Automating Morals – On the Morality of Automation Technology, Ironies of Automation and Responsible Research and Innovation

Christian Herzog, né Hoffmann*

**Institute for Electrical Engineering in Medicine, Ethical Innovation Lab, University of Lübeck, Germany (Tel: 0049 451-3101-6211; e-mail: christian.herzog@uni-luebeck.de).*

Abstract: The prevalence and impact of morals in technology design is increasingly better understood. Likewise, advances in machine learning, systems theory and control continue to push the boundary with respect to the applications in which automation may be considered. The present paper is intended to act as a precursor to a lively debate about professional ethics within the control community regarding automation in morally charged situations and beyond. First, the paper provides a primer on the actualities of applications in which morals already play a significant role. It further claims that–in contrast to typical expositions–within the scope of systems in which automation is employed, there is a continuum between addressing morally charged contexts to actually performing a kind of automated moral deliberation, though technically and philosophically there may be a vast difference. Second, from this perspective, the paper presents a first indication about potential new and persistent “ironies” within the context of automating morals. Third, the paper draws conclusions, essentially calling for the community to open up and engaging in participatory research and development settings as a matter of professional ethics.

Keywords: Ethics, Value systems, Sustainability.

1. INTRODUCTION

The field of machine ethics has sparked much debate about whether humankind should or should not build machines capable of moral deliberation, i.e., so-called artificial moral agents (AMAs). This has also led many to criticize the mere discussion about this issue as premature, see, e.g., (TU/e Cursor, 2019), or distracting from another issue, cf. (AlgorithmWatch and Bertelsmann Stiftung, 2019, p. 14), namely, that in some highly morally salient situations, decision-making processes are already being automated with the risk that a proper consideration of the ethical issues surrounding them is lacking. Yet others have pushed the notion that a mere incorporation of ethical analyses all-to-often only serves the purpose of a white- or ethics-washing and that the actual and persistent effects at hand, e.g., social inequality, should be put to the fore within the discussions of the community (Sloan, 2019). A common notion to these discussions, however, is the importance of addressing both ethical and societal implications, typically at a relatively deep scholarly level directly within technology development. This paper maintains that the field of automation and control can and should contribute to the discussion based on the observation that the merging of theory and methods from both machine learning and control offers a rich set of tools that allows automation in domains, that may be denoted as morally charged: Areas of application, in which–often implicitly–moral questions are being addressed.

The present paper aims at three interrelated objectives. It first tries to sensitize to the actualities of morals in technology designs, ranging from morally charged decision-support to automated moral decision-making. In doing so, the paper aims at providing a fruitful precursor to a debate on the notion of technological neutrality, i.e., whether and to what extent technology is morally neutral. The present paper argues for the case that, more often than not, morals do play a significant, albeit implicit, role in the conception of research directions, the selection of methods and even the composition of development teams. More explicitly, technology–including control theory and systems technology–is explicitly used in morally charged applications. The paper therefore tries to drag the discussion about the feasibility, desirability and potential benefits of AMAs onto a highly exigent level that does not neglect a present, already transpiring transition towards an actual “automation of morals”. In the paper, it is therefore made a case for increasing the efforts to establish inter- and perhaps also transdisciplinary development teams mandated by the ever-growing potential of methods from automation, machine learning and control to be directly applied to societally relevant challenges. At the same time, relevant arguments from the debate about AMAs can and should be considered now, recognizing that the transition from automation used in morally charged situations to actually automated moral deliberation, at least from a layperson’s perspective, might not be as abrupt and far-off as it may appear.

Second, considering the notion of a continuum between automation used in morally charged situations and automated moral deliberation–at least in terms of public perception–, the paper discusses analogies and differences to some of the so-called “ironies of automation”, cf. (Bainbridge, 1983). While in many examples of automation used in morally charged situations, a straightforward analogy may be drawn due to excessive demands that are imposed on human operators or supervisors, actual, if possible, and partly automated moral deliberation will present humans with novel challenges. The paper draws on the concept of "technological opacity" (Burrell, 2016; Herzog, 2019b) as a framework that may facilitate the analysis of the "ironies" and further implications.

Third, the paper aims at drawing conclusions and formulating recommendations that specifically aim at encouraging debate about the professional ethics of control theorists and practitioners. In reference to the paper's first objective, methods from "Responsible Research & Innovation" (RRI) (Grunwald, 2011) can be considered that aim at integrating ethical and societal analysis within the development process by means of participative approaches including perspectives from all stakeholders. While such an approach will potentially not prove to be a panacea in addressing all societal challenges, at the very least, it will provide the transparency and stakeholder involvement necessary to monitor and decide which route to take on automating morals or moral deliberation on a societal level. In reference to the paper's second objective, a case is made for specialists from science and technology studies, philosophy of technology, technology assessment, law and ethics to contribute accessible methods, tools and primers as well as engaging in interdisciplinary education on undergraduate, graduate and research levels to facilitate the integration of an analysis of ethical, legal and societal aspects (ELSA) directly within development processes. Empowering technology specialists with a set of preliminary methods to assess the basics of ELSA is promised to lead to more sustainable and societally desirable developments, essentially establishing ethics as a driver for innovation with analyses being performed *ex ante*, instead of ethics acting primarily as a means of critique *ex post*. The purpose is not to devalue the work of scientists from humanities, such as philosophy, ethics or sociology, who provide highly valuable independent analyses and grounds for debate. However, the paper argues that the speed and pervasiveness of automation technology, its development and dissemination, warrants a strong proactive commitment to try and guarantee that it serves the interests and needs of of society in its entirety. This, it seems, is an especially pressing agenda when considering the automation of morals.

The paper is structured as follows: Section 2 discussed the actualities in automating morals. Section 3 analyses "ironies" with respect to automating morals, while section 4 compiles suggestions about the professional ethics of control theorists and practitioners in handling morally charged control problems.

2. THE ACTUALITIES OF AUTOMATING MORALS

Even though there are valid doubts that the current methods of artificial intelligence (AI) are sufficient to construct AMAs (Brożek and Janik, 2019), it can be maintained that–at the very least–from the perspective of laypersons, the transition from morally charged decision-support to automated moral decision-making is a continuous one. For the purpose of this paper, denote "decision-support systems" to be "morally charged", if they propose decisions, which, even if they may be completely deterministically computed based on input data, would frequently challenge the moral compass of some human operator tasked with evaluating the proposed decision. In contrast, "automated moral decision-making" denotes closing-the-loop using an AMA, routinely omitting human oversight. This implies that "decision-support" versus "decision-making" concerns whether human oversight is put in place on a decision-by-decision basis, or whether it is routinely omitted, respectively, potentially except for a few random samples.

An example of an automated morally charged decision-making tool can be found in the everyday use of pacemakers equipped with capabilities to resuscitate (J. L. Millar, 2015). What is, in fact, a deep moral question about whether or not the life of a specific person should end, is answered once prior to implantation of a technological tool, which is put in place to try and enforce that available medical procedure is staying true to that answer (J. Millar, 2015). Obviously, such procedure is only performed with a patient's consent. However, the choice offered, or rather, the decision taken is attributed a permanence, which is possible to enforce exclusively due to technological advances.

Likewise, and perhaps more subtle than the often referenced trolley problem (Bonnefon, Shariff and Rahwan, 2019), autonomous cars could be conceived of being designed with automated routines that evaluate whether or not the breach of traffic rules will potentially help mitigate dangerous situations or might be allowed in order to improve traffic flow. Himmelreich, 2018, has highlighted the ethical implications of rather mundane driving situations as perhaps more relevant than dilemma situations due to the high frequency by which they occur. Salient trade-offs are found when balancing safety, mobility, ecology, etc. Decisions taken during the design of pervasive technologies have far-reaching ethical implications due to the "challenge of scale". How to exactly specify an automation procedure, in turn, suffers under the "challenge of specificity". This, of course, extends beyond autonomous driving. In fact, recent reports have given rise to assert the already relatively wide-ranging deployment of morally charged decision-support systems (*The Guardian*, 2019). Further pieces of literature even indicate the existence of some morally charged control loops, i.e., in absence of case-by-case human oversight (Lin, 2018).

Security screenings (Hall, 2017), border control, terrorism detection, (AlgorithmWatch and Bertelsmann Stiftung, 2019) and others mark highly sensitive areas in which automated decision-making is already employed. Perhaps one of the farthest developed areas of such automation can be found in finance, where high-frequency trading is automating buy or sell decisions with moral implications (Pasquale, 2015). Often the concept of a control loop enters these applications in a more nuanced way. Decision-making tools could be employed to improve key performance indices, respect funding limits or immigration targets. While the associated algorithms will, at least currently, not necessarily be designed to alter control variables online to meet desired set points, the means by which they function, e.g., even simple rule sets, will be designed for that purpose.

If decision-support or -making adheres to enacted laws, automation might be designed to exploit available margins of discretion to do so. These "judgement calls", cf. (Lin, 2018), abound in many further applications ranging from traffic control to social, or care robots.

The purpose of this paper is not to criticize the existence of automation in morally charged contexts. On the contrary, the benefits from medical implants to automated driving–if carefully designed–will often be highly desirable. Rather the purpose is to showcase the saliency of morals associated with these technologies, which, for itself, appears to be reason enough to engage with a wide range of stakeholders in progressing their development. This includes also considering the limits of the technologies at hand by interweaving their designs with proposed and promising non-technological means, cf. (Baker, 2019), e.g., in the case of autonomous driving and urban planning, or, e.g., accounting for epistemological limits to data-driven architectures, see, e.g. (Caliskan-islam, Bryson and Narayanan, 2016; Herzog, 2019a; Topol, 2019).

3. THE RELEVANCE OF THE DEBATE ABOUT ARTIFICIAL MORAL AGENTS

Besides highlighting the necessity of appropriate training for designers to be able to shoulder the responsibility to program devices that can cope with morally charged situations, cf. programs such as the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2018, the above-mentioned examples indicate that automation incorporating morals is not a fiction but a reality. Technological neutrality, at least in absolute terms regarding the entirety of technology, i.e., without wide-ranging exceptions, in turn, is a myth (Verbeek, 2006; J. L. Millar, 2015).

However, while acknowledging that technically, cf. (Fisher *et al.*, 2016), and philosophically, cf. (Brożek and Janik, 2019), there may still appear to be an unsurmountable gap from automation in morally charged situations to actual automated moral deliberation, it can be maintained that the debate about whether AMAs should or should not be built is highly relevant for at least two interrelated reasons: First, a layperson may either not be informed enough, or may not be exposed to automated decision-making extensively enough, such that she or he may not be able to tell the difference. Second, some ethical frameworks, such as a strict, quantified version of utilitarianism, may lend itself more easily to implementation with current programming paradigms, making moral deliberation–at least part of it–already possible.

With respect to the first argument, it is often stated that humankind's pursuit of AMAs will lead to or requires a proper definition of "intelligence". By defining "intelligence" to "do the right thing at the right time", Bryson, 2018, claims that the dichotomy of AMAs and automation acting in morally charged situations is a misleading depiction. Instead, it is a continuum and if it is perceived that way, it is suggested that it will allow designers to more readily recognize the responsibility that is involved in the design process. Admittedly, and even though there appears to be no coherent definition of "intelligence" in the literature, Bryson's attempt may seem too simplistic from the perspective of, e.g., cognitive and neurosciences, or psychology, cf., e.g., (Legg and Hutter, 2007; Weinbaum (Weaver) and Veitas, 2017). However naïve, it may quite accurately formulate an intuition that, at least from brief encounters with any agent, captures a layperson's benchmark of the term "intelligent". The definition therefore bears appeal for being practically relevant in preliminary evaluations of ethical and societal implications, precisely, because it highlights how an autonomous system may be perceived in its day-to-day encounters. Consumers of technology and customers with no working in-depth knowledge about artificial intelligence will likely only evaluate based on a few "data points", such as phone calls with automated assistants like Google Duplex (Chen and Metz, 2019), service chats, etc.. Many of these encounters may not even last long enough to successfully evaluate, whether the Turing Test would have been passed, however debatable its meaning in defining "intelligence" may be, let alone it being a useful test for "automated moral deliberation" (Arnold and Scheutz, 2016).

The ethical and societal implications from automation technology that only seems to explicitly act intelligently, or morally, does not preclude that researchers should be content to perceive artificial moral agents merely from an input-output perspective. "Mind-less morality" (Floridi and Sanders, 2004) or taxonomical approaches such as those presented by (Cave *et al.*, 2018) or (Beavers, 2011) compress the one put forward by (Moor, 2006) by disposing of any relevance of the moral interiority of an agent. Within their approaches only consequences are deemed relevant, but not the qualities of the process involved. Such a narrowing of the concept of moral agency is perhaps a result of machine ethics' focus on what may be grasped within the confined realm of current ways of implementation, cf. (Tonkens, 2009), but, in the author's opinion, it remains epistemologically lacking. That being said, it is a different matter what ethical and societal implications arise even from illusions of moral agency. As so often, it can be meaningfully maintained that assuming the perspective of the most vulnerable will highlight salient ethical aspects.

While legislation such as the GDPR[1], may grant individuals some certainty with respect to the fact that fully automated decision-making processes are prohibited, at least when concerning personalized data (Dreyer and Schulz, 2019), there is likely to be a high level of uncertainty on the extent of automation and what it will be capable of. For instance, margins of discretion are an important concept in administrative issues (Hall, 2017) and it is important for citizens to know at which point those margins are being utilized to facilitate due process. The main point then refers to the concept of "technological opacity", which may be categorized in different ways, e.g., by the source it emanates from, cf. (Burrell, 2016; Herzog, 2019b). Vulnerabilities are often generated by power gradients and a prominent source of

[1] General Data Protection Regulation

these stems from differences in knowledge, expertise and understanding related to technology.

With respect to the second argument, e.g., basic utilitarian implementations of moral deliberation might be crude, overly simplistic and will not do justice to the term "deliberation", but they constitute a sometimes-necessary form of an evaluative method, nonetheless. In that regard, it may even be argued that rapid progress in the field of AMAs utilizing only this sort of limited way of offsetting quantifiable costs with benefits, could even do harm to the perceived meaning of "moral deliberation", cf. (Herzog, 2020). In fact, arguments abound that machines could be more objective in evaluating facts, means and perhaps even morals, cf., e.g., (Gips, 1991). While this may not be entirely wrong, it is mandatory to reclaim that the value of an actual moral deliberation lies in a fair evaluation of arguments that is charitable to opposing views and notions and therefore aims at just conclusions, rather than being overburdened by subjective human biases. This entails value being attributed to actual debate, persuasion and discourse, which rather precludes top-down automated moral governance in the name of increasing efficiency.

This paper's claim that, at least in terms of perceptive and conceptual vagueness, there may not be a clear distinction whether an algorithm is deterministically operating in morally charged situations or whether it performs actual moral deliberation is not meant to cater to a notion that designing AMAs is unavoidable. It clearly is, as van Wynsberghe and Robbins, 2018, elaborate on. Therefore, it is a decision whether AMAs should be built, and society should answer that question through an open and discursive process.

4. IRONIES OF AUTOMATING MORALS

From a control viewpoint it should be clear that many of the above examples can be easily pictured as feedback loops or, at least, as feedforward control designs. These would incorporate not just classical continuous-valued variables, but also, perhaps owing to the use of neural network architectures, highly unstructured data. Decision-support systems can be thought of as additional sensory input to a human actuator, while the latter can be replaced, e.g., by an automated decision logic or an actual AMA. From this familiar perspective of control, it is not a far stretch then, to assume that instances of the above, potentially simplistic, model of control loops, be they morally charged or incorporating actual AMAs, may soon assume the role of a "hidden technology" similar to the way classical control loops have (Åström, 1999; Craig, 2018). Similar to the point about the nature of moral deliberation as a fair process made above, it appears vital that the control objectives underlying the architectures are chosen with care, questioning the viability of purely quantified, potentially reductionist set-points against more accurate multi-variate or fuzzy metrics.

It appears that it is the "hiddenness" that presents at least as many grounds for a scientific analysis of the ethical and social implications of morally charged control applications as the more classical automation schemes and control loops that Bainbridge's seminal paper "The Ironies of Automation" (Bainbridge, 1983) has addressed. Bainbridge, 1983, and others after her, cf. (Baxter *et al.*, 2012), have largely focussed on manual control, cognitive and monitoring skills in human-computer interaction. The identified "ironies" therefore mainly consisted in vastly increased requirements with respect to these skills on behalf of human operators and supervisors, despite the fact that automation should be employed to ease the burden. Especially when considering how humankind's dependence on technology has evolved to constructions of systems of systems, technologists, psychologists and human-machine interface specialists have all worked hard to mitigate the ironies and harness the benefits of automation. However, the general issue remains relevant and warrants specialized training for both operators and developing teams (Baxter *et al.*, 2012).

The present paper attempts to outline some ironies associated with morally charged and automated moral control loops. In doing so, it presents some non-technical limitations that an "automation of morals" is likely to face.

Delegating moral decisions might challenge the moral deliberation process of a human operator unduly. This does not only concern the relevant and often discussed notion of moral deskilling (Vallor, 2015). Instead, as indicated by a study by Krawczyk and Sylwestrzak, 2018, humans appear to require additional deliberation time to overcome instinctive impulses, such as envy, and to more carefully evaluate a morally charged situation. There is a strong perceived difference to the ironies of automation and those associated with automating morals as the output will usually not be a quantified value, say of a controlled concentration that should be maintained within prespecified limits, even though some demand that "AI researchers and ethicists need to formulate ethical values as quantifiable parameters" (Polonski, 2017). It is doubtful both whether this is possible and desirable, perhaps only so in a few select cases. Instead we will more likely be dealing with concrete decisions, which may be accepted (or not) based on moral intuition. This intuition, some researchers argue, appears to fail with regard to recent technological systems, which have little precedence or analogies within the "natural", i.e., non-technologically augmented realm (Klein, 2016). However, this does not suffice as a supervisory concept as the moral deliberation process is more involved, potentially requires different perspectives and, again, time.

The point that moral deliberation is a complex endeavour, however, applies equally well to the calculations performed during classical control. Moral deliberation may, at some point, be artificially computed orders of magnitude faster than is possible for humans, similar to the control algorithms of today. Therefore, the main irony does not concern the processing, but the evaluation of the outcome, which, as described above, may, or perhaps even should, be qualitatively different if moral considerations are involved.

Even still, the question remains, whether humans will actually want to delegate moral deliberation. If not, challenging every decision output from an algorithm by means of human moral deliberation will require vast amounts of time. Overall, the irony here is that the increased efficiency promised by automation may be rendered void, as human oversight might require a limiting time constant to be imposed such that operator supervision is actually possible. The danger remains, though, that increases in efficiency are too tempting, such that human oversight is dispensed with.

A common refute to even the classical ironies of automation proposed by Bainbridge, 1983, usually consists in requiring better, more reliable and even more complete automation. If the need for human oversight can be significantly reduced, then the burden on a human operator is lifted in equal measure. However, with morally charged situations and especially with the actual automation of moral deliberation, humans will not only be involved in terms of supervision, they will also be the recipients of the decisions insofar as that someone will have to eventually accept the automated decisions on his/her behalf or with respect to his/her person. Studies indicate, however, that the delegation of moral decisions to machines incurs severe reservations (Gogoll and Uhl, 2018). People will want to scrutinize moral decisions and they will need their proper time to do so. It is this insight, that, I believe, imposes a significant limitation to attempts at increasing our output per time of moral decisions, i.e., “moral efficiency”. Professional moral discourse might be required or even trials might ensue for which it may not be possible to increase efficiency at the same pace as AMAs would allow to increase the “moral efficiency”. Drastic effects have already become visible with moderators of social media platforms, see, e.g., (Barnett and Hollingshead, 2012; Chen, 2014), for which calls for machine augmented moderation–arguably partly including a moral deliberation process–have appeared, cf. (Ruckenstein and Turunen, 2019). The sheer magnitude of incidences in social media moderation is already stressing the limits of public discourse as well as appropriate and timely legal action. The construction of additional moral entities other than humans that humans need to hold to account, hence, may be seen as a danger to any society’s legal system infringing on the rule of law. Such dangers may be averted, but until humankind has figured out how, it might be ill-advised to introduce further moral entities.

Others have proposed a so-called “Socratic assistant” for realizing moral enhancement (Lara and Deckers, 2019). The idea is not to automate the moral deliberation, but to design a companion that aids humans with the deliberation process in posing, as was Socrates’ habit, the right questions. A potential irony may consist in the development of automation biases (Goddard, Roudsari and Wyatt, 2011), i.e., a propensity to simply trust the assistant’s suggestions. Such drawbacks may be easier to mitigate by designing systems to withhold final verdicts. Consequently, increasing “moral efficiency” may not be the overall goal, rather than increasing “moral effectivity”–facilitating more carefully reflected moral decisions. Such an approach, however, is clearly limited to situations that do not require the automation of deciding on “judgement calls” with margins of discretion.

5. CONCLUSIONS

Within this paper, the following considerations have been addressed, or touched upon, so far: First, there is a need to an increased sensitization to the morals implicitly imbued in automation technology and the perceived continuum of employing automation in morally charged situation to fully delegating moral deliberation to algorithms in certain cases. Second, and because there is an abundance of automation utilized in morally charged situations, a brief introduction to possible ironies, ethical and societal implications of automating morals have been given.

The questions posed and implications sketched warrant a more resourceful and rigorous treatment within the automation and control profession with the aim to guide the good and purposeful intentions of the community towards providing society with helpful tools that meet societal challenges without accidentally increasing them. In response, while control and automation experts may rightfully be proud because it is their science that enables many applications and makes things work, maybe it is time to ditch the pride associated with the notion of automation and control as “the hidden technology that society cannot live without”, cf. (Craig, 2018), at the very least when considering morally charged automation and the road where this is leading. The simple reason, why we need more debate, openness and participation in developing such systems, consists in the fact that the “hiddenness”, or “technological opacity”, cf. (Burrell, 2016; Herzog, 2019b), severely inhibits a society’s capability to scrutinize.

Incorporating methods from “Responsible Research & Innovation” (Grunwald, 2011) will likely incur more development time and costs, but promise to lead to solutions (and marketable products) that effectively tackle societal challenges, generate social desirability and advance the underlying science due to intriguing new requirements in a sustainable way. It is also morally charged control and automation that can strongly benefit from such an approach. In following down such a route, it also generates the stakeholder involvement and participation that is needed in addressing the questions about the desirability of a potentially impending automation of moral deliberation.

REFERENCES

AlgorithmWatch and Bertelsmann Stiftung (2019) *Automating Society: Taking Stock of Automated Decision-Making in the EU*.

Arnold, T. and Scheutz, M. (2016) ‘Against the moral Turing test: accountable design and the moral reasoning of autonomous systems’, *Ethics Informat. Technol*. Springer Netherlands, 18(2), pp. 103–115.

Åström, K. J. (1999) ‘Automatic Control - The Hidden Technology’, in Frank, P. M. (ed.) *Adv. Contr.*, pp. 1–28.

Bainbridge, L. (1983) ‘Ironies of Automation’, *Automatica*, 19(6), pp. 775–779.

Baker, P. C. (2019) ‘Collision course: why are cars killing more and more pedestrians?’, *The Guardian*, 3 October.

Barnett, E. and Hollingshead, I. (2012) ‘The Dark Side Of Facebook Memes’, *Telegraph*, pp. 8–11.

Baxter, G. *et al.* (2012) ‘The ironies of automation … still going strong at 30?’, in *Proc. ECCE 2012 Conf.*. Edinburgh, North Britain, pp. 65–71.

Beavers, A. F. (2011) ‘Could and Should the Ought Disappear from Ethics?’, in Heider, D. and Masanari, A. (eds) *Digital*

Ethics: Research and Practice. New York: Peter Lang, pp. 197–209.

Bonnefon, J.-F., Shariff, A. and Rahwan, I. (2019) ‘The Trolley, the Bull Bar, and Why Engineers Should Care About the Ethics of Autonomous Cars’, *Proc. IEEE*, 107(3), pp. 502–504.

Brożek, B. and Janik, B. (2019) ‘Can artificial intelligences be moral agents?’, *New Ideas in Psychology*. Elsevier Ltd, 54(April 2017), pp. 101–106.

Bryson, J. J. (2018) ‘Patiency is not a virtue: the design of intelligent systems and systems of ethics’, *Ethics Informat. Technol.*. Springer Netherlands, 20(1), pp. 15–26.

Burrell, J. (2016) ‘How the Machine “Thinks:” Understanding Opacity in Machine Learning Algorithms’, *Big Data Soc.*, 3(1), pp. 1–12.

Caliskan-islam, A., Bryson, J. J. and Narayanan, A. (2016) ‘Semantics derived automatically from language corpora necessarily contain human biases’, *Sci.*, 356(April), pp. 183–186.

Cave, S. J. *et al.* (2018) ‘Motivations and Risks of Machine Ethics’, *Proc. IEEE*, 107(3).

Chen, A. (2014) ‘The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed’, *WIRED*.

Chen, B. X. and Metz, C. (2019) ‘Google’s Duplex Uses A.I. to Mimic Humans (Sometimes)’, *The New York Times*, 22 May.

Craig, I. (2018) ‘Automatic control: The hidden technology that modern society cannot live without’. University of the Witswatersrand, Johannesburg, South Africa.

Dreyer, S. and Schulz, W. (2019) *The General Data Protection Regulation and Automated Decision-making: Will it deliver?* BertelsmannStiftung.

Fisher, M. *et al.* (2016) ‘Engineering Moral Agents - from Human Morality to Artificial Morality’, *Dagstuhl Reports*, 6(5), pp. 114–137.

Floridi, L. and Sanders, J. W. (2004) ‘On the Morality of Artificial Agents’, *Minds and Machines*, 14(3), pp. 349–379.

Gips, J. (1991) ‘Towards the Ethical Robot’, *Android Epistemology*, (May), p. 13.

Goddard, K., Roudsari, A. and Wyatt, J. C. (2011) ‘Automation Bias – A Hidden Issue for Clinical Decision Support System Use’, *Intl. Perspect. Health Informat. Studies in Health Technol. Informat.*, 164, pp. 17–22.

Gogoll, J. and Uhl, M. (2018) ‘Rage against the machine: Automation in the moral domain’, *J. Behav. Exp. Econ.* Elsevier, 74(March), pp. 97–103.

Grunwald, A. (2011) ‘Responsible Innovation: Bringing together Technology Assessment, Applied Ethics, and STS Research’, *Enterpr. Work Innov. Studies*, 7, pp. 9–31.

Hall, A. (2017) ‘Decisions at the data border: Discretion, discernment and security’, *Sec. Dialogue*, 48(6), pp. 488–504.

Herzog, C. (2019a) ‘Ethical and Epistemological Challenges of AI in Medical Systems’, in *41st Intl. Eng. Med. Biol. Conf., Symp. ‘Ethical Design Considerations for MedTec’*. Berlin, Germany.

Herzog, C. (2019b) ‘Technological Opacity of Machine Learning in Healthcare’, in *2nd Weizenbaum Conf.* Berlin, Germany.

Herzog, C. (2020) ‘Should We Build Artificial Moral Agents?’, *Sci. Eng. Ethics*. submitted.

Himmelreich, J. (2018) ‘Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations’, *Eth. Theory Moral Pract*. Springer Netherlands, 21(3), pp. 669–684.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2018) *Ethically Aligned Design - Version 2*.

Klein, W. E. J. (2016) ‘Problems with moral intuitions regarding technologies’, *IEEE Potentials*. IEEE, 35(5), pp. 40–42.

Krawczyk, M. and Sylwestrzak, M. (2018) ‘Exploring the role of deliberation time in non-selfish behavior: The double response method’, *J. Behav. Exp. Econ.*, 72(Dec. 2017), pp. 121–134.

Lara, F. and Deckers, J. (2019) ‘Artificial Intelligence as a Socratic Assistant for Moral Enhancement’, *Neuroethics*. Springer Nature.

Legg, S. and Hutter, M. (2007) ‘A Collection of Definitions of Intelligence’, in Goertzel, B. and Wang, P. (eds) *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*. IOS Press, pp. 17–23.

Lin, P. (2018) ‘The Moral Gray Space of AI Decisions’, *The Ethical Machine - Big Ideas for Designing Fairer AI and Algorithms*.

Millar, J. (2015) ‘Technology as Moral Proxy: Autonomy and Paternalism by Design’, *IEEE Technol. Soc. Mag.* IEEE, 34(2), pp. 47–55.

Millar, J. L. (2015) ‘Technological Moral Proxies and the Ethical Limits of Automating Decision-Making In Robotics and Artificial Intelligence’.

Moor, J. H. (2006) ‘The Nature, Importance, and Difficulty of Machine Ethics’, *IEEE Intell. Syst.*, 21(4), pp. 18–21.

Pasquale, F. (2015) *The Black Box Society - The Secret Algorithms That Control Money and Information*. Cambridge, Massachusetts; London, England: Harvard University Press.

Polonski, V. (2017) ‘Can we teach morality to machines? Three perspectives on ethics for artificial intelligence’, *medium*.

Ruckenstein, M. and Turunen, L. L. M. (2019) ‘Re-humanizing the platform: Content moderators and the logic of care’, *New Media Soc.*

Sloan, M. (2019) ‘Inequality Is the Name of the Game: Thoughts on the Emerging Field of Technology, Ethics and Social Justice’, in *2nd Weizenbaum Conf.* Berlin, Germany.

The Guardian (2019) ‘Automating Poverty Series’.

Tonkens, R. (2009) ‘A challenge for machine ethics’, *Minds Mach.*, 19(3), pp. 421–438.

Topol, E. J. (2019) ‘High-performance medicine: the convergence of human and artificial intelligence’, *Nat. Med.* Springer US, 25(1), pp. 44–56.

TU/e Cursor (2019) *Report on Symposium ‘Artificial Moral Agents?’* Available at: https://www.tue.nl (Accessed: 11 November 2019).

Vallor, S. (2015) ‘Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character’, *Phil. & Technol.*, 28(1), pp. 107–124.

Verbeek, P. (2006) ‘Materializing Morality: Design Ethics and Technological Mediation’, *Sci., Technol. Human Val.*, 31(3), pp. 361–380.

Weinbaum (Weaver), D. and Veitas, V. (2017) ‘Open ended intelligence: the individuation of intelligent agents’, *J. Exp. & Theor. Artif. Intell.*, 29(2), pp. 371–396.

van Wynsberghe, A. and Robbins, S. (2019) ‘Critiquing the Reasons for Making Artificial Moral Agents’, *Sci. Eng. Ethics*. Springer Netherlands, 25(3), pp. 719–735.