical framework fails because it ignores the profit motive. For example, one may rightly critique the racial and gender bias inherent in many medical datasets but the fact that those datasets will be used anyway at many places is simply because readymade datasets are freely obtained and its not in the mandate of some small research team of a company to gather data for prohibitive costs. The reasons regulations are avoided throughout nations and

ethical frameworks are pushed is itself political economic in nature, there is a fear of hampering innovation and thus missing the bus compared to the competing team, company, industry or company. Thus, the logic of finance capital, the particulars of industrial processes in production of AI, the relationship of the tech-industries with their workers, etc. cannot be ignored when designing a framework for ethics.

In this symposium the four speakers represent one aspect of this puzzle on how to interrogate and improve ethical frameworks. Prof. Bidisha Chaudhuri works on the future of works and ethics, and will explore the techno-social aspect of rooting ethics in materiality, Vidushi Marda is an expert in AI and human rights and will advocate augmenting the ethical normative framework with HR based regulation, Prof. Shishir Jha will connect the problem of AI ethics with that of the precariat labour in platformisation, and Prof. Anupam Guha who works in AI policy and labour will moderate the panel and suggest a framework for combining the political economic lens with the ethical one.

The tentative idea for the symposium slot is to have 4 talks by the panelists on the topic for 10 minutes each, followed by a panel discussion of thirty minutes. This will be followed by a twenty minutes question answer session with the audience

Keywords

interrogating-ethics political-economy platformisation

Primary authors: Prof. GUHA, Anupam (Centre for Policy Studies, Indian Institute of Technology Bombay); MARDA, Vidushi (ARTICLE 19); Prof. CHAUDHURI, Bidisha (IIIT Bangalore); Prof. JHA, Shishir (CPS, IIT Bombay)

Presenters: Prof. GUHA, Anupam (Centre for Policy Studies, Indian Institute of Technology Bombay); MARDA, Vidushi (ARTICLE 19); Prof. CHAUDHURI, Bidisha (IIIT Bangalore); Prof. JHA, Shishir (CPS, IIT Bombay)

Contribution ID: **98** Type: **Individual Paper**

Representation and Machine Agency

Theories of action tend to require agents to have mental representations. A common trope in discussions of Artificial Intelligence (AI) is that they don't, and so cannot be agents. Properly understood there may be something to the requirement, but the trope is badly misguided. Here we provide an account of representation for AI that is sufficient to underwrite attributions to these systems of ownership, action, and responsibility. Existing accounts of mental representation tend to be too demanding and unparsimonious. We offer instead a minimalist account of representation that ascribes only those features necessary for explaining action, trimming the "extra"features in existing accounts (e.g., representation as a "mental"phenomenon). Our account makes 'representation'whatever it is that, for example, the thermostat is doing with the thermometer. The thermostat is disposed to act as long as the thermometer is outside a given range of parameters. Our account allows us to appraise decision-making machines, with respect to their moral agency, and so address the 'responsibility gap', a new type of problem raised by the actions of sophisticated machines: because nobody has enough control over the machine's actions to be able to assume responsibility, conventional approaches to responsibility ascription are inappropriate.

Keywords

Representation, Mental Mental Representations, Agency, Responsibility Gap, Machine Learning

Primary authors: CIBRALIC, Beba (Georgetown University); Dr MATTINGLY, James (Georgetown University)

Presenters: CIBRALIC, Beba (Georgetown University); Dr MATTINGLY, James (Georgetown University)

Session Classification: Synthetic Minds and Consciousness

Contribution ID: **100** Type: **Individual Paper**

Luke, I'm NOT your father: beyond technological paternalism, towards mutual cooperation between patients, medical staff and AI

This paper explores how the introduction of AI systems in medical practice alters doctor-patient therapeutic relationships. Our claim is that although current medical AI systems increase effectiveness in diagnostic and treatment, they have the potential to introduce another layer of paternalism in doctor-patient relations, which we call technological paternalism. Consequently, we argue that medical AI should be designed so as to foster patient-centered therapeutic relationships, which have been proven to increase adherence to treatment, maximize autonomy for both doctors and patients and decrease the general discomfort of medical services and treatments.

Keywords

therapeutic relations, doctors, patients, artificial intelligence, technological paternalism

Primary authors: Dr VOINEA, Cristina (Bucharest Univeristy of Economic STudies, Department of Philosophy and Social Sciences); Dr VICA, Constantin (University of Bucharest, Department of Philosophy); Dr DRAGOMIR, Alexandru (University of Bucharest, Department og Philosophy)

Session Classification: AI and Medical Practices I

Contribution ID: **103** Type: **Individual Paper**

What to Make of Functions in Computer Science? A Case Study on the Basis of Computer Programs

This study aims to fill a gap in the literature about the role of "functions"in computing, particularly focusing on the case of computer programs. For exemplifying the problem, I focus on recent work, Piccinini's 2015 Physical Computation: A mechanistic account and Turner's 2018 Computational Artifacts, and compare them. Based on that (i) I argue that the notions of functions employed in the domains of physical computation and artifacts, subsequently come together in the case of computer programs. However, (ii) since momentarily there is no consensus on the nature of functions, the integration of different kinds of functions potentially also spells trouble for our understanding of computer programs. So, (iii) does it turn out that the notion of function in one of the previous accounts has to be revised (revision of existing accounts)? Is it possible to unify the different notions of functions (unification of existing account)? Do we have to take a pluralistic stance (co-existence; pluralism)? In case of computer programs, I aim to show, how the notion of a stratified ontology of computer programs might be able to resolve these worries, accommodating the different kind of functions "on different levels".

Keywords

Computer Programs, Functions, Computational Artifacts

Primary author: WIGGERSHAUS, Nick

Session Classification: Philosophy of Computing and Machine Learning

Contribution ID: **104** Type: **Individual Paper**

Governance of AI Ethics

Wednesday, 7 July 2021 17:15 (30 minutes)

The technical and economic benefits of artificial intelligence (AI) are counterbalanced by legal, social and ethical issues. It is challenging to conceptually capture and empirically measure both benefits and downsides. Furthermore, while many groups have published AI ethics principles, the wider discussion in AI ethics is towards operationalisation and governance. We provide an account of the findings and implications of a multi-dimensional study of AI undertaken by the SHERPA Project, comprising 10 case studies, five scenarios, an ethical impact analysis of AI, a human rights analysis of AI and a technical analysis of known and potential threats and vulnerabilities. Based on our findings, we analyse governance and mitigation measures take by organisations currently using AI-assisted systems.

Keywords

AI, ethics, governance, privacy, code of conduct

Primary authors: MACNISH, Kevin; Prof. STAHL, Bernd; Dr RYAN, Mark; Dr ANTONIOU, Josephina; Dr JIYA, Tilimbe

Session Classification: Governance of AI

Contribution ID: **105** Type: **Individual Paper**

AI Personhood Mapped Via a Continuum

It is an important and pressing question whether personhood ought to be adopted as a legal status for artificial intelligence systems (AI systems). I argue that conceptualizing artificial intelligence agency on a continuum rather than in binary terms allows the implications of the adoption of legal personhood status for artificial agents to be more appropriately framed.

This paper is thus motivated first as an extension of the recent conceptual work on AI systems begun by scholars. It is additionally a response to recent European Parliament commission on AI ethics and related scholarly work.

This paper is laid out as follows. I introduce the definitions of what is meant by intentional agency (agency) in 1.1. In Section 2 I introduce these essential aspects of agency as a conceptual framework for how to think about artificial agents. In Section 3 I provide an example of how various current AI systems might be mapped to this AI continuum. I respond to counterarguments in Section 4 and conclude in Section 5.

Keywords

AI, Artificial Intelligence, Personal Identity, Political Philosophy, Law

Primary author: REIMER, Karl (University of Zurich)

Session Classification: Artificial Moral Agents

Contribution ID: **107** Type: **Individual Paper**

Is there a civic duty to support biomedical AI development by sharing personal health data?

With AI and big data on the rise, there is growing hope that medical research will be able to take huge innovative steps towards a healthier future. To do this, however, it is essential that AI systems have access to personal medical and non-medical Big Data. Most theorists recognize an ethical problem in the trade-off between the individual right to privacy and the potential social benefits of innovative medical AI. I belief that the ethical problem in this matter cannot be described sufficiently by reference to this trade-off scenario. Instead I argue that human individuals must have sovereignty over their digital personality in order to maintain their status as morally autonomous subjects. The same applies to the moral autonomy of individuals, who reciprocally understand themselves as authors and addressees of the same social structures. In my conclusion, there can only be a civic duty to interact with biomedical AI systems, if those interactions are actually demanded by the citizens.

Keywords

AI, personal health data, civic duty, autonomy

Primary author: MÜLLER, Sebastian

Session Classification: Data, AI and Responsibility II

Contribution ID: **109** Type: **Individual Paper**

An Ethics by Design Approach for Artificial Intelligence

Short Abstract
In this paper, we will present an Ethics by Design approach for AI. Ethics by design approaches include ethical considerations in a systematic way in design processes by allowing ethical guidelines to be translated into concrete design practices. Our approach can be described in a five-layer model, going from abstract to concrete. It starts with (1) a total of six Ethics by Design values, which are translated into (2) twenty-eight AI-specific ethics requirements, which are then translated into (3) eighty-five Ethics by Design guidelines which map onto the different phases of a generic model of the design process for AI. Designers will need to map this generic model onto their preferred AI development methodology (4) before being able to apply the ethics guidelines to different steps in the design process. Finally, Ethics by Design employs specialized tools and methods (5) to help designers with the mapping and application process. We argue that our Ethics by Design approach enables developers to engage with ethical considerations in a proactive way, instead of the reactive approach offered by research ethics and ethics assessment. It moreover builds on existing ethics guidelines, including those of the EU High-Level Expert Group on AI, IEEE and OECD.

For Long Abstract, see the attached file.

Keywords

Ethics by Design; Design for Values; Ethics guidelines

Primary authors: BREY, Philip (University of Twente); Dr DAINOW, Brandt (Ayni Ltd)

Session Classification: AI, Values, and Design

Contribution ID: **110** Type: **Individual Paper**

Conceptualising Solidarity in the Context of AI

While several scholars have proposed that solidarity should be considered as an ethical principle for Artificial Intelligence (AI), it is far from clear what solidarity would actually mean when applying it to AI. This paper aims to fill the gap and explores the concept of solidarity in the context of Artificial Intelligence. Based on a conceptual analysis, it argued in this paper that solidarity can take different forms, each operating at different levels –micro, meso, and macro level (as similarly argued by Prainsack & Buyx, 2012), and consequently, each also transcending into varying forms of social relations, collective goals and duties. In line with Axel Honneth and Luengo-Oroz (2019) and by taking a more collectivist perspective, this paper claims that solidarity bust be based on mutual recognition and the normative overall goal to equally share the benefits and risks of AI technologies. Furthermore, this paper lays out in how far solidarity does or can potentially work at the three levels of analysis and where it might lead to.

Keywords

AI Ethics, Ethics of AI, Artifical Intelligence, Solidarity

Primary author: RUDSCHIES, Catharina (Universität Hamburg)

Contribution ID: **111** Type: **Individual Paper**

Assessing the Ethical Risks of AI

Despite the surge in research on ethical risks of artificial intelligence (AI) of the past few years, there is still a clear need for methodologies and practical strategies to assess ethical risks of AI applications. Especially identifying normative issues related to voluntariness, justice and power imbalances remains a challenge. The current article proposes a supplementary assessment method to address these ethical issues of AI, by building on the three-party model for ethical risk analysis. In this relational model, developed by Hermansson and Hansson (2007), ethical risks are evaluated from the relationship between three critical parties or roles present in all risk related decisions: the decision-maker, the risk-exposed and beneficiary. Adjustments to the three-party model are advanced to enhance its operability in AI contexts, firstly by including first and second order roles and secondly by integrating explicability as an important feature of the AI ethics field. A credit risk scoring model from the financial industry is used as an example to indicate how the modified three-party model can be used to identify salient ethical features.

Keywords

Ethical Risk Assessment, Philosophy of Risk, AI Ethics, Risk-based Regulation

Primary author: KRIJGER, Joris

Session Classification: Risks of AI

Contribution ID: **112** Type: **Individual Paper**

Extended and autonomous agency - capacity, control and coherence

By employing AI systems, we are increasingly able to enhance our capacity as agents in many different areas, including information processing and goal-oriented behavior. Concerns with the impact of this development on individuals have focused on possible deskilling and transfer of control from human being to technology, leaving individuals less autonomous. Such concerns, however, presume that the human agent is distinct from the AI systems she uses. As shown by the extended mind hypothesis, it is not obvious how to draw the line between a human being and the tools she employs. We may consider an extended agent to consist of both a human user and the tools she employs for enhancing her agency, including AI systems. While this perspective is in one way pro-technological, it also brings forth three new concerns with individual autonomy. First, enhancement of one ability need not improve an agents overall capacity to perform well in all areas of life. Second, extended agents may be more dependent on external maintenance, and so may be less in control of their own future. Third, to the extent that extended agents consists of partly distinct deliberative systems, the extended agent may be less internally coherent.

Keywords

Extended agency, extended mind, autonomy, control, reliability

Primary author: GRILL, Kalle (Umeå University)

Session Classification: Nature and Ethics of Autonomous Systems

Contribution ID: **114** Type: **Individual Paper**

Ethics of AI

This paper argues that technical rationality is never just calculative and is always embedded in self-interpretive practices with competing goods, such that technical decisions are more authentically understood as hermeneutic interpretations of practical predicaments. The critical role of technology ethics is to unpack and account for the full ethical content of the decisions faced by the designer and engineer, including the competing internal moral visions of the engineer at play in these decisions. Such an account would restore the sense of ethical agency that the notion of technology as applied science espoused by engineering educators conceals from engineers.

Contemporary technology, including the development of probability and statistics up to the present state of AI, is oriented by one of two self-interpretations, two goods internal to technology –the self as responsible to one's experience, rather than being bound by external authority, and the self as applier of exact and universally valid science, rather than superstition or subjective prejudice. These two goods exist in tension throughout the history of probability and statistics, and animate competing directions of the future of AI.

Calls for ethical guidance of technology and AI from the outside reinforce the narrow agency afforded engineers and statisticians. Nothing less than an appeal to competing internal goods within technical reasoning itself, and a concomitant reframing of technical knowledge and training, will change the current trajectory of AI.

Keywords

Primary author: ARCHER, Ken

Session Classification: Data, AI and Responsibility II

Contribution ID: **116** Type: **Individual Paper**

When Does Enhancement Become Restriction? The Case for Contestable Intelligent Space

Artificial intelligence has transformed the built environment. The function of space and how it is navigated has adapted with predictive technology. One can now be "lost"when their phone dies, but "know where they are"upon seeing their GPS coordinates on a map while stranded in an unknown locale. Moreover, physical space "understands"when to light streets and alley ways, heat or cool a building, or let drivers cross the intersection. Unfortunately, only when predictions go wrong does the restrictive and paternalistic nature of artificially intelligent space become uncomfortably apparent. Designing for contestability might pave a way to harness the enhancement of algorithmic decision making without disposing of human agency.

Keywords

Primary author: CAMMERS-GOODWIN, Sage

Session Classification: Human Machine Relations I

Contribution ID: **117** Type: **Individual Paper**

Autonomous Weapon Systems and the Claim-Rights of Innocents on the Battlefield

Much of the debate about autonomous weapon systems is "pre-implementational."As such, questions like "can AWS be in compliance with the regulations of the Geneva Conventions?"are often the focus. As such, the debate centers on whether the system itself would be able to comply with current or future possible IHL. What this debate does not address is, if AWS can conform to IHL and the JWT, how does this change the duties, rights, and responsibilities inherent on the battlefield? To do this, we must shift the debate from pre-implementation to post-implementation. In support of this, I focus on the claim-rights of innocents against combatants to use AWS in order to reduce instances of unjust killing and collateral damage.
Leif Wenar, in "The Nature of Claim-Rights,"argues that the most philosophically robust form of the claim-right is the "Kind-Desire"as opposed to Will and Interest claim-rights. Wenar defines Kind-Desire claim-rights as, "some systems of norms refers to entities under descriptions that are kinds ("parent,""journalist,""human,"etc.). Within such a system, claim-rights correspond to those enforceable strict duties that the members of the relevant kind want to be fulfilled."Using Wenar's concept, I argue that , if and only if autonomous weapon systems reach a point in their development where they are capable of being both more discriminating in their target selection and more proportionate in their response to threats, then innocents on the battlefield (as a kind) have a claim-right to not be unjustly harmed against both sets of combatants, claiming that AWS be used in place of human soldiers. This claim is a fulfillment of the "strict duties that the members of the relevant kind want fulfilled,"namely to not be killed, injured, nor have their property destroyed unjustly. This application of Wenar's Kind-Desire Theory of claim rights also accords with Convention IV Article 27, "Protected persons are entitled···to respect for their persons···They shall at all times be humanely treated, and shall be protected especially against all acts of violence or threats thereof···"(ICRC 1949). I conclude that with the implementation of sophisticated AWS, innocents will have a claim-right on their use by combatants and combatants have a duty to use AWS in fulfillment of this claim-right.

Keywords

Primary author: CANTRELL, Hunter

Session Classification: Ethics and Automated Warfare

Contribution ID: **118** Type: **Individual Paper**

A Perfect Storm for Epistemic Injustice: Algorithmic Targeting and Sorting on Social Media

Over the past decade, feminist philosophers have gone a long way toward identifying and explaining the phenomenon that has come to be known as epistemic injustice. Epistemic injustice is injustice occurring within the domain of knowledge (e.g., knowledge production and transmission), which impacts structurally marginalized social groups. In this paper, we argue that, as they currently work, algorithms on social media exacerbate the problem of epistemic injustice. In other words, we argue that algorithms on social media recreate and reify the conditions that lead to some groups being systematically denied the full status of knowers, thereby corrupting the epistemic terrain, and with it, systems of social trust and cooperation. We argue that algorithms do this in two ways, namely, what we are calling algorithmic targeting and algorithmic sorting.

Keywords

Epistemic injustice, social media, algorithmic sorting

Primary authors: STEWART, Heather (Western University); CICHOCKI, Emily (Western University); Prof. MCLEOD, Carolyn (Western University)

Session Classification: AI, Data, and the Pandemic

Contribution ID: **119** Type: **Individual Paper**

What Confucian Ethics Can Teach Us About Designing Caregiving Robots for Geriatric Patients

Caregiving robots are often lauded for their potential to assist with geriatric care. While it is tempting to treat seniors as wise and mature, possessing valuable life experience, they also present a variety of ethical challenges, from prevalence of behaviors associated with racism and sexism, to troubled relationships, histories of abusive behavior, and more recent onset aggression, mood swings and impulsive behavior associated with cognitive decline. I draw on Confucian ethics, especially the concept of filial piety, to address these issues. While filial piety is sometimes thought to require children's unquestioning obedience, and a willingness to sacrifice their own needs and interests to support their parents, this is a misunderstanding. Confucian scholars have developed a rich set of theoretical resources for dealing with beloved but imperfect elders, and navigating the challenges of supporting seniors whose ethical commitments are unreliable. These resources provide a way to reconcile two important but conflicting desiderata: to value and care for seniors, but also to clear-mindedly deal with their moral shortcomings. In particular, they articulate a duty to remonstrate with our elders when they err, a duty that can helpfully inform technology design and use in geriatric care.

Keywords

Primary author: ELDER, Alexis (University of Minnesota Duluth)

Session Classification: AI and Medical Practices II

Contribution ID: **121** Type: **Individual Paper**

Trust in IoT Systems in Among Users with Disabilities: A Framework for Inclusive and Trustworthy Development

This session will present results from an interdisciplinary mixed-methods research project on trust and access in the use of Internet of Things (IoT) devices by disabled persons. The year-long project is directed by professors in Philosophy, Education, and Modeling and Simulation, and will produce a trust framework as well as a prototype app. This presentation will focus on research outcomes for computer ethics: conclusions from our study of how trust in IoT is generated or eroded for users with disabilities that tell us about how systems can be built to be both more trusted and more worthy of trust. We will outline concerns having to do with justice, security, and privacy as articulated through our theoretical research based in feminist ethics of care, postphenomenology, and disability studies, and our empirical research based in universal design and universal design learning.

Keywords

Primary authors: WITTKOWER, D.E. (Old Dominion University); BLACKMON, Stephanie (W&M School of Education); RECHOWICZ, Krzysztof (Old Dominion University); HERDEGEN, Hanna (Virginia Tech)

Session Classification: Human Life and Trust

Contribution ID: **122** Type: **Individual Paper**

Machine Learning as Model & as Metaphor

Over human intellectual history, philosophers have made progress on under- standing natural systems by analogising them with technological advances. New technologies also serve as formal, physical, and conceptual models of target sys- tems, and their development is informed by empirical discoveries in the sciences. Emerging technologies in machine learning promise a radical disruption of scien- tic practice. With the increasing prevalence and in uence of machine learning comes an imperative to understand how these techniques come into play in an empirical context. Machine learning poses a unique puzzle in this respect, as the history of innovation in this domain is intrinsically tied to the history of discovery in the cognitive and neurosciences. Thus the relation between the empirical study of mind and brain and the development of techniques in artic- ial intelligence is one of bidirectional in uence. In this essay, I map out this nascent area of study for the philosophy of science, making several key distinc- tions by way of a relevant case study: the concept of the brain as a predictive engine, as exemplied by a recent mathematical framework known as the free energy principle. I delineate three ways of interpreting the notion of brain and body as engaging in Bayesian inference: (approximate) realism, metaphor, and model or research heuristic.

Keywords

Primary author: ANDREWS, Mel (University of Cincinnati)

Session Classification: Philosophy of Computing and Machine Learning

Contribution ID: **123** Type: **Individual Paper**

A Bayesian approach to trust: metacognition, confidence and autonomy

We take 'trust'as a belief of a human H (the 'trustor') in an agent A (or the 'trustee') when H expects that A 'will take care of the things'(A. Baier, K. Jones, K. Hawley, A. Carter). When is H's trust in A rational? How much does H need to know about A in order to entrust it? We argue here that H needs to know something about A's internal nature and can conditionalize its trust by imposing some requirements on A. The present formal approach to the epistemology of trust is based on: (i) metacognitive requirements inspired by recent studies in cognitive neuroscience (F. Meyniel, S. Fleming, N. Daw) and (ii) a number of Bayesian requirements on A's process of confidence assessment (B. Timmermans). It is part of the argument to show that a genuine trust of H implies a number of requirements on A's internal structure (cognitive and metacognitive), rather than on A's behavior or actions. We are interested in differentiating cognitive features of A (e.g. the accuracy of its representation of the world) and a number of metacognitive features of A. The trustworthy A is the one that instantiates cognitive processes of mode the world (the cognitive component) and, independently, processes of assessing the confidence level in this model (metacognitive component). We discuss two formal definitions o f c onfidence as an independent factor from accuracy and we define trust as a conditionalization on confidence levels. We end our analysis by assuming that H can trust A, among other things, when A implements a metacognitive computational process independent of the cognitive process. This analysis complements and auguments A. Carter's (2019) recent bi-level approach to trust with a naturalized epistemology (inspired by metacognitive studies) and a Bayesian formalism.

Keywords

Primary author: MUNTEAN, Ioan (University of North Carolina, Asheville & UT Rio Grande Valley)

Session Classification: Human Life and Trust

Contribution ID: **124** Type: **Individual Paper**

Human Rights to Health vs. Human Rights to Privacy in the Covid-19 Pandemic: A Consequential Evaluation in the Big Data Era

In the era of globalization, many health issues are not confined by national boundaries, but also problems for the whole world, such as the Covid-19 pandemic. One of the most important questions in global health ethics is about the ethics of human rights to health. Some people question that human rights to health have no correlative perfect obligation and these rights are impossible to be satisfied in the current conditions of the world. In addition to these questions, another controversy is the conflicts between human rights to health with other human rights, such as human rights to privacy. In this big data era, when the information of our health may be found on the internet and revealed to others without our consent, government and other organizations may use our information for many good or bad purposes. How to handle and balance the conflict between human rights to health and human rights to privacy is a big topic that everyone should think about. And it is important for philosophers to provide an analytical and moral framework to solve such a conflict.

My research project seeks to investigate such a framework for the ethics of human rights to health and privacy in the big data era. Particularly, this project aims at seeking out a non-utilitarian consequential evaluation for these human rights by evaluating and comparing ideas from Allen Buchanan, Amartya Sen, and William Talbott. This account of consequential evaluation of human rights is not only theoretical important but also have practical implications. In the wake of the Covid-19 pandemic, everyone's health is being threatened by the virus, and the global impact of Covid-19 has been profound. In this paper, a practical issue of public health in the Covid-19 pandemic will be investigated. It is about the conflicts between individual freedom and public goods in lockdown policies. It is arguably that some policies against the present pandemic conflict with individual freedom and human rights. Some of these controversies around Covid-19 involve new technologies such as tracing device and immunity apps, and also health code, immunity passport, and other kinds of collecting health information that may have issues about the ethics of privacy. New technologies bring new improvement and innovative changes to traditional methods, but they also bring new challenges to the core ethical concepts such as privacy, consent of data owner and responsibility. I argue in this presentation that the consequential evaluation of human rights is a key to solve these issues. Although one presentation cannot answer all questions, at least some preliminary but important philosophical investigation will be addressed in the presentation.

Keywords

Primary author: CHAN, Benedict S. B. (Hong Kong Baptist University)

Session Classification: AI, Data, and the Pandemic

Contribution ID: **125** Type: **Individual Paper**

Digital Surveillance in a Pandemic-Response: what bioethics needs to learn from Indigenous perspectives

Our paper interrogates the ethics of digital pandemic surveillance from Indigenous perspectives. The Covid-19 pandemic has showed that Indigenous people are among communities most negatively affected by a pandemic infectious disease spread. During the pandemic, like other racialised subpopulations, Indigenous peoples have faced strikingly high mortality rates owing to structural marginalisation and comorbidities, yet, intensified by the past and present colonial dominance that Indigenous subpopulations have been subject to. At the same time, digital technologies that have been promoted as effective tools to suppress pandemic infectious spread also carry disproportionately negative implications for Indigenous subpopulations. Critical race scholars have warned that Indigenous people are most vulnerable regarding digital surveillance and some bioethicists have argued that certain types of pandemic surveillance tools will lead to the employment of disproportionate profiling, policing and criminalisation of marginalised population subgroups. Building on this work, our paper investigates whether and under which conditions could digital pandemic surveillance tools help protect Indigenous lives, without exacerbating the structural vulnerabilities of Indigenous people. As part of the investigation, we identify lessons that bioethics debates ought to learn from and incorporate so that the guidance regarding health interventions and policy generated through them will be relevant to and benefit Indigenous subpopulations.

Keywords

Primary authors: Dr HENDL, Tereza (Ludwig-Maximilians-University in Munich); Dr ROXANNE, Tiara (DeZIM Institut Berlin)

Session Classification: AI, Data, and the Pandemic

Contribution ID: **126** Type: **Individual Paper**

Rule-Based Robots: Why Autonomous Machines Can and Must be Governed by Rules of Right

Some machine ethicists claim that any rule-based approach to implementing morality in autonomous machine agents must fail because rules are either too vague or, alternatively, too brittle and conflicted to determine what action the agent should take in particular cases. In this paper, I argue that autonomous machine agents not only can be governed by explicit moral rules, but must be governed by such rules, if they are to act morally. The most important moral rules that should govern moral machines specify what Immanuel Kant calls "duties of right"("legal"duties), which are rightfully enforceable duties to which everyone can consent. I argue that all such duties of right must be precisely specified to remove ambiguity or conflict; otherwise, universal consent to them is impossible to secure. I thus provide a novel partial solution to a longstanding problem of judgment in Kantian theory that is a practically complete solution for moral machine agents. I then set forth a technical corollary: Any "bottom-up,"pure machine learning approach to implementing morality in autonomous machine agents must fail because such an approach cannot meet the key technical requirement that the agent's behavior can be formally verified to respect relevant legal and safety regulations.

Keywords

Primary author: THOMAS WRIGHT, Ava (California Polytechnic State University San Luis Obispo)

Session Classification: Nature and Ethics of Autonomous Systems

Contribution ID: **131** Type: **Individual Paper**

Orthogonality and Existential Risk from AI: Can We Have it both Ways?

There is a hole in the standard argument to the conclusion that AI constitutes an existential risk for the human species –even if its two premises are true: (1) AI may reach superintelligent levels, at which point we humans lose control (the 'singularity claim'); (2) Any level of intelligence can go along with any goal (the 'orthogonality thesis'). We find that the singularity claim uses a notion of 'general intelligence', while the orthogonality thesis uses a notion of 'instrumental intelligence' . If that is true, they cannot be joined to support the conclusion that AI constitutes an existential risk. To repair the situation, we try to find a unified notion of intelligence that can be used in both premises, but we fail.

Keywords

Primary author: MÜLLER, Vincent C.

Session Classification: Risks of AI

Contribution ID: **132** Type: **Individual Paper**

Medical Interfaces with Emotion AI: Shaping Public Narratives and Perceptions of Nonverbal Patients with Degenerative Diseases

Affect recognition, a subset of emotion artificial intelligence that is sometimes more aptly called human perception technology, has been gaining increasing media attention in recent years. Although news coverage has focused on education, hiring, and law enforcement contexts, researchers have also spent the past two decades investigating how this technology might improve health care and clinical research. My argument focuses on a particular use-case of human perception technology and its implications for epistemic injustice: assessing pain and needs in nonverbal patients, an application that has been tested in persons with late-stage dementia.

Late-stage diseases are fraught with hermeneutical injustice by virtue of the narrative deficiency surrounding those experiences in the public sphere. I weigh the complicated implications for epistemic injustice against two relevant criteria in considering the ethics of automated pain assessment: (1) the prospect of improving the lives of persons unable to communicate remediable discomfort to caregivers, and (2) the prolific, prejudiced history of pain assessment. I conclude by assessing how the history of pain assessment might shape an understanding of hermeneutical injustice that takes seriously the role of narrative foreclosure in shaping the perceptions, valuation, and epistemic status of nonverbal patients at the end of life.

Keywords

Emotion AI; Pain assessment; Epistemic injustice

Primary author: BOLO, Isabel

Session Classification: Issues of Facial and Emotion Analysis

Contribution ID: **133** Type: **Individual Paper**

An exploration on the emergence of Machine Consciousness, and the Risk of Robocentrism

This paper presents a deep dive into the possible emergence of machine consciousness, and subsequent development of Robocentric ethical and moral values.

Drawing on biologist Mario Vaneechoutte's (2000) evolutionary approach to experience, awareness, and consciousness, as well as aspects of Braitenberg's Vehicles (1984), this paper demonstrates that without programming specific behaviours, machine consciousness can emerge simply by the nature of evolutionary processes and a human need for general purpose AI.

As machines become more advanced and capable of autonomous self-learning and improvement, it is reasonable to assume that they will develop their own sense of ethical and moral standards in relation to their surroundings. The final section discusses the implications of such developments for future society, ultimately opening the debate for further discussion and comment from conference attendees.

Keywords

Primary author: DAVIS, Matthew

Session Classification: Synthetic Minds and Consciousness

Contribution ID: **134** Type: **Individual Paper**

Artificial artificial intelligence - The hidden microwork in AI

Hidden behind the promises of "artificial" intelligence or "machine" learning is a large workforce of microworkers managed by internet platforms doing the necessary data work that powers the AI industry. Human labour is not only needed to train the algorithmic models but also to verify them. In some cases, humans even replace the algorithms, creating what has been called "fauxtomation". While microwork is essential for developing AI, it often goes unrecognised in the narratives around it that frequently focus on the technological aspects. One the one hand full-time employees at tech companies who celebrate their liberalism and egalitarianism receive high wages and a lot of benefit, microworkers on the other hand do not even earn minimum wage nor receive any benefits. The following paper analyzes microwork as an instance of 'invisible labour'a concept originally stemming from feminist scholarship on unpaid housework. As I argue, microwork is made invisible through a variety of strategies such as geographical distancing, the design of the platforms and particular legal classifications. The paper argues that the implications of the precariousness of microwork should be taken into account in the discussions around fairness in machine learning, the introduction of potential biases into data sets, and the future of work and automation.

Keywords

Primary author: BILSING, Charlotte

Session Classification: Digital Life and Ethics

Contribution ID: **135** Type: **Individual Paper**

Philosophical Foundations of AI Regulation

Questions of regulation are intricately tied to questions of autonomy and personhood. In this paper, while exploring the philosophical foundations of AI regulation, I argue that if Artificial Intelligence (AI) is considered to be morally autonomous then the search for regulation would lead us towards the existing regulation of natural persons and formulation of 'reasonable algorithm' standard. However, if AI is considered to be only computationally autonomous, then it's regulation would be based on existing standards of legal person regulation arising out of group autonomy. This dichotomy would have bearing on liability standards as well, which may vary between joint and several. From an ethical perspective, AI regulation would further need to account for the distinction between embodied and disembodied AI.

Keywords

Primary author: PURI, Anuj (University of St. Andrews)

Presenter: PURI, Anuj (University of St. Andrews)

Session Classification: AI and Regulation

Contribution ID: **136** Type: **Individual Paper**

The Deontic Logic of East-Asian Moral Cognition, for Robots

In a sizable body of work, Nisbett et al. claim there is a fundamental difference between Occidental versus East-Asian human reasoning. Encapsulated, the idea is that in the former case reasoning is often highly sensitive to classical inconsistency (which revolves around contradictions of the shape *P* & not-*P*), while in the latter case things are —and here we quote —"dialectical."Now, taking note of the fact that, over the past two decades, more than a few researchers —in what is variously e.g. called 'machine ethics'or 'robot ethics'—have toiled toward the production of machines/robots that are *themselves* ethically correct (or at least competent), we ask the following question: What would it take to build such a robot that is ethically correct within the East-Asian dialectical paradigm? Under the assumption that the ethical theory to be used in such building in Confucian in nature, we answer this question, after first showing how the Nisbettian East-West dichotomy is dissolved by way of a novel kind of logic-based adjudication that captures dialectical reasoning (including the Nisbettian/East-Asian variety) through time. Computational simulations that support and concretize our answer are provided.

Keywords

Primary authors: Prof. BRINGSJORD, Selmer (Rensselaer Polytechnic Institute); SUNDAR G, Naveen (Rensselaer Polytechnic Institute); GIANCOLA, Michael (Rensselaer Polytechnic Institute)

Session Classification: Nature and Ethics of Autonomous Systems

Contribution ID: **137** Type: **Symposium/Panel Proposal**

Dissecting digital complexity to inform policy choices in the shaping of COVID-19 digital infrastructure

Many countries worldwide turned to digital approaches for fighting the pandemic. These approaches were hailed as smart solutions for contact tracing, infection risk warning, even quarantine surveillance. While often simple on the outside, many of these digital approaches are embedded in complex socio-technical ecosystems. These ecosystems reflect critical design choices for digital infrastructures in our society.

In this panel, we explore how a dialog between informatics, law, and philosophy can help to dissect the complexity of these infrastructures to uncover the policy choices made explicitly or implicitly in designing these infrastructures. We debate if and how such knowledge can help us to better understand the interplay between such design choices and normative reasoning about our digital future.

Keywords

Primary authors: BÖHMANN, Tilo (Universität Hamburg); KOREN-ELKIN, Niva (Tel Aviv University); RIEDER, Gernot (Universität Hamburg)