Workshop

Machines of Change: Robots, AI and Value Change

1 – 3 February, 2022

Advances in Artificial Intelligence and robotics stand to change many aspects of our lives, including our values. If trends continue as expected, many industries will undergo automation in the near future, calling into question whether we can still value the sense of identity and security our occupations once (ideally) provided us. Likewise, the advent of social robots appears to be shifting the meaning of numerous, long-standing values associated with interpersonal relationships, like care, friendship and privacy. Furthermore, powerful actors' and institutions' increasing reliance on AI to make decisions that affect how people live their lives, has given rise to the development of new values such as algorithmic transparency, meaningful human control and explainability.

During the *Machines of Change: Robots, AI and Value Change* workshop, we will explore how the deployment of Artificial Intelligence and robots leads to value change and how we can study value change, as a phenomenon, via these technologies. The workshop will centre around the following three, major themes:

1. How do AI and /or robotics contribute to value change?
2. How can we study value change via AI and / or robotics?
3. How should AI and / or robotics deal with value change?

Program

Tuesday, February 1, 2022
Theme: **How do AI and robotics contribute to value change?**

Time	Session and speakers
10:00 – 10:15	Introduction by Prof. Dr. Ibo van de Poel
10:15 – 11:15	**Cecilie Eriksen** "AI and Fundamental Value Change in Public Administration and the Welfare State" Commentary by J. *Hopster; M. Maas*
11:15 – 12:15	**Jeroen Hopster; Matthijs Maas** "Triaging the Technology Triad: Disruptive AI, Regulatory Gaps and Value Change" Commentary by *Cecilie Eriksen*
12:15 – 13:00	*Lunch break*
13:00 – 14:00	Keynote speech by **John Danaher**
14:00 – 14:15	*Break*
14:15 – 15:15	**Zachary Daus** "How the unhealthy division of labor in AI calcifies value change: A Durkheimian analysis" Commentary by *Filippo Santoni de Sio*
15:15 – 16:15	**Seppe Segers** "Care robots and the value of veracity: Disruption or entrenchment?" Commentary by *Mehrdad Rahsepar Meadi*
16:15 – 17:15	Closing remarks and informal discussion

Wednesday, February 2, 2022
Theme: **How can we study value change with the help of AI and robotics?**

Time	Session and speakers
10:00 – 10:15	Introduction to Day 2
10:15 – 11:15	**Henk J. van Gils-Schmidt; J.-C. Põder, A. Räder; J. Wegner** "A mixed-method approach to investigate value change by technological innovations" Commentary by *Edmund Terem*
11:15 – 12:15	**Deivide Garcia da S. Oliveira** "AI's invisibility: Distrusting AI's choices" Commentary by *Vipra Chopra*
12:15 – 13:00	*Lunch break*
13:00 – 14:00	**Michał Sikorski** "Non-epistemic values and automation of science" Commentary by *M. van Lier; A. López Incera; S. Styger*
14:00 – 15:00	**Maud van Lier; Andrea López Incera; Sahra Styger** "AI-Scientists and their Impact on Science" Commentary by *Michał Sikorski*
15:00 – 15:15	*Break*
15:15 – 16:15	Keynote speech by **Arianna Betti**
16:15 – 16:30	Closing remarks and informal discussion

Thursday, February 3, 2022
Theme: **How should AI and robotics deal with value change?**

Time	Session and speakers
10:00 – 10:15	Introduction to Day 3
10:15 – 11:15	**Vipra Chopra** "Value Change/Transformation in Artificial Agents: Understanding the nature of Artificial Values" Commentary by *Deivide Garcia da S. Oliveira*
11:15 – 12:15	**Filippo Santoni de Sio** "Artificial Intelligence and 'broad meaningful human control'" Commentary by *Zachary Daus*
12:15 – 13:00	*Lunch break*
13:00 – 14:00	**Edmund Terem Ugar** "Analysing technological colonialism in Sub-Saharan Africa: A case for many-value AI" Commentary by H. *van Gils-Schmidt; J.-C. Põder, A. Räder; J. Wegner*
14:00 – 15:00	**Mehrdad Rahsepar Meadi** "When will the robot therapist take over? Ethical reflection on how artificial intelligent psychotherapy should take shape" Commentary by *Seppe Segers*
15:00 – 15:15	*Break*
15:15 – 16:15	Keynote speech by **Bertram Malle**
16:15 – 16:30	Closing remarks and informal discussion

Keynotes

The Normative Significance of Value Change
John Danaher (Senior Lecturer of Law, Galway University)

Since moral revolutions have occurred in the past, it seems plausible to suppose that they will occur again in the future. But what is the normative significance of this possibility? Does the fact that significant moral disruption may occur in the future have any normative implications for us right now, or could it have some normative significance once we enter into a period of moral disruption? This paper attempts to answer these questions. It does so by identifying, describing, and evaluating eight potential normative implications of future moral revolutions. It also considers the role that judgments of certainty and uncertainty play in understanding those normative implications, and the potential for metaethical commitments to modulate their impact.

To be announced
Arianna Betti (Professor (Chair) of Philosophy of Language, the University of Amsterdam)

Norm for humans and robots
Bertram F. Malle (Professor, the Department of Cognitive, Linguistic, and Psychological Sciences, Brown University)

Robots of the very near future will occupy roles in human communities that require them to represent, learn, and conform to social and moral norms. I introduce a theoretical model that identifies a number of properties of human norms, and I present empirical support for this model. I then discuss how these kinds of properties could be implemented in robots. I introduce a human-machine interaction paradigm that enables research and safe implementation of human norm teaching and machine norm learning, including continuous trust assessments. I present initial results from human experiments and algorithm development of symbiotic norm teaching-learning. Finally, I identify challenges that any norm-competent agent (human or machine) faces, and that designers must grapple with as well: the context and community specificity of norms, their graded strength, and their dynamic, sometimes rapid change.

Abstracts

Day 1: How do AI and robotics contribute to value change?

AI and Fundamental Value Change in Public Administration and the Welfare State
Cecilie Eriksen

Technology has been shown to be able to be a significant factor in the co-creation of moral changes (Morris 2013; Swierstra 2013). However, the literature on moral revolutions has so far not covered this aspect in any depths (Appiah 2010; Pleasants 2018; Baker 2019; Eriksen 2020; Kitcher 2021). This paper therefore investigates how the use of digital technology can cause changes to core values and ideals in the current 'moral paradigm' of public administration in Denmark and thus cause fundamental moral changes in the Danish welfare state.

The paper investigates how the use of digital technologies in the Danish public administration, especially AIs in case management, can radically change some of the fundamental moral values, norms, and ideals, which are informing how public administration so far has been done in the Danish welfare state. Examples of this are that decisions has to be made with a view to what (within the bounds of the law) is best for the citizen, the principle that 'no assessment must be put under a rule', that the citizen has a right to know what the reasons for the decision in their case where, and that citizens has to be treated and answered in 'kind and forthcoming ways' (Jensen, Jensen and Motzfeldt 2020; Hundebøl, Pors and Sørensen 2020; Motzfeldt and Abkenar 2019).

Whether these fundamental moral changes will all be hindered, and whether they will – if not hindered – add up to something deserving to be categorized as a 'moral revolution' in public administration is something the future as well as philosophical debate on the conceptualization of moral revolutions will decide. But understanding how AI has the potential to change not only the efficiency and economic cost of current public administration but also the fundamental moral values to far informing it, underscores the importance of broader investigations of consequences prior to the introduction of AI – something which is unfortunately often lacking today.

References

Appiah, A.K. (2010). The Honor Code. How Moral Revolutions Happen. W. W. Norton & Company.

Baker, R. (2019). The Structure of Moral Revolution. Studies of Changes in the Morality of Abortion, Death, and the Bioethics Revolution. Cambridge, MA: The MIT Press. Box, R.C. (2014). Public Service Values. Oxon: Routledge.

Eriksen, C. (2020). Moral Change: Dynamics, Structure and Normativity. New York: Palgrave Macmillan.

Hundebøl, J., Pors, A., Sørensen, L.H. (eds). (2020). Digitalisering i offentlig forvaltning. Copenhagen: Samfundslitteratur

Jensen, J, Jensen, M, and Motzfeldt, H.M. (2020). Grundlæggende forvaltningsre. 8 udgave. Copenhagen: Samfundslitteratur. Kitcher, P. (2021). Moral Progress. Oxford: Oxford University Press.

Morris, I. (2015). Foragers, farmers, and fossil fuels: How human values evolve. Princeton University Press.

Motzfeldt, H. and Abkenar, A. T. (eds) (2019). Digital forvaltning. Copenhagen: Djøf Forlag. Swierstra, T. (2013). “Nanotechnology and Technomoral Change”. Etica & Politica, 15(1), 200–219.

Pleasants, N. (2018). “The Structure of Moral Revolutions”, Social Theory and Practice 44(4): 567-592.

Triaging the Technology Triad: Disruptive AI, Regulatory Gaps and Value Change

Jeroen Hopster; Matthijs Maas

In technology ethics, a common approach to anticipating and coping with value change is to look at the mutual shaping of emerging technologies and morality, in a dyadic or reciprocal relation between the two. Simultaneously, in the field of technology law (or 'TechLaw'), the focus is on the mutual shaping of emerging technologies and particular regulatory systems. In this paper, we propose to integrate these two dyadic models, to instead shift focus to the triadic relations and mutual shaping of values, technology, and regulation. We argue that a triadic values-technology-regulation model is more descriptively accurate, as it allows the mapping of second-order impacts of technological changes (on values, through changes in legal systems; or on legal systems, through changes in values). Simultaneously, it serves to highlight a broader portfolio of ethical, technical, or regulatory interventions that can enable effective ethical triage of – and a more resilient response to – Socially Disruptive Technologies, such as Artificial Intelligence. In the paper we sketch this triadic model, ground it in existing work, and then apply it to the case study of the differential societal impacts of AI. We highlight three strengths of the approach. First, to technology ethicists, a triadic model foregrounds the multiple realizability of ethical interventions and facilitates a shift from a reactive stance in the face of emerging AI, to a problem-solving and problem-finding orientation. Second, to technology lawyers, a triadic approach shifts away from debates over technological exceptionalism, and examines AI in conjunction with the broader dynamics of social change and value change in which it is implicated, grounding more tailored and resilient regulatory frameworks. Third, the triadic model enables triage amongst many technological changes, helping to identify where these may be most disruptive, and where ethical and regulatory interventions are most urgently needed.

How the Unhealthy Division of Labor in AI Calcifies Value Change: A Durkheimian Analysis

Zachary Daus

It has been argued by numerous authors that users of AI technologies are not merely consumers but prosumers, whose disclosure of personal data increases the profits of the owners of the technologies that they consume (Fuchs 2014, 2019; Zuboff 2019). It has also been argued that the division of labor between owners and prosumers of AI technologies is worryingly unequal, demonstrated perhaps most clearly by the claim that the majority of prosumers are largely unaware of the extent to which their personal data is extracted for the sake of the profits of owners. I argue that this imbalance in division of labor is not only questionable from the perspective of privacy rights or social alienation, as argued by authors like Christian Fuchs and Shoshana Zuboff, but is also detrimental to liberal societies that seek to foster a flexible plurality of value systems. In order to argue why a healthy division of labor in AI is significant for the maintenance of a liberal society — and the flexibility of values that such societies make possible — I will turn to the thought of Émile Durkheim. In The Division of Labor in Society (1984), Durkheim argues that the healthy interdependence of modernity's increasingly complex divisions of labor allows for the existence of a plurality of value systems which nonetheless remain united in solidarity behind the common interests of their shared labor. However, when divisions of labor become non-contractual or non-transparent, as in the case of digital labor, this dynamic pluralism becomes endangered. To ameliorate this danger, I argue that the division of labor between owners and prosumers of AI technologies must become more transparent, and consequently offer preliminary ways through which such transparency can be achieved. References

References

Durkheim, Émile. The Division of Labor in Society. Translated by W.D. Halls. New York: The Free Press, 1984.

Fuchs, Christian. Digital Labour and Karl Marx. London: Routledge, 2014.

Fuchs, Christian. Rereading Marx in the Age of Digital Capitalism. London: Pluto Press, 2019.

Zuboff, Shoshana. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. London: Profile Books, 2019.

Care robots and the value of veracity: Disruption or entrenchment?

Seppe Segers

It is frequently asserted that robot use in a care setting will be a transformative development. Accordingly, the introduction of care robotics in healthcare is identified by some authors as a disruptive innovation, in the sense that it will upend the praxis of care. It is an open ethical question to what extent this alleged disruption will also have a disruptive impact on well-rooted ethical concepts and principles. One

particularly prevalent worry is that the implementation of care robots, for instance in geriatrics, will turn deception into a routine component of elderly care, at least to the extent that these robots will function as simulacra for something that they are not (i.e. human caregivers). At face value, this may indeed seem to indicate a concern for how this technology may upend existing practices and relationships within a care setting. Yet, on closer inspection, this reaction may rather point to a rediscovery and a revaluation of a particularly well-entrenched value or virtue, i.e. veracity, which has received a predominant role within bioethics since the 1970's (and which was traditionally ignored by older codes of medical ethics). The virtue of veracity is one of the values that is mobilized to argue against a substitution of human caregivers (while a combination of care robots and human caregivers is much more accepted). Thus, the subject of this paper is to explore how the moral panic surrounding care robots should not so much be interpreted as an anticipated and probable disruptor in a care setting, but rather as a sensitizing – in a way conservationist – argument that identifies veracity as an established value that is supposed to be protected and advanced in a future care setting.

Day 2: How can we study value change with the help of AI and robotics?

Robots, AI and Value Change A Mixed-Method Approach to Investigate Value Change by Technological Innovations
Henk J. van Gils-Schmidt; J.-C. Põder, A. Räder; J. Wegner

Current approaches for analyzing value change by technological innovations can broadly be classified in two categories: those that develop scenarios to anticipate the influence of technologies on our value frameworks (e.g., Schwierstra et al., 2009) and those that see technological innovations as social experiments to be guided by ethical evaluation from within our current framework (e.g., van der Poel, 2013). However, the former approaches are rather speculative while the latter lack an instrument to anticipate value change caused by the technology (Kudina & Verbeek, 2019). In our paper, we develop an approach that combines the strengths of both, alternating empirical methodologies and ethical reasoning to bring domains of potential value change into view. We do so in context of the VoluProf research project that develops a mixed reality avatar and AI-based interaction for educational purposes. First, a foresight analysis is performed to develop both a catalogue of (potential) ethical considerations and predictive scenarios of the technological innovation (cf. Brey, 2012). Second, a mixed-method study with experts and potential users investigates the empirically validity of the catalogue and, additionally, aims to uncover additional ethical considerations. The catalogue and scenarios constitute the framework to investigate the change the technological innovation potentially causes in our value framework. The third step employs a questionnaire, a discussion panel, and qualitative interviews. The questionnaire, by presenting the scenarios as short vignettes (Paul et al., 2019), aims to provide insight into how a diverse group of potential users estimates the value change caused by the technology. The results are subsequently explored and concretized by a discourse-ethical, participative panel discussion with experts (Weber 2016) and qualitative interviews with a group of users, who undergo a demonstrator-experience of the technological innovation. The fourth step evaluates these results against the background of ethical theories, further specifying the catalogue and predictive scenarios while simultaneously laying the foundation for a future-oriented, empirically-grounded ethical guideline for policymakers and technology developers.

References

Brey, P. A. E. (2012). Anticipatory Ethics for Emerging Technologies. NanoEthics, 6(1), 1–13. https://doi.org/10.1007/s11569-012-0141-7

Kudina, O., & Verbeek, P.-P. (2019). Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy. Science, Technology, & Human Values, 44(2), 291–314. https://doi.org/10.1177/0162243918793711

Paul, L. A., McCoy, J., & Ullman, T. (2019). Modal Prospection. In J. McCoy, L. A. Paul, & T. Ullman, Metaphysics and Cognitive Science (S. 235–267). Oxford University Press. https://doi.org/10.1093/oso/9780190639679.003.0010

Swierstra, Tsjalling, Dirk Stemerding, and Marianne Boenink. 2009. “Exploring Techno-moral Change: The Case of the Obesitypill.” In Evaluating New

Technologies, edited by Paul Sollie and Marcus Duwell, 119-38. Dordrecht, the Netherlands: Springer.

van de Poel, Ibo. 2011. “Nuclear Energy as a Social Experiment.” Ethics, Policy & Environment 14(3): 285-90. Weber, K. (2016). Ein erweitertes Modell zur ethischen Evaluierung soziotechnischer Arrangements. Technische Unterstützungssysteme, 11.

AI’s invisibility: Distrusting AI’s choices
Deivide Garcia da S. Oliveira

We usually see AI as something that could compete with us or, as in sci-fi movies, as something that could exterminate humanity. However, there is, at least, a third option, where AI can cooperate with us, being an extension of our minds, socially and individually (Clark & Chalmers, 1998). This view of AI as an extension of our minds has many corollaries, such as how and if AI can be trusted to fulfill the expected task. This paper focuses on this issue, its strengths and weaknesses, exploring how it can change the content of our minds, as beliefs and tastes (Susser, 2019; Van Den Eede, 2011). To do this, we will examine a famous experiment about how recommendation systems slightly manipulate and change our personal preferences (Adomavicius, Bockstedt, Curley, & Zhang, 2013, 2018; Fleder & Hosanagar, 2009). In this account, we propose a principle of invisibility of AI (henceforth PIAI), which is here understood as the conscious delegation of our choices to AI. However, PIAI also has a byproduct, i.e., it gains our trust and drives us away from our original goals, with implications for our values. In conclusion, it will be clear why the application of AI, as an extension of our minds, does a good job in anticipating our choices, although it may result in an effective change of preferences, and eventually, things like our values. (Angwin, Larson, Mattu, & Kirchner, 2016). Thus, despite artificial intelligence being of great help, we still do not fully understand how AI solves problems (the so-called black-box of AI), which brings concerns about where and how AI is desirable. Keywords: AI invisibility, trustiness, beliefs change, values change, extension of mind.

Non-epistemic Values and Automation of Science
Michał Sikorski

Due to a number of influential arguments, it is now almost universally agreed in the philosophy of science that non-epistemic values should play a role in science. Arguments such as inductive risk argument (Rudner 1953) or underdetermination argument (Longino 1990) convinced the majority of the philosophers that non-epistemic values should influence practically relevant research. Such values should guide a scientist when she chooses the studied hypothesis, decides which experimental design to use or when the collected evidence is sufficient to accept the hypothesis, or makes other methodological decisions crucial for the results of the procedure. On the other hand, the possibility of value-free science seems to be supported by development in the automation of science. For example, ‘Robot Scientist’ presented in King et al. (2004), “[..]automatically originates hypotheses to explain observations, devises experiments to test these hypotheses, physically runs the experiments using a laboratory robot, interprets the results to falsify hypotheses inconsistent with the data, and then repeats the cycle.” (King et al. 2004 p. 247-248)

But how can a system that does not possess any non-epistemic values (we know that no such values were implemented) can perform tasks that traditionally were believed to require the use of such values? In my presentation, I will try to answer this question by arguing that the existence of systems like 'Robot Scientist' and more advanced systems, for example, developed in response to the "Nobel Turing Grand Challenge" (see ATI 2020), suggests that value-free science is possible. When, as in the case of such systems, a research question and methodology to be used is well defined, there is no need for value-laden choices. Consequently, it seems that the orthodox view concerning the role of non-epistemic values in science should be rethought in light of developments in the automation of science.

References

King, Ross D., et al. 2004. "Functional genomic hypothesis generation and experimentation by a robot scientist." Nature 427:247–252.

Longino, Helen E. 1990. Science as Social Knowledge: Values and Objectivity in Scientific Inquiry. Princeton University Press. 1 Rudner, Richard. 1953. "The Scientist Qua Scientist Makes Value Judgments." Philosophy of Science 20 (1): 1–6.

The Alan Turing institute. 2020. "AI Scientist Grand Challenge: Summary of Discussion." https: //www.turing.ac.uk/sites/default/files/2021- 02/summary_of_ discussion_workshop_2020_ai_scientist_grand_challenge_clean. pdf. 2

AI-Scientists and their Impact on Science

Maud van Lier; Andrea López Incera; Sahra Styger

The automation of science is a focus area of current AI research. Starting with automated research assistants, the eventual aim is to build a proper AI-Scientist that is able to independently conduct research, thereby contributing to our overall body of knowledge (see Kitano 2021). The aim of this paper is to make and defend three claims. The first two are about the impact on science of these AI-Assistants/Scientists. The third states what future research should pay attention to.

First, as Kitano (2021, 9) argues, the integration of AI-Assistants/Scientists increases the chance of finding something of value in research domains that were deemed of "low-value" by us before. AI-Assistants/Scientists can, given their ability to process huge amounts of data, focus on testing hypotheses seen as not worthy of our time. These tests might eventually lead to interesting and valuable findings.

Second, a benefit of integrating these AI-Assistants/Scientists in the research process is that they approach data with far less biases (except for the biases that we implicitly/explicitly give them). They might therefore explore directions that we would not have (immediately), as we would have found them counter-intuitive.

Third, current research on the automation of science has a rather limited view of science. First, it sees scientific research as 'generating and testing hypotheses' (Sparkes et al. 2010). Philosophers of science have shown that this notion of science

is too narrow and exclusive. If we want to challenge AI-research and create a truly general AI-Scientist, we should base it on a realistic and broad view of science. Second, it assumes that AI-Scientists/Assistants can be unbiased in their approach (Kitano 2021, 1). Even though the AI-Scientists/Assistants might be unbiased, scientists must be careful not to underestimate the bias that they themselves introduce in the design, the data, or the evaluation of the output.

References

Kitano, H. (2021). Nobel Turing Challenge: creating the engine for scientific discovery. npj Systems Biology and Applications, 7(1), 1-12.

Sparkes, A., Aubrey, W., Byrne, E., Clare, A., Khan, M. N., Liakata, M., ... & King, R. D. (2010). Towards Robot Scientists for autonomous scientific discovery. Automated Experimentation, 2(1), 1-11.

Day 3: How should AI and robotics deal with value change?

Value Change/Transformation in Artificial Agents: Understanding the nature of Artificial Values

Vipra Chopra

Ongoing research on value embodiment and technology in Philosophy has been exponential and has a vast scope for further research. Research is conducted on different aspects of value embodiment and my concern is the human understanding of the value embodied by the artificial agent. The question I am working on is not whether the artificial agent is capable of moral agency or intentionality but a question preliminary to a possibility of embodiment. Socio-technical systems are composed of technical artefacts, human agents, institutions, artificial agents and certain technical norms. Values represent the evaluative content of moral norms without placing focus on its descriptive content.

In the unlikely event of either a sound argument or an empirical evidence proving the outcome of artificial agents having the ability to possess moral agency and intentionality would not amount to forming the exact identical structure of the concept of value. A human agent has an understanding of the concept of value, a concept which applies to human agency and intentionality. The human agent should also have a distinct understanding of the concept of artificial value, a concept which applies to values embodied by an artificial agent. Once a human value is embedded into an artificial agent, the embodied value goes through a conceptual transformation and should not be misunderstood as the same as the human concept of value. My concern is not how the artificial agent understands this value but how the human agent understands this embedded value.

There are three factors which explain why we should have an understanding based on distinctive concepts of value for human agents and for artificial agents. The fundamental differences between the human agent and the artificial agent lead to three factors of variance. The first being Programming as an artificial agent can be programmed to embody values but a human agent cannot be. The second is the factor of Doubt. An artificial agent will not change its programing based on negative or unsuccessful outcomes by following a certain value even with infinite amount of negative outcomes. While a human agent will reconsider upgrading their value system even with a single negative outcome. The third and the most important is the factor of Time pressured capability possessed by the artificial agent. The artificial agent has an ability to respond much faster than human agents in time-pressured situations. The artificial agent has the possibility to embody and disembody a value an infinite number of times within a mere second which the human agent lacks. A value embodied and disembodied numerous times in a second questions the representation of an evaluative content of a norm. If a human agent would accept and reject the same value a number of times within an hour, it would give rise to questioning the value itself. To avoid any challenges which might be raised towards our existing understanding of the concept of value, I give a solution by having distinct concepts of human value and artificial value.

Artificial Intelligence and "broad meaningful human control"
Filippo Santoni de Sio

Control has always been a big theme in philosophy of technology, since thematized by Winner (1977) and Collingridge (1980), and stigmatized by Foucault (1975) and Deleuze (1992). Recent developments in AI and Robotics have prompted the more specific question about the (im)possibility to maintain (moral) control of complex and dynamic intelligent systems, such as autonomous weapon systems (Horowitz & Scharre, 2015). Keeping (intelligent) technology aligned to human intentions and values is also difficult because human values themselves are complex and dynamic (Van de Poel, 2018).

This paper critically reviews three existing concepts of control over technology: Collingridge's social control over technology, the Responsible Innovation approach (RI), and the recent "meaningful human control" (MHC) over AI systems. It argues that notwithstanding their merits, none of these concepts satisfactory addresses the different challenges of human control over (intelligent) systems in the light of value change. Collingridge recognizes the importance of designing for uncertainty and the normative plurality of expert opinions, but overlooks the emergence of meaning and value in the interaction with the technology (Kiran, 2012). The Responsible Innovation approach promotes a technological process that responds to the values of a broader range of stakeholders (Stilgoe et al., 2013). However, it does not offer specific design principles to keep AI systems aligned to (changing) human reasons. Recent theories of "meaningful human control" over AI (Santoni de Sio & van den Hoven, 2018) try to fill this gap but overlook the intrinsic uncertainty of the technological process, and the need for proactively including underrepresented voices in this process (Blok, 2014).

The paper concludes by proposing a new concept of control – "broad meaningful human control". This combines Collingridge's attention to uncertainty, RI's call for inclusivity, and MHC's specific focus on design principles for AI systems that respond to changing reasons.

References

Blok, V. (2014). Look who's talking: responsible innovation, the paradox of dialogue and the voice of the other in communication and negotiation processes. Journal of Responsible Innovation, 1(2), 171–190. https://doi.org/10.1080/23299460.2014.924239

Collingridge, D. (1980). The Social Control of Technology. Frances Printers.

Deleuze, G. (1992). Postscript on the Societies of Control*. Surveillance, Crime and Social Control, 35–39. https://doi.org/10.4324/9781315242002-3

Foucault, M. (1995). Discipline and punish: the birth of the prison (2nd Vintag). Vintage Books.

Horowitz, M. C., & Scharre, P. (2015). Meaningful Human Control in Weapon Systems: A Primer. Kiran, A. H. (2012). Does responsible innovation presuppose

design instrumentalism? Examining the case of telecare at home in the Netherlands. Technology in Society, 34(3), 216–226. https://doi.org/10.1016/j.techsoc.2012.07.001

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful Human Control over Autonomous 2 Systems: A Philosophical Account. Frontiers in Robotics and AI, 5, 15. https://doi.org/10.3389/frobt.2018.00015

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. Research Policy, 42(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Van de Poel, I. (2018). Design for value change. Ethics and Information Technology 2018 23:1, 23(1), 27–31. https://doi.org/10.1007/S10676-018-9461-9

Winner, L. (1977). Autonomous Technology. Technics-out-of-Control as a Theme in Political Thought. MIT Press.

Analysing Technological Colonialism in Sub-Saharan Africa: A Case for Many-Value AI
Edmund Terem Ugar

Since the introduction of the Fourth Industrial Revolution (4IR) technologies, especially AI technology and robotics, many ethical guidelines have been published. These ethical guidelines ensure that the values embedded in AI and robotics are in line with human values, i.e., those that promote the flourishing of the human species. I argue that current ethical guidelines of AI are mainly based on Western individualism/democracy or Asian Confucianism/autocracy and that, as such, these ethical guidelines have the potential to engender a technological colonialism in places that are not at the forefront of the technical development of these technologies, such as sub-Saharan Africa.

Once I have established this, I argue that 4IR technologies, specifically AI technology, should instead be undergirded by my novel approach, which is, a combination of relevant Western individualist values, Asian Confucian values, and Afro-communitarian values. My view is that such an approach will allow the engineering of AI technologies to promote inclusivity and diversity in machine learning. Furthermore, I apply my novel approach to AI care robots to show the socioeconomic benefits of having an AI that understands diversity.

When will the robot therapist take over? Ethical reflection on how artificial intelligent psychotherapy should take shape.
Mehrdad Rahsepar Meadi

Ever since chatbot ELIZA in the 1960s showed promising counselling skills one could wonder whether the artificial intelligence therapist could and should replace the human therapist. Currently this question is voiced even louder with the rapid rise of commercial psychotherapy AI-chatbots like Woebot and Wysa, that any person possessing a smartphone and an internet connection can consult. Today artificial

intelligence is seen as a way to keep mental health care accessible and affordable. With AI-chatbots, patients can possibly benefit from the online disinhibition effect and the ‘no-eyebrows’ effect. On the other hand, they might suffer from discriminatory biases based on ‘black box’ algorithms. When only the privileged few could keep access to human-to-human therapeutic contact, they can worsen not only social but also health inequalities (for example by lacking adequate crisis management). It can also lead to drastic changes in important values about medical responsibility, accountability (especially in the case of fully autonomous AI-therapists) and the patient-doctor relationship. This raises questions about whether an AI-therapist can be trusted, whether we should be concerned it will replace human-to-human contact and which boundaries and limits the AI-therapist should have. And hence, ethical guidelines are necessary on how to develop and implement these technologies. To prevent that the isolation of researchers within their own discipline will produce biases, we will conduct a broad systematic review in PubMed, Embase, APA PsychInfo, Web of Science, Scopus, Philosopher’s Index and ACM Digital Library with the aim of identifying and normatively analysing potential ethical worries. Moreover, we will argue that AI-psychotherapy can be considered a Socially Disruptive Technology and translate potential lessons from other SDTs. With the results we will undertake an ethical reflection on how to develop and implement AI-psychotherapy and give recommendations for further research.