Ethics of Autonomous Systems

Annotated bibliography of recommended materials

Lars S. Laichter
QUC, ELTE
lars.laichter@gmail.com

April 25, 2019

Contents

Introduction

This bibliography has been compiled for the Self-Driving Cars Research Group at the Eötvös Loránd University (ELTE) in Budapest. Resources in this document are organised under various motivating topics where each topic constitutes a section. Each section is comprised of various subsections. Articles are placed in the most suitable subsections but are also often relevant to other topics in the subsections in the same section or even subsections in other sections.

Most articles have been sourced from PhilPapers based on their "impact" rating. Most papers have been selected out of the top 100 most impactful papers in their correspondent category. Further papers were then selected based on apparent relevance of the references of these papers.

This bibliography is accompanied by an online Zotero collection. This collection can be accessed at `https://www.zotero.org/groups/2190549/ethics_of_autonomous_systems`

1 Agenthood: Can we view autonomous systems as agents?

1.1 Status: Are autonomous systems agents in any sense?

- Bernd Carsten Stahl. 2004. "Information, Ethics, and Computers: The Problem of Autonomous Moral Agents." *Minds and Machines* 14, no. 1 (February 1): 67–83. ISSN: 0924-6495, 1572-8641, accessed June 27, 2018. doi:10.1023/B:MIND.0000005136.61217.93. http://link.springer.com/article/10.1023/B:MIND.0000005136.61217.93

  In this paper, Stahl discusses the possibility of computers being artificial moral agents. He centres the discussion around the concept of information. He defines information as data endowed with meaning. Given that computers in their current form are unable to capture the meaning of information and therefore fail to reflect morality in anything but a most basic sense of the term. He discusses this shortcoming in the context of the Moral Turing Test. The paper ends with a consideration of which conditions computers would have to fulfil to be able to use information in such a way as to render them capable of acting morally and reflecting ethically. He concludes that computers in their current form do not appear to be candidates for moral agency, mainly because they do not capture the meaning of the data they process. In addition, he poses the question of whether this result is a fundamental and which changes would be necessary to render the am autonomous moral agents. He claims that computers would have to understand meaning they would have to be in the situation, to be in the world in a Heideggerian sense (Heidegger, 1993), to share a life-world with others. Furthermore, the agent would need a capacity to learn. This paper is very much in conversation with other frequently cited authors on AMAs, such as Floridi, Sanders, and Allen.

- Deborah G. Johnson. 2006. "Computer systems: Moral entities but not moral agents." *Ethics and Information Technology* 8, no. 4 (December 5): 195–204. ISSN: 1388-1957, 1572-8439, accessed May 25, 2018. doi:10.1007/s10676-006-9111-5. http://link.springer.com/10.1007/s10676-006-9111-5

  Johnson in her article opposes the general conception of an artificial moral agent as disconnected from human intentionality. She criticizes authors, such as Floridi and Saunders, for falsely attributing computer systems mental states and claims that the failure to recognize the intentionality of computer systems and their connection to human intentionality and action, hides the moral character of computer systems. On the other hand, she argues that computer systems have intentionality, and because of this, they should not be dismissed from the realm of morality in the same way that natural objects are dismissed. She pays a close attention to the role of intentionality in relation to the artefact designer, artefact, and artefact user, which, as she claims, are at work when there is an action and all three should be the focus of moral evaluation. Furthermore, her argument relies on two fundamental distinctions, the distinction between natural phenomena or natural entities and human-made entities, and the distinction between artefacts and technology, to which she dedicates a major part of the discussion. Her argument is grounded in the work of Johnson and Powers who provide an account of the intentionality of artefacts in which the intentionality of artefacts is connected to their functionality. Overall, her argument is an interesting opposition to the general conception of artificial moral agents in the literature, while bringing in important definitional distinctions and grounding the computer as a human-made artefact which appears to be frequently overlooked.

- Luciano Floridi and J. W. Sanders. 2004. “On the Morality of Artificial Agents.” *Minds and Machines* 14, no. 3 (August 1): 349–379. ISSN: 0924-6495, 1572-8641, accessed June 27, 2018. doi:10.1023/B:MIND.0000035461.63578.9d. http://link.springer.com/article/10.1023/B:MIND.0000035461.63578.9d

  In this paper, the authors investigate to which extent do ethics lie within the human domain. To do so, the authors argue that insisting on the necessarily human-based nature of the agent means undermining the possibility of understanding another major transformation in the ethical field, the appearance of artificial agents (AAs). They analyze the concept of an agent, for which they introduce the fundamental ‘Method of Abstraction’(LoA). The new concept of moral agent is used to argue that AAs, though not intelligent and fully responsible, can be fully accountable sources of moral action. Moreover, they propose that morality is captured as a ‘threshold’ defined on the observables determining the LoA under consideration. An agent is morally good if its actions all respect that threshold; and it is morally evil if some action violates it. The use of the Method of Abstraction, LoAs and thresholds enables responsibility and accountability to be separated when the levels of abstraction involve numerical variables, as is the case with digital AAs. They conclude that there is substantial and important scope, particularly in Computer Ethics, for the concept of a moral agent not necessarily exhibiting free will, mental states or responsibility. Overall, this paper contains one of the most frequently cited conceptions of moral personhood for artificial agents.

- Shane Legg and Marcus Hutter. 2007. "A Collection of Definitions of Intelligence." *arXiv:0706.3639 [cs]* (June 25). Accessed August 5, 2018. arXiv: `0706.3639`. `http://arxiv.org/abs/0706.3639`

  This paper is a survey of a large number of informal definitions of "intelligence" that the authors have collected over the years. The 70-odd definitions presented in this survey are, to the authors' knowledge, the largest and most well-referenced collection there is.

- Luc Steels. 1995. "When are robots intelligent autonomous agents?" ISSN: 10.1016/0921-8890(95)00011-4, accessed August 5, 2018. `https://digital.csic.es/handle/10261/127979`

  Steel explores a biologically inspired definition of intelligent autonomous agents. In his account, intelligence is related to the behaviour of a system which contributes to its self-maintenance; behaviour becomes more intelligent when it is capable to create and use representations, and dynamics at various levels of intelligent systems play an essential role in forming representations. Steel defines agent to be a system within another agent and/or a system. He emphasizes the capacity of self-maintenance which makes his conception of a system closely tied to any biological system. He highlights autonomy as another property of being an agent and defines it as 'being relatively independent of' something. AI systems are in his view not as autonomous as their objectives and motivations are only those of their designers. He proceeds to discuss some other conceptions of machine intelligence, such as the definition based on a comparison to a human via that Turing test, or definitions based on knowledge and intentionality. Finally, he discusses the role of representation which, as he concludes, are not necessarily explicit but may be implicit.

- Patrick J. Hayes, Kenneth M. Ford, and Neil Agnew. 1994. "On babies and bathwater: A cautionary tale." *AI magazine* 15 (4): 15

  The authors argue that as in the old aphorism, one should not throw out the baby with the bathwater, the popular recent positions in AI theory have done just that by rejecting the "useful" idea of mental representations. According to the authors, various “situated” perspectives correctly emphasize that agents live in a social world, using their environments to help guide their actions without needing to always plan their futures in detail; but they incorrectly conclude that the very idea of mental representation is mistaken. They discuss these ideas and disputes in the form of an illustrated fable concerning nannies and babies. Overall, the paper is a good discussion of the representational versus situated conception of an agent, however, might be outdated given it is from 1994.

- George Kiss. 1990. “Autonomous Agents, AI and Chaos Theory”

  Kiss argues that the theory of dynamical systems, specifically chaos theory, should be used for descriptions of artificial agents. In this paper, it is proposed that a natural reinterpretation of agent-theoretic intentional concepts like knowing, wanting, liking, etc., can be found in process dynamics. It takes the first steps in providing an interpretation of agent theoretic concepts in terms of process dynamics concepts. If successful, this approach could have the double advantage that it provides both a mathematical theory which is already known to be successful in other application domains like physics and biology, and it also gives a hint at implementation strategy.

- Jose C. Brustoloni. 1991. *Autonomous Agents: Characterization and Requirements*

  Brustoloni dismisses the traditional dichotomy between reactive and symbolic architectures and instead provides a classification based on the amount of knowledge embedded in the system. He defines an autonomous agent as a system capable of autonomous, purposeful action in the real world. In his account, autonomous agents must be reactive, goal-directed, and must have a corresponding hierarchy to their goals, as well as the capacity to find out how to satisfy their goals. Brustoloni discusses various types of agents, including the regulations agents, planning agents, and adaptive agents. He introduces the concept of a drive to address the issues of where an agent’s goals come from. Finally, Brustoloni proposes and discuss a parallel architecture that embodies some of the ideas presented.

1.2 Criteria: What are the criteria for an autonomous system to be considered an agent?

- Deborah G. Johnson and Thomas M. Powers. 2001. "Computers as Surrogate Agents." In *Information Technology and Moral Philosophy,* edited by Jeroen van den Hoven and John Weckert, 251–269. Cambridge: Cambridge University Press. ISBN: 978-0-511-49872-5 978-0-521-85549-5 978-0-521-67161-3, accessed July 5, 2018. doi:10.1017/CBO9780511498725.014. https://www.cambridge.org/core/product/identifier/CBO9780511498725A020/type/book_part

  The authors engage with the possibility of computers being moral agents. They argue that human agency is a good model for understanding the moral agency of computers and that human surrogate agency is a good model for understanding the moral agency of computers. They examine the structural parallels between human surrogate agents and computer systems to reveal the moral agency of computers. The paper is divided into four parts in which (1) they discuss the "role morality" of human surrogate agency and the nature of agency relationships; (2) specify more carefully what they mean by computers, computer programs, and robots; (3) draw parallels between human surrogate agents and computer systems and maps the moral framework of human surrogate agency onto the agency of computer systems; (4) they review the account they have given and assess its implication. They conclude that computer systems have a certain kind of moral agency and this agency and the role of this agency in morality should not be ignored.

- Stan Franklin and Art Graesser. 1996. "Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents." In *International Workshop on Agent Theories, Architectures, and Languages,* 21–35. Springer

  The authors propose a formal definition of an autonomous agent, offer a natural kinds taxonomy of autonomous agents, and discuss possibilities for further classification. They also discuss sub-agent and multi-agent systems. They discuss various other definitions of what is an agent from the literature. Subsequently, their provide their own definition: "An autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future."

- John P. Sullins. 2006. "When is a robot a moral agent?"

Sullins argues that in certain circumstances robots can be seen as real moral agents. In his account, a robot does not have to have personhood to be a moral agent. He proposes three requirements for a robot to be seen as a moral agent: (1) it has to be significantly autonomous; (2) its behaviour can be only explained by ascribing to it some predisposition or 'intention' to do good or harm; (3) the robot to behaves in a way that shows an understanding of responsibility towards some other moral agent. He reviews various positions, such as the by Dennett 1998, Bringsjord 2007, Irrgang 2006, Nedeau 2006, or Floridi and Sander 2004. He concludes that robots are moral agents when there is a reasonable level of abstraction under which we must grant that the machine has autonomous intentions and responsibilities.

- J. H. Moore. 2001. "The Turing Test: Past, Present and Future." *Minds and Machines* 11 (1)

  The Turing test has been one of the longest standing proposed criteria for machine intelligence. Moor discusses the role of the Turing test in the present understanding of artificial intelligence. He discusses the proposed a gender imitation test by Sterrett 2000 and argues that the test is not an operational definition. He proceeds to discuss the progress that has been made on the Truing test through the Loebner contest. He concludes with three different arguments for the future of the Turing test: (1) the Intelligence Attribution Argument; (2) the Methodology Argument; and (3) the Visionary Argument. Overall, the article offers a critical analysis of what should be the role of the Turing test in the current understanding of AI.

- Thomas Hellström. 2013. "On the moral responsibility of military robots." *Ethics and Information Technology* 15, no. 2 (June): 99–107. ISSN: 1388-1957, 1572-8439, accessed May 25, 2018. doi:10.1007/s10676-012-9301-2. http://link.springer.com/10.1007/s10676-012-9301-2

  This article discusses mechanisms and principles for the assignment of moral responsibility to intelligent robots, with a special focus on military robots. Hellström introduces the concept of autonomous power as a new concept, and use it to identify the type of robots that call for moral considerations. It is furthermore argued that autonomous power, and in particular, the ability to learn, is decisive for the assignment of moral responsibility to robots. The article includes an overview of existing battlefield robots, classifies them based on autonomous power. analyzes in what way moral responsibility may be applicable to robots, and how it relates to autonomous power. Finally, it analyses assignments of moral responsibility in hypothetical war scenarios with and without robots. Hellström concludes that the introduced concept of autonomous power, defined as a combination of self-ruling and capacity for actions, interactions and decisions, extends the traditional concept of autonomy such that weapons and military robots can be classified in a meaningful way for ethical considerations. Furthermore, autonomous power seems to be a decisive factor when assigning moral responsibility to other agents.

- Robert Sparrow. 2007. "Killer Robots." *Journal of Applied Philosophy* 24 (1): 62–77. ISSN: 1468-5930, accessed May 25, 2018. doi:10.1111/j.1468-5930.2007.00346.x. http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-5930.2007.00346.x

  This paper considers the problem of responsibility attribution for the actions of autonomous systems, specific weapon systems that might be involved in an atrocity of the sort that would normally be described as a war crime. Sparrow argues that it is not satisfactory to hold either the persons who designed or programmed the system, the commanding officer who ordered its use, or the machine itself, for actions that occur in the course of the war. In other words, he remains sceptical whether an AI could ever be considered a full moral agent. In the subsequent discussion, he highlights the lack of clarity when it comes to the level of autonomy that various systems might have, as he sees autonomy and moral responsibility go hand in hand. Ultimately, he argues that the more these machines are held to be autonomous the less it seems that those who program or design them, or those who order them into action, should be held responsible for their actions. He concludes, that for the foreseeable future, the deployment of weapon systems controlled by artificial intelligence in warfare is therefore unfair either to potential casualties in the theatre of war or to the officer who will be held responsible for their use.

1.3 Responsibility: Does agenthood imply moral responsibility?

- Colin Allen, Gary Varner, and Jason Zinser. 2000. "Prolegomena to any future artificial moral agent." *Journal of Experimental & Theoretical Artificial Intelligence* 12, no. 3 (July 1): 251–261. ISSN: 0952-813X, accessed June 27, 2018. doi:10.1080/09528130050111428. https://doi.org/10.1080/09528130050111428

This paper surveys the ethical disputes that characterize the challenge of building artificial moral agents (AMAs). The authors identify two areas of disagreement when it comes to building AMAs: (1) what standards moral agents ought to follow; (2) what does it mean to be a moral agent. The paper outlines the Moral Turing Test and the Comparative Moral Turing Test, which could serve to determine the moral capacity of an AMA. They also discuss various moral frameworks, primarily the utilitarian and the deontological, and asses the difficulties for each in terms of serving as a good framework for AMAs. In addition to the ethical difficulties, they also consider computational challenges and limitations. They conclude that building a morally praiseworthy agent is essentially the task of giving it enough intelligence to assess the effects of its actions on sentient beings and to use those assessments to make appropriate choices. This paper is a good general overview of the different challenges to be faced while building AMAs within different ethical frameworks.

- Andreas Matthias. 2004. "The responsibility gap: Ascribing responsibility for the actions of learning automata." *Ethics Inf Technol* 6, no. 3 (September 1): 175–183. ISSN: 1388-1957, 1572-8439, accessed June 27, 2018. doi:10.1007/s10676-004-3422-1. http://link.springer.com/article/10.1007/s10676-004-3422-1

  Matthias argues that autonomous, learning machines, based on neural networks, genetic algorithms, and agent architectures, create a new situation, where the manufacturer/operator of the machine is in principle not capable of predicting the future machine behaviour any more, and thus cannot be held morally responsible or liable for it. He claims that if we want to avoid the injustice of holding men responsible for actions of machines over which they could not have sufficient control, we must find a way to address the responsibility gap in moral practice and legislation. The society must decide between not using this kind of machine any more (which is not a realistic option), or facing a responsibility gap, which cannot be bridged by traditional concepts of responsibility ascription. This paper is an accessible introduction to the problems of attributing responsibility to humans for machine learning and other algorithms.

- "When Hal Kills, Who's to Blame? Computer Ethics - Tufts Digital

Library." 2018. Accessed June 27. https://dl.tufts.edu/catalog/tufts:ddennett-1997.00011

In this excerpt, Dennett considers various criteria for artificial agents to gain moral responsibility. He particularly frames the problem in the context of HAL and its comparison to the Deep Blue. He makes some interesting references to other authors who have considered this issue, such as Clark or Damasio. One of the arguments is that an agent would have to become a higher-order intentional system, capable of framing beliefs about its own beliefs, desires about its desires, beliefs about its fears about its thoughts about its hopes, and so on. Although this is not a rigorous academic argument, it is an accessible and general introduction to the problem of moral responsibility of artificial agents.

- William Bechtel. 1985. "Attributing Responsibility to Computer Systems." *Metaphilosophy* 16 (4): 296–306. ISSN: 1467-9973, accessed July 5, 2018. doi:10.1111/j.1467-9973.1985.tb00176.x. http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9973.1985.tb00176.x

  While Moor (1979) has provided a convincing argument that computers are able to make a decision, Bechtel confronts the question of whether computers can be held responsible for their decision. Bechtel questions the common assumption that computers cannot be morally responsible for their actions, by exploring possible conditions under which we might attribute the responsibility to a computer system. He argues that responsibility must be directed at those designing and using the machines. In addition, Bechtel claims that an important reason to consider the moral status of computers might be that humans are unable to meet the responsibility that now rests on them. He bases a notable part of his argument about intentionality on the work of Dennett. In his account, all a system needs in order to be teleological is to be appropriately embedded within its environment so as to respond to demands of that environment. In the end, he highlights on connectionist or parallel distributed processing systems, as possible systems that could potentially overcome these challenges, which might be explainable by the historical context of this paper. He concludes that currently available computers do not meet the conditions for being responsible agents. Despite the immense challenges of making a machine responsible, he still emphasizes the need of attempting to make such a machine.

- Bernhard Irrgang. 2006. "Ethical Acts in Robotics." *Ubiquity* 2006

(September): 1:2–1:16. ISSN: 1530-2180, accessed August 13, 2018. doi:10.1145/1164069.1164071. http://doi.acm.org/10.1145/1164069.1164071

Irrgang considers the question whether computers/robots can act morally. He conducts his analysis within the continental tradition, in particular in reference to the work of Paul Ricoeur. Overall, this article might be an example of resolving the issues arising from an increasingly autonomous computational system. Irrgang concludes that the difference between humans and robots cannot be erased due to human physicality, however, cautions against the trouble that might arise in the case of cyborgs.

1.4 Rights: Does agenthood warrant some type of rights

- Blay Whitby. 2008. “Sometimes it’s hard to be a robot: A call for action on the ethics of abusing artificial agents.” *Interacting with Computers* 20, no. 3 (May): 326–333. ISSN: 09535438, accessed August 4, 2018. doi:10.1016/j.intcom.2008.02.002. https://academic.oup.com/iwc/article-lookup/doi/10.1016/j.intcom.2008.02.002

  This paper is a call for an informed debate on the ethical issues raised by the forthcoming widespread use of robots, particularly in domestic settings. Whitby lists three questions that the designers of robots need to take an ethical stance on: (1) is it acceptable to treat artefacts (particularly human-like artefacts) in ways which we would consider morally unacceptable to treat humans?; (2) if so, just how much sexual or violent ’abuse’ of an artificial agent should we allow before we censure the behaviour of the abuser?; and (3) is it ethical for designers to attempt to ‘design out’ abusive behaviour by users? Whitby discusses these questions and concludes with various suggestions that might be adopted to prevent negative consequences. One of such is the proposition that it "may often be unethical to build robots that are inappropriately pleasant to their users". Overall, this argument is very application and industry oriented.

- Mark Coeckelbergh. 2014. “The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics.” *Philosophy and Technology* 27 (1): 61–77

  Coeckelbergh explores the implications of a relational approach to moral standing for thinking about machines, in particular autonomous, intelligent robots. His approach focuses on moral relations and on the conditions of the possibility of moral status ascription. Moreover, he provides a way to take a critical distance from what he calls the "standard" approach to thinking about moral status and moral standing, which is based on properties. It does not only overcome epistemological problems with the standard approach, but can also explain how we think about, experience, and act towards machines—including the gap that sometimes occurs between reasoning and experience. He also articulates the non-Cartesian orientation of his "relational" research program and specifies the way it contributes to a different paradigm in thinking about moral standing and moral knowledge.

- Mark Coeckelbergh. 2010. "Robot rights? Towards a social-relational justification of moral consideration." *Ethics and Information Technology* 12, no. 3 (September): 209–221. ISSN: 1388-1957, 1572-8439, accessed August 4, 2018. doi:10.1007/s10676-010-9235-5. http://link.springer.com/10.1007/s10676-010-9235-5

  Coeckelbergh discusses whether we should grant rights to artificially intelligent robots. He argues that most current and near-future robots do not meet the hard criteria set by deontological and utilitarian theory. In his account, virtue ethics can avoid this problem with its indirect approach. However, both direct and indirect arguments for moral consideration rest on ontological features of entities, an approach which incurs several problems. In response to these difficulties, Coeckelbergh proposes a framework which could grant some degree of moral consideration to some intelligent social robots: he sketches an argument for moral consideration based on social relations. It is suggested that we need a social ecology, which may be developed by engaging with Western ecology and Eastern worldviews. Although this relational turn raises many difficult issues and requires more work, this paper provides a rough outline of an alternative approach to moral consideration that can assist us in shaping our relations to intelligent robots.

2 Ethical norms: Which norms should govern autonomous systems?

2.1 Human vs. Non-Human

- Colin Allen. 2002. “Calculated morality: Ethical computing in the limit.” *Cognitive, emotive and ethical aspects of decision making and human action* 1:19–23

  Allen discusses issues with “calculated” morality. By "calculated" he means a tendency to deliberately exploit the altruism of others in order to further his own well-being. He argues that we should worry that the calculated nature of decisions made by any artificially intelligent system will result in behaviour that is not recognizably moral. He proposes two ways to surmount this problem. One is to try to write high-level rules that could be used to filter out problematic behaviours. Kant’s categorical imperative is an example in this category. The other approach is to abandon the search for explicitly rule-based approaches and seek other means of guiding behaviour. For example, providing artificial moral agents with human-like emotions might serve to keep certain anti-social tendencies in check. Each approach has strengths, but both are deeply problematic. He concludes that we should expect that our artificial moral agents will be as subject to making ethical mistakes as the next person.

- Wendell Wallach and WW Associates. n.d. “Artificial Morality: Bounded Rationality, Bounded Morality and Emotions”: 5

  Wallach discusses whether some of the help we derive in making decisions from emotions and an understanding of the semantic content of values, function as compensations for our limited ability to comprehensively analyze challenges we face. He claims that emotions and values are essential to facilitate the bounded rationality of human beings, but will be less essential, and often unnecessary to the calculated morality of artificial moral agents. This, of course, presumes that the potential comprehensive rationality of a computer is truly functional in meeting challenges fraught with real-world tensions, and not merely limited to the bounded moral environments in which artificial moral agents will initially act. The paper overall takes on mostly a predictive rather than normative stance, however, it outlines well some of the problems faced when trying to account for emotions and other human-like qualities in artificially moral systems.

- Steve Torrance. 2014. “Artificial Consciousness and Artificial Ethics: Between Realism and Social Relationism.” *Philos. Technol.* 27, no. 1 (March 1): 9–29. ISSN: 2210-5433, 2210-5441, accessed August 5, 2018. doi:10.1007/s13347-013-0136-5. http://link.springer.com/article/10.1007/s13347-013-0136-5

  Torrance compares a ‘realist’ with a ‘social–relational’ perspective on our judgments of the moral status of artificial agents (AAs). He develops a realist position according to which the moral status of a being (particularly in relation to moral patiency attribution) is closely bound up with that being’s ability to experience states of conscious satisfaction or suffering (CSS). For a realist, both moral status and experiential capacity are objective properties of agents. A social relationist denies the existence of any such objective properties in the case of either moral status or consciousness, suggesting that the determination of such properties rests solely upon social attribution or consensus. Torrance claims that a wide variety of social interactions between us and various kinds of artificial agents will proliferate in future generations, and the social–relational view may well be right that the appearance of CSS features in such artificial beings will make moral role attribution socially prevalent in human–AA relations. But there is still the question of what actual CSS states a given AA is capable of undergoing, independently of the appearances. This is not just a matter of changes in the structure of social existence that seem inevitable as human–AA interaction becomes more prevalent. The social world is itself enabled and constrained by the physical world, and by the biological features of living social participants. Properties analogous to certain key features in biological CSS are what need to be present for nonbiological CSS. Working out the details of such features will be an objective scientific inquiry. Torrence proposes a list of such features. Overall, the paper contains a thorough discussion of the realist and social-relational position and an account of issues of attributing moral responsibility to moral agents.

- Eric Dietrich. 2001. “Homo sapiens 2.0: why we should build the better robots of our nature.” *Journal of Experimental & Theoretical Artificial Intelligence* 13, no. 4 (October): 323–328. ISSN: 0952-813X, 1362-3079, accessed August 13, 2018. doi:10.1080/09528130110100289. http://www.tandfonline.com/doi/abs/10.1080/09528130110100289

  In this essay, Dietrich argues that since robots are not limited by the same evolutionary factors as humans, we can build them to be better moral agents than humans. He motivates his argument with the hypothesis that humans are bad in part because of our evolutionary history. He considers four cases: child abuse, sexism, rape and racism. In each, he provides an outline of how such behaviour could have been caused evolutionary. Subsequently, he proposes that we should not just try to "teach our children to behave better", but that we should implement in robots the best of our moral theories. These are the theories that see morality as comprising universal truths, applying fairly to all sentient beings. He concludes by considering three various counter-arguments to his position.

2.2 Deontological vs. Utilitarian

- Sven Nyholm and Jilles Smids. 2016. “The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?” *Ethical Theory and Moral Practice* 19 (5): 1275–1289

  The authors explore the analogy between the accident-scenarios of self-driving cars and the dilemmas associated with the trolley problem. They identify three important ways in which the ethics of accident- algorithms for self-driving cars and the philosophy of the trolley problem differ from each other. These concern: (i) the basic decision-making situation faced by those who decide how self- driving cars should be programmed to deal with accidents; (ii) moral and legal responsibility; and (iii) decision-making in the face of risks and uncertainty. The authors cite Lin (2015, 78), Wallach and Allen (2009, 14), and Bonnefon et al. (2015, 3), as examples of literature where there is an equation made between the trolley problem and accident-scenarios of self-driving. Among other things, they highlight the role of uncertainty and limitations of inference in the case of self-driving cars in comparison to the trolley problem dilemma. Overall, the paper provides a good general overview of the reasons why the trolley problem does not map well on the accident-scenarios of self-driving cars, as other writers might have expected.

- Luciano Floridi and Jeff W. Sanders. 2001. “Artificial evil and the foundation of computer ethics.” *Ethics and Information Technology* 3 (1): 55–66

  The authors outline two traditionally distinguished types of evil: moral (ME) and natural (NE). ME is the product of human agency (e.g. war, torture, psychological cruelty) while NE is the product of nonhuman agency (e.g. floods, famine, disease). Then there are also combinations of ME and NE. They argue that we need to consider autonomous agents as a new type of evil. They refer to it as artificial evil (AE). The AE forms a component of the foundation of Computer Ethics (CE). The behaviour of artificial agents in cyberspace is one of the primary concerns of CE. The behaviour of artificial agents can be morally good or evil even in the absence of biologically sentient participants and thus allows artificial agents not only to perpetrate evil (and for that matter good) but conversely to 'receive' or 'suffer from' it. The thesis defended is that the notion of entropy structure, which encapsulates human value judgment concerning cyberspace in a formal mathematical definition, is sufficient to achieve this purpose and, moreover, that the concept of AE can be determined formally, by mathematical methods. Based on this thesis, the authors propose a theory called Information Ethics (IE). It is argued that the uniqueness of IE is justified by its being non-biologically biased and patient-oriented: IE is an Environmental Macroethics based on the concept of a data entity rather than life. Finally, they relate IE back to CE.

- Colin Allen, Iva Smit, and Wendell Wallach. 2005. "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches." *Ethics Inf Technol* 7, no. 3 (September 1): 149–155. ISSN: 1388-1957, 1572-8439, accessed August 5, 2018. doi:10.1007/s10676-006-0004-4. http://link.springer.com/article/10.1007/s10676-006-0004-4

  This paper discusses strategies for implementing artificial morality and the differing criteria for success that are appropriate to different strategies. In particular, the authors discuss the philosophical roots and computational possibilities of top-down and bottom-up strategies for designing artificial moral agents (AMAs). This paper directly builds on a paper by Allen et al. (2000). By top-down approaches, the authors mean an implementation of rules, based on moral principles and theories, to the selection of appropriate actions of AMAs. Among they discuss various deontological and utilitarian approaches. By 'bottom-up' approaches to the development of AMAs, they mean those that do not impose a specific moral theory, but which seek to provide environments in which appropriate behaviour is selected or rewarded. The paper contains a discussion of various bottom-up approaches, as well as hybrid approaches. The authors conclude with questions regarding the evaluation of machine and morality.

2.3 Particularist vs. Universalist

- Vincent Conitzer et al. 2017. “Moral Decision Making Frameworks for Artificial Intelligence.” In *AAAI,* 4831–4835

  The authors discuss possibilities for a general decision-making theory, similar to the generality and domain-independence of theories in decision and game theory. They argue that moral dilemmas will need to be more abstractly represented, and as is generally the case in AI research, the choice of representation scheme is extremely important. They consider two paradigms for designing general moral decision-making methodologies: extending game-theoretic solution concepts to incorporate ethical aspects and using machine learning on human-labelled instances. They conclude that the machine learning approach to automating moral judgments is perhaps more flexible than a game-theoretic approach, but the two can complement each other.

- M. Guarini. 2006. “Particularism and the Classification and Reclassification of Moral Cases.” *IEEE Intelligent Systems* 21, no. 4 (July): 22–28. ISSN: 1541-1672. doi:10.1109/MIS.2006.76

  Guarini confronts the following question: “Is it possible to classify cases as morally acceptable or unacceptable without using moral principles?.” The recent debate on principle versus case-based approaches to moral reasoning has come under the headings generalism and particularism, respectively. Particularism is often defined in terms of its rejection of moral principles. The philosopher Jonathan Dancy has suggested that moral reasoning (including learning) could be done without using moral principles and that neural network models could help us understand how to do this. Guarini proposes a neural network model of classification and explores the possibility of case-based moral reasoning (including learning) without recourse to moral principles. Implementation results show that nontrivial case classification might be possible but reclassification is more problematic.

- John Mikhail. 2007. “Universal moral grammar: theory, evidence and the future.” *Trends in Cognitive Sciences* 11, no. 4 (April): 143–152. ISSN: 13646613, accessed August 5, 2018. doi:10.1016/j.tics.2006.12.007. http://linkinghub.elsevier.com/retrieve/pii/S1364661307000496

  This article discusses approaches to the psychology and biology of human morality. In particular, it discusses universal moral grammar

(UMG). UMG seeks to describe the nature and origin of moral knowledge by using concepts and models similar to those used in Chomsky's program in linguistics. This approach is thought to provide a fruitful perspective from which to investigate moral competence from computational, ontogenetic, behavioural, physiological and phylogenetic perspectives. In this article, Mikhail outlines a framework for UMG and describe some of the evidence that supports it. Mikhail also proposes a novel computational analysis of moral intuitions and argue that future research on this topic should draw more directly on legal theory.

- Jeroen Van Den Hoven. 1997. "Computer Ethics and Moral Methodology." *Metaphilosophy* 28, no. 3 (July 1): 234–248. ISSN: 1467-9973, accessed August 5, 2018. doi:10.1111/1467-9973.00053. http://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9973.00053

  The premise of this article is that in computer ethics, as in other branches of applied ethics, the problem of the justification of moral judgment is still unresolved. Hoven argues that the method which is referred to as "The Method of Wide Reflective Equilibrium" (WRE) offers the best solution to it. It does not fall victim to the false dilemma of having to choose either case-based particularist or principle-based universalist approaches to the problem of moral justification. He claims that WRE also provides the best model of practical moral reasoning available for computer ethics. It does not pretend to provide quasi-algorithmic procedures for moral decision making, but neither does it abandon the regulative ideal of communicative transparency in discursive public justification.

2.4 Epistemology: How should we answer these ethical questions?

- Jan Gogoll and Julian F. Müller. 2017. "Autonomous Cars: In Favor of a Mandatory Ethics Setting." *Science and Engineering Ethics* 23 (3): 681–700

  The authors confront the question whether we should implement a mandatory ethics setting (MES) for the whole of society or, whether every driver should have the choice to select his own personal ethics setting (PES). This problem is discussed in the context of the trolley problem and the authors argue that a PES would most likely result in a prisoner's dilemma. They argue that MES in the better interest of everybody. Furthermore, they make the case that the classic trolley problem is conceptually inadequate for discussing the case of ethics settings. The reason for this is that the trolley problem fails to model three important structural aspects of the traffic dilemma discussed: strategic interaction, iteration as well as the varying position an individual might occupy. Overall, this paper provides a general argument about how should self-driving cars deal with dilemma situation. Cites Sandberg and Bradshaw (2013) as the authors to argue for PES, while Lin (2014b) argues against.

- Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016a. "The social dilemma of autonomous vehicles." *Science* 352, no. 6293 (June 24): 1573–1576. ISSN: 0036-8075, 1095-9203, accessed June 17, 2018. doi:10.1126/science.aaf2654. arXiv: 1510.03346. http://arxiv.org/abs/1510.03346

  The authors argue that moral algorithms will need to accomplish three potentially incompatible objectives: being consistent, not causing public outrage, and not discouraging buyers. They argue to achieve these objectives, manufacturers and regulators will need psychologists to apply the methods of experimental ethics to situations involving AVs and unavoidable harm. They focus on whether an AV should save lives by sacrificing its owner, and provide insights into (i) the perceived morality of this self-sacrifice, (ii) the willingness to see this self-sacrifice being legally enforced, (iii) the expectations that AVs will be programmed to self-sacrifice, and (iv) the willingness to buy self-sacrificing AVs. Although this is an empirical study, it constitutes a well-formulated introduction of the problem intuitions with accident scenarios. In addition, it is a frequently cited study in the literature.

- Selmer Bringsjord. 2008. "Ethical robots: the future can heed us." *AI & Soc* 22, no. 4 (April 1): 539–550. ISSN: 0951-5666, 1435-5655, accessed August 13, 2018. doi:10.1007/s00146-007-0090-9. http://link.springer.com/article/10.1007/s00146-007-0090-9

  Bringsjord provides a criticism of Bill Joy's "Why the Future Doesn't Need Us." Although Joy discusses also generics and nanotechnology, Bringsjord in his critique focuses primarily on the part concerning robotics. He argues that Joy's justification of why we should fear robots and AI fails. Bringsjord analyses three of Joy's arguments about the dangers of robots. Firstly, that increasingly sophisticated machines will supersede humans and render them unnecessary. He claims that this argument is unsound and invalid. Secondly, he challenges Joy's fear of replicating robots. Thirdly, he questions the argument that humans will find it irresistible to download themselves into robotic bodies to prolong their life-span. Bringsjord concludes that we should not fear robots, but rather what some of us may do with robots. Overall, this paper is a succinct response to some of the commonly misconceived options about the dangers of AI, mainly stemming from the work of writers, such as Moravec or Kurzweil.

3 Human interaction: How should people approach autonomous systems?

3.1 Theory of mind: Should autonomous systems posses some type of theory of mind / empathy?

- R W Picard. 1997. "Affective Computing": 16

  This paper presents and discusses key issues in "affective computing," computing that relates to, arises from, or influences emotions. Picard postulates that those affective computers should not only provide better performance in assisting humans but also might enhance computers' abilities to make decisions. In this paper, the author defines important issues in affective computing. She suggests models for affect recognition, and present her ideas for new applications of affective computing to computer-assisted learning, perceptual information retrieval, arts and entertainment, and human health and interaction. She also describes how advances in affective computing, especially combined with wearable computers, can help advance emotion and cognition theory. Although is not primarily targeted at philosophy audience, it raises some of the ethical and empirical issues that come about with affective computing and provides a good overview of the discipline of affective computing.

- Rosalind W. Picard. 2003. "Affective computing: challenges." *International Journal of Human-Computer Studies* 59 (1): 55–64

  This article raises and responds to several criticisms of affective computing, articulating state-of-the-art research challenges. Picard responds to critique regarding sensing and recognizing emotion, affect modelling, emotion expression, ethics, and the utility of considering affect in HCI. This pare appears to be a valuable addition to the basics of affective computing outlined in Picard 1997.

- Michael Anderson and Susan Leigh Anderson. 2007. "Machine Ethics: Creating an Ethical Intelligent Agent." *AI Magazine* 28, no. 4 (December 15): 15. ISSN: 2371-9621, accessed August 5, 2018. `https://www.aaai.org/ojs/index.php/aimagazine/article/view/2065`

  In this article, the authors discuss the importance of machine ethics, the need for machines that represent ethical principles explicitly, and the challenges facing those working on machine ethics. They also give an example of current research in the field that shows that it is possible, at least in a limited domain, for a machine to abstract an ethical principle from examples of correct ethical judgments and use that principle to guide its own behaviour. They are primarily concerned with the ethical decision making itself, rather than how a machine would gather the information needed to make the decision and incorporate it into its general behaviour. They provide an example of their own approach to implementation of ethics in a machine. In addition, provide a discussion of various ethical frameworks suitable for such an implementation, as well as the role of emotions in machine ethics.

3.2 Attitude towards autonomous systems: How should we relate to autonomous systems?

- J. H. Moor. 2006. "The Nature, Importance, and Difficulty of Machine Ethics." *IEEE Intelligent Systems* 21, no. 4 (July): 18–21. ISSN: 1541-1672. doi:10.1109/MIS.2006.80

  Moor's paper is a general introduction to the problem of machine ethics. He claims that we can evaluate machines both in terms of design norms, as well as ethical norms. He outlines various instances in which we do need a machine to act ethically, such as a money transfer of flying a plane. One of the approaches which one could take in implementing machine ethics is to restrict machine's actions to avoid unethical outcomes. He claims that we should, therefore, focus on developing limited explicit ethical agents. Especially since more powerful machines need more powerful machine ethics. He offers three reasons why we cannot be too optimistic about our ability to develop machines which are going to be explicit ethical agents: (1) we have a limited understanding of what a proper ethical theory is; (2) we need to understand learning better; and (3) major problem might be computers' absence of common sense and world knowledge. He concludes by emphasizing the need to dedicate much more effort to making progress in this domain. Overall, the argument does not go into much depth but outlines some of the more practical issues faced when developing machine ethics.

- Clifford Nass and Youngme Moon. 2018. "Machines and Mindlessness: Social Responses to Computers." *Journal of Social Issues* 56 (1): 81–103. ISSN: 1540-4560, accessed June 27. doi:10.1111/0022-4537.00153. http://spssi.onlinelibrary.wiley.com/doi/abs/10.1111/0022-4537.00153

  This paper reviews a series of experimental studies that demonstrate that individuals mindlessly apply social rules and expectations to computers. The first set of studies illustrate how individuals over-use human social categories, applying gender stereotypes to computers and identifying with computer agents that share their ethnicity. The second set of experiments demonstrate that people exhibit over-learned social behaviours such as politeness and reciprocity with respect to computers. In the third set of studies, premature cognitive commitments are demonstrated: A television set labelled a specialist is perceived as providing better content than a television set that provides multiple types of content. A final series of studies demonstrate the depth of social responses with respect to computer "personality." Alternative explanations for these findings, such as anthropomorphism, intentional social responses, and demand characteristics, cannot explain the results. The authors conclude with an agenda for the next generation of research. Although this paper is not primarily philosophical, it provides an introduction to the study of human responses to computers.

- Peter M. Asaro. 2006. "What Should We Want From a Robot Ethic." *International Review of Information Ethics* 6 (12): 9–16

  Asaro argues that there are at least three things we might mean by "ethics in robotics": the ethical systems built into robots, the ethics of people who design and use robots, and the ethics of how people treat robots. He argues that the best approach to robot ethics is one which addresses all three of these, and to do this it ought to consider robots as socio-technical systems. By so doing, it is possible to think of a continuum of the agency that lies between amoral and fully autonomous moral agents. Thus, robots might move gradually along this continuum as they acquire greater capabilities and ethical sophistication. It also argues that many of the issues regarding the distribution of responsibility in complex socio-technical systems might best be addressed by looking to legal theory, rather than moral theory. This is because our overarching interest in robot ethics ought to be the practical one of preventing robots from doing harm, as well as preventing humans from unjustly avoiding responsibility for their actions. Specifically, he stresses the need for us to figure out how to relate to autonomous weapon systems and their moral standing. He relates his conclusions to authors, such as Allan at al. 2000. Overall, the paper offers a sparse set of references to literature from the field.

- R. Parasuraman, T. B. Sheridan, and C. D. Wickens. 2000. "A model for types and levels of human interaction with automation." *IEEE*

Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 30, no. 3 (May): 286–297. ISSN: 1083-4427. doi:10.1109/3468.844354

This paper aims to address the question concerning which system functions should be automated and to what extent. The authors outline a model for types and levels of automation that provides a framework and a basis for making such choices. They propose that automation can be applied to four broad classes of functions: 1) information acquisition; 2) information analysis; 3) decision and action selection, and 4) action implementation. Within each of these types, automation can be applied across a continuum of levels from low to high, i.e., from fully manual to fully automatic. A particular system can involve automation of all four types at different levels. The human performance consequences of particular types and levels of automation constitute primary evaluative criteria for automation design. Secondary evaluative criteria include automation reliability and the costs of decision/action consequences, among others. Examples of recommended types and levels of automation are provided to illustrate the application of the model to automation design. Although not specifically philosophical, this paper outlines some of the conceptual challenges in automation and human interaction with autonomous systems.

3.3 Mixed traffic: How should self-driving cars relate to human drivers?

- Sven Nyholm and Jilles Smids. n.d. "Automated Cars Meet Human Drivers: Responsible Human-Robot Coordination and The Ethics of Mixed Traffic." *Ethics and Information Technology*

  This paper discusses responsible human-robot coordination within mixed traffic: i.e. traffic involving both automated cars and conventional human-driven cars. They do three main things: (1) They explain key differences in robotic and human agency and expectation-forming mechanisms that are likely to give rise to compatibility-problems in mixed traffic, which may lead to crashes and accident; (2) They identify three possible solution-strategies for achieving better human-robot coordination within mixed traffic; (3) They identify important ethical challenges raised by each of these three possible strategies for achieving optimized human-robot coordination in this domain. Overall, this paper highlights well the problems that might arise within mixed traffic and proposes a general path to a solution and the development. The paper engages with the topic from a wide variety of perspectives, including the ethical, legal, and technological perspective. Cites Goodall 2014a, b; Lin 2015; Hevelke and Nida-Rümelin 2014; Gurney 2016; Gogoll and Müller 2016; Nyholm and Smids 2016; Nyholm forthcoming to define the field of ethics of automated driving.

- Milos N. Mladenovic and Tristram McPherson. 2016. "Engineering Social Justice into Traffic Control for Self-Driving Vehicles?" *Sci Eng Ethics* 22, no. 4 (August 1): 1131–1149. ISSN: 1353-3452, 1471-5546, accessed June 17, 2018. doi:10.1007/s11948-015-9690-9. http://link.springer.com/article/10.1007/s11948-015-9690-9

  This paper makes a case for the importance of addressing questions of social justice in this transformation and sketches a preliminary framework for doing so. The authors explain how new forms of traffic control technology have potential implications for several dimensions of social justice, including safety, sustainability, privacy, efficiency, and equal access. The central focus is on efficiency and equal access as desiderata for traffic control design. This paper is interesting because of its success in pointing out the problems and importance of ethical consideration of the broader network infrastructure that might be necessary for the development of autonomous vehicles.

- Quan Yuan, Yan Gao, and Yibing Li. 2016. "Suppose Future Traffic Accidents Based on Development of Self-driving Vehicles." In *Man-Machine-Environment System Engineering,* 253–261. Lecture Notes in Electrical Engineering. Springer, Singapore, October 21. ISBN: 978-981-10-2322-4 978-981-10-2323-1, accessed August 5, 2018. doi:10.1007/978-981-10-2323-1_28. http://link.springer.com/chapter/10.1007/978-981-10-2323-1_28

  This article discusses various forms of road accidents that might occur if there are more self-driving vehicles on the road. In particular, it focuses on the on the co-existence of humans and autonomous vehicles. They argue that collision avoidance should be focused on protecting vulnerable users. The paper analyzes and forecasts the possible complexity of the future accidents, and suggests that in future the characteristics of accidents form change should be revealed in depth, so as to work out plans for preventing and handling the related accidents. The analysis is not specifically philosophical. It considers the problematic from a variety of perspectives, including technological, social, or geographical.

4 Legal norms: What kind of legal norms should apply to autonomous systems, their creators, and users?

4.1 Regulation: Should society regulate development of autonomous systems and how?

- James A. Stieb. 2008. “A Critique of Positive Responsibility in Computing.” *Science and Engineering Ethics* 14, no. 2 (June): 219–233. ISSN: 1353-3452, 1471-5546, accessed August 4, 2018. doi:10.1007/s11948-008-9067-4. http://link.springer.com/10.1007/s11948-008-9067-4

  Stieb provides an argument about the general ethics that should govern the conduct of programmers. He opposes the stance that that (1) computer professionals should be held responsible for an undisclosed list of “undesirable events” associated with their work and (2) most if not all computer disasters can be avoided by truly understanding responsibility. Stieb claims that programmers, software developers, and other computer professionals should be defended against such vague, counterproductive, and impossible ideals because these imply the mandatory satisfaction of social needs and the equation of ethics with a kind of altruism. He discusses both the concept of “social needs” and “positive responsibility” and argues that both of these are unclear concepts with no authority to define and apply them. He concludes that insisting that every “bug” or “computer error” is an ethical lapse, runs the risk of confusing efficiency with ethics to the detriment of both. Stieb’s argument is specifically interesting in contrast to articles such as Johnson 2006.

- Bernd Carsten Stahl. 2006. “Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency.” *Ethics and Information Technology* 8, no. 4 (December 5): 205–213. ISSN: 1388-1957, 1572-8439, accessed August 4, 2018. doi:10.1007/s10676-006-9112-4. http://link.springer.com/10.1007/s10676-006-9112-4

  In this paper, Stahl confronts the question of whether computers can be responsible. Instead of approaching the subject in terms of agency or personhood, he addresses the question from the perspective of social responsibility. Stahl provides an analysis of the concept of social responsibility and argues that the analysis shows that it is a social construct of ascription which is only viable in certain social contexts and which serves particular social aims. If this is the main aspect of responsibility then the question whether computers can be responsible no longer hinges on the difficult problem of agency but on the possibly simpler question whether responsibility ascriptions to computers can fulfil social goals. As a solution, Stahl suggests the introduction of a new concept, called "quasi-responsibility" which will emphasize the social aim of responsibility ascription and which can be applied to computers. Overall, his paper offers a well-contextualized discussion of responsibility attribution and he juxtaposes his argument to other prominent researchers in the field (Floridi, Sanders, Johnson, etc.).

- P. A. Hancock. 2014. "Automation: how much is too much?" *Ergonomics* 57, no. 3 (March 4): 449–454. ISSN: 0014-0139, accessed August 6, 2018. doi:10.1080/00140139.2013.816375. https://doi.org/10.1080/00140139.2013.816375

  Hancock confronts the question of when is it for humans appropriate to automate. He suggests that unlimited automation of all technical functions will eventually prove anathema to the fundamental quality of human life. To support his claim, he provides examples of tasks, pursuits and past-times that should potentially be excused from the automation imperative. He concludes by discussing the question of balance in the cooperation, coordination and potential conflict between humans and the machines they create. Overall, this paper is not written for scholars in philosophy, but it provides a general discussion of the questions one might confront when considering the upsides and downsides of automation.

- F. S. Grodzinsky, K. Miller, and M. J. Wolf. 2012. "Moral Responsibility for Computing Artifacts: "the Rules" and Issues of Trust." *SIGCAS Comput. Soc.* 42, no. 2 (December): 15–25. ISSN: 0095-2737, accessed August 11, 2018. doi:10.1145/2422509.2422511. http://doi.acm.org/10.1145/2422509.2422511

  This paper uses the document called "The Rules" to examine the issues of trust. The Rules is a collaborative document (started in March 2010) that states principles for responsibility when a computer artefact is designed, developed and deployed into a sociotechnical system. The first part of this paper presents The Rules. The Rules document currently includes five rules that are intended to serve "as a normative guide for people who design, develop, deploy, evaluate or use computing artefacts." Next, the authors briefly examine a model of trust and the relationship between The Rules and society through the lens of trust. Then, they examine each rule with respect to the sociotechnical system and trust. The authors claim that the power and complexity of computing artefacts and the growing sophistication of sociotechnical systems require us to be more dependent on trust relationships, not less. In the last section of the paper, they illustrate their claim by applying The Rules to the paradigms of quantum and cloud computing. The paper offers an interesting approach to evaluating and instituting ethical systems, as well as they ground their argument in the relevant literature, such as the work of Floridi and Sanders.

4.2 Legal status: What should be the legal status of autonomous systems? Should we view autonomous systems as legal persons?

- Lawrence B. Solum. 1992. "Legal Personhood for Artificial Intelligences." *North Carolina Law Review* 70:1231

  In his essay, Solum confronts the question of legal personhood for artificial intelligence from the legal perspective. He explores this question through a series of thought experiments that transform the theoretical question whether artificial intelligence is possible into legal questions. The essay is divided into VI parts. Part I of this Essay recounts some recent developments in cognitive science and explores the debate as to whether artificial intelligence is possible. Part II puts the question in legal perspective by setting out the notion of legal personhood. Parts III and IV explore two hypothetical scenarios. Part III examines the first scenario-an attempt to appoint an Al as a trustee. The second scenario, an AI's invocation of the individual rights provisions of the United States Constitution, is the subject of Part IV. The results are then brought to bear on the debate over the possibility of artificial intelligence in Part V. In conclusion, Part VI takes up the question whether cognitive science might have implications for current legal and moral debates over the meaning of personhood.

- Peter M Asaro. n.d. "Robots and Responsibility from a Legal Perspective": 5

  This paper considers how legal theory, or jurisprudence, might be applied to robots. This is done with the intention of determining what concepts and approaches to robot ethics might be gained from taking a legal perspective. In many cases, legal theory is suggestive of possible approaches to problems that will require further work to evaluate. It concludes that legal theory does allow us to define certain classes of ethical problems that correspond to traditional and well-defined legal problems, while other difficult practical and meta-ethical problems cannot be solved by legal theory alone.

- Hin-Yan Liu. 2012. "Categorization and legality of autonomous and remote weapons systems." *International Review of the Red Cross* 94, no. 886 (June): 627–652. ISSN: 18163831, accessed August 5, 2018. doi:10.1017/S181638311300012X. http://proxy.lib.sfu.ca/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=88366770&site=ehost-live

  This article reconsiders the status and legality of both autonomous and remote weapons systems under international humanitarian law. Technologically advanced unmanned military systems are being introduced into the modern battlespace with insufficient recognition of their potential challenge to international humanitarian law. The article questions the understanding of both autonomous and remote weapons systems as 'weapons' and seeks to consider how their use may impact existing legal categories. Their use is then specifically situated to consider the legality of their deployment in certain contexts. Finally, the article raises the question of impunity for the use of both autonomous and remote weapons systems that arise from the inability to attribute responsibility for the harm they cause. In the author's view, it is imperative that law and policy are developed to govern the development and deployment of these advanced weapons systems to forestall these likely situations of impunity.

4.3 Legal norms: What legal norms should be applied to self-driving cars?

- Alexander Hevelke and Julian Nida-Rümelin. 2015. "Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis." *Science and Engineering Ethics* 21 (3): 619–630

  This article discusses who should be held responsible for accidents of fully autonomous cars from a moral standpoint. The authors argue that both the duty to intervene and a responsibility of the driver as a form of a "strict liability" seem like viable options. The paper is overall a rather general introduction to the problematic but makes some interesting arguments based on the capacity of the driver to intervene.

- J. Christian Gerdes and Sarah M. Thornton. 2016. “Implementable Ethics for Autonomous Vehicles.” In *Autonomous Driving,* 87–102. Springer, Berlin, Heidelberg. ISBN: 978-3-662-48845-4 978-3-662-48847-8, accessed June 17, 2018. doi:10.1007/978-3-662-48847-8_5. https://link.springer.com/chapter/10.1007/978-3-662-48847-8_5

  In this book chapter, the authors pose the question whether automated vehicles can be designed a priory to embody not only the laws but also the ethical principles of the society in which they operate. In particular, can ethical frameworks and rules derived for human behaviour be implemented as control algorithms in automated vehicles? They argue that direct analogies can be drawn between the frameworks of consequentialism and deontological ethics in philosophy and the use of cost functions or constraints in optimal control theory. In their account, the challenge then becomes determining which principles are best described as a comparative weighing of costs from a consequentialist perspective and which form the more absolute rules of deontological ethics. They also pose the question of whether it is sufficient for vehicles to simply try to avoid collisions. The authors claim it to be more actionable than avoiding harm and propose a set of rules, similar to those of Asimov’s three rules of robotics. The chapter concludes with examples of ethical constraints implemented as control laws and a reflection on whether human override and the ubiquitous “big red button” are consistent with an ethical automated vehicle.

- Sabine Gless, Emily Silverman, and Thomas Weigend. 2016. “If Robots Cause Harm, Who Is to Blame: Self-Driving Cars and Criminal Liability.” *New Crim. L. Rev.* 19:412–436. Accessed August 5, 2018. https://heinonline.org/HOL/P?h=hein.journals/bufcr19&i=418

  In this paper, the authors discuss the novel issues in criminal law raised by the increasing prevalence of robots, especially self-driving cars. Robots can malfunction and cause serious harm. But as things stand today, they are not suitable recipients of criminal punishment, mainly because they cannot conceive of themselves as morally responsible agents and because they cannot understand the concept of retributive punishment. Humans who produce, program, market, and employ robots are subject to criminal liability for an intentional crime if they knowingly use a robot to cause harm to others. A person who allows a self-teaching robot to interact with humans can foresee that the robot might get out of control and cause harm. This fact alone may give rise to negligence liability. In light of the social benefits associated with the use of many of today s robots, the authors argue in favour of limiting the criminal liability of operators to situations where they neglect to undertake reasonable measures to control the risks emanating from robots. They conclude that each society must answer for itself the question of whether investment in chances for a better life should be rewarded with an exemption from criminal responsibility for some of the risks involved and what these risks are. This article provides a succinct introduction to some of the legal and ethical issues connected to the legal norms that should be connected to the use of autonomous agents.

5 Limits of imagination

5.1 Trolley scenarios

- Jamy Li et al. 2016. “From Trolley to Autonomous Vehicle: Perceptions of Responsibility and Moral Norms in Traffic Accidents with Self-Driving Cars.” April 5. doi:10.4271/2016-01-0164

  This paper presents a study that assessed the participant’s moral intuitions in self-driving car scenarios. The study consists of two online experiments which assessed lay perceptions of moral norms and responsibility for traffic accidents involving autonomous vehicles. In Experiment 1, 120 US adults read a narrative describing a traffic incident between a pedestrian and a motorist. In different experimental conditions, the pedestrian, the motorist, or both parties were at fault. Participants assigned less responsibility to a self-driving car that was at fault than to a human driver who was at fault. Participants confronted with a self-driving car at fault allocated greater responsibility to the manufacturer and the government than participants who were confronted with a human driver at fault did. In Experiment 2, 120 US adults read a narrative describing a moral dilemma in which a human driver or a self-driving car must decide between either allowing five pedestrians to die or taking action to hit a single pedestrian in order to save the five. The “utilitarian” decision to hit the single pedestrian was considered the moral norm for both a self-driving and a human-driven car. Moreover, participants assigned the obligation of setting moral norms for self-driving cars to ethics researchers and to car manufacturers. Although empirical, this research reveals some patterns of public perception of autonomous cars and provides an example of an empirical study concerning the trolley problem.

- Selmer Bringsjord and Atriya Sen. 2016. “On Creative Self-Driving Cars: Hire the Computational Logicians, Fast.” *Applied Artificial Intelligence* 30, no. 8 (September 13): 758–786. ISSN: 0883-9514, accessed August 6, 2018. doi:10.1080/08839514.2016.1229906. https://doi.org/10.1080/08839514.2016.1229906

  The authors provide an argument for the need for computational logician on the engineering teams of self-driving vehicles. That is, the computational logicians must be recruited to design and implement logic that are connected to the operating-system level of the self-driving cars, and that ensure these cars meet all of their moral and legal obligations, never do what is morally or legally forbidden, invariably steer clear of the invidious, and, when appropriate, perform what is supererogatory.

- Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016b. "The social dilemma of autonomous vehicles." *Science* 352, no. 6293 (June 24): 1573–1576. ISSN: 0036-8075, 1095-9203, accessed August 6, 2018. doi:10.1126/science.aaf2654. http://science.sciencemag.org/content/352/6293/1573

  The authors posit that autonomous vehicles (AVs) should reduce traffic accidents, but they will sometimes have to choose between two evils, such as running over pedestrians or sacrificing themselves and their passenger to save the pedestrians. The authors found that participants in six Amazon Mechanical Turk studies approved of utilitarian AVs and would like others to buy them, but they would themselves prefer to ride in AVs that protect their passengers at all costs. The study participants disapprove of enforcing utilitarian regulations for AVs and would be less willing to buy such an AV. Accordingly, regulating for utilitarian algorithms may paradoxically increase casualties by postponing the adoption of a safer technology. Overall, this data-driven approach highlights how the field of experimental ethics can provide some insights into the moral, cultural, and legal standards that people expect from autonomous driving algorithms.

- Dennis Moberg and David F. Caldwell. 2007. "An Exploratory Investigation of the Effect of Ethical Culture in Activating Moral Imagination." *J Bus Ethics* 73, no. 2 (June 1): 193–204. ISSN: 0167-4544, 1573-0697, accessed August 6, 2018. doi:10.1007/s10551-006-9190-6. http://link.springer.com/article/10.1007/s10551-006-9190-6

  This study explores the factors that may engage decision makers in a morally imaginative decision process. Moral imagination is a process that involves a thorough consideration of the ethical elements of a decision. The authors sought to explore what might distinguish moral imagination from other ethical approaches within a complex business simulation. Using a three-component model of moral imagination, they sought to discover whether organization cultures with a salient ethics theme activate moral imagination. Finding an effect, they sought an answer to whether some individuals were more prone to being influenced in this way by ethical cultures. The authors found that employees with strong moral identities are less influenced by such cultures than employees whose sense of self is not defined in moral terms.

5.2 Other

- David Danks and Alex John London. 2017. “Algorithmic Bias in Autonomous Systems,” 4691–4697. International Joint Conferences on Artificial Intelligence Organization, August. ISBN: 978-0-9992411-0-3, accessed August 6, 2018. doi:10.24963/ijcai.2017/654. https://www.ijcai.org/proceedings/2017/654

  The authors of this paper discuss the role and different meanings of the term ‘bias’ in the discussion of algorithm analysis and design. They claim that the both purely descriptive, as well as pejorative uses of the term can promote confusion and hamper discussions about when and how to respond to algorithmic bias. In this paper, they first provide a taxonomy of different types and sources of algorithmic bias, with a focus on their different impacts on the proper functioning of autonomous systems. They then use this taxonomy to distinguish between algorithmic biases that are neutral or unobjectionable and those that are problematic in some way and require a response. In some cases, there are technological or algorithmic adjustments that developers can use to compensate for problematic bias. In other cases, however, responses require adjustments by the agent, whether a human or autonomous system, who uses the results of the algorithm. There is no “one size fits all” solution to algorithmic bias. They conclude that we need to think about algorithmic bias (with respect to various norms) in terms of the whole system, including the consumer–human or machine–of the algorithm output.

- Batya Friedman. 1995. ““It’s the computer’s fault”: reasoning about computers as moral agents,” 226–227. ACM Press. ISBN: 978-0-89791-755-1, accessed August 13, 2018. doi:10.1145/223355.223537. http://portal.acm.org/citation.cfm?doid=223355.223537

  Friedman presents a study about error attribution in human-computer interactions. Friedman posits that typically tool use poses few confusions about whom we understand to be the moral agent for a given act. But when the “tool” becomes a computer, do people attribute moral agency and responsibility to the technology (“it’s the computer’s fault”)? Twenty-nine mate undergraduate computer science majors were interviewed. Results showed that most students (83%) attributed aspects of agency—either decision-making and/or intentions—to computers. In addition, some students (21 %) consistently held computers morally responsible for an error. Discussion includes implications for computers system design.

- Youngme Moon and Clifford Nass. 1998. “Are computers scapegoats? Attributions of responsibility in human–computer interaction.” *International Journal of Human-Computer Studies* 49, no. 1 (July): 79–94. ISSN: 10715819, accessed August 13, 2018. doi:10.1006/ijhc.1998.0199. http://linkinghub.elsevier.com/retrieve/pii/S1071581998901999

  The paper presents a study investigating how people make attributions of responsibility when interacting with computers. In particular, two questions were addressed: Under what circumstances will users blame computers for failed outcomes? And under what circumstances will users credit computers for successful outcomes? The first prediction was that similarity between a user’s personality and a computer’s personality would reduce the tendency for users to exhibit a “self-serving bias” in assigning responsibility for outcomes in human-computer interaction. The second predication was that greater user control would lead to more internal attributions, regardless of the outcome. A 2 x 2 x 2 balanced, between-subjects experiment (N=80) was conducted. Results strongly supported the predictions: When the outcome was negative, participants working with a similar computer were less likely to blame the computer and more likely to blame themselves, compared to participants working with a dissimilar computer. When the outcome was positive, participants working with a similar computer were more likely to credit the computer and less likely to take the credit themselves, compared to participants working with a dissimilar computer. In addition, when users were given more control over outcomes, they tended to make more internal attributions, regardless of whether the outcome was positive or negative.

6 Books

- Wendell Wallach and Colin Allen. 2010. *Moral Machines: Teaching Robots Right from Wrong.* 1 edition. Oxford: Oxford University Press, June 3. ISBN: 978-0-19-973797-0

- Wendell Wallach. 2015. *A Dangerous Master: How to Keep Technology from Slipping Beyond Our Control.* New York: Basic Books, June 2. ISBN: 978-0-465-05862-4

- S. Hansson. 2013. *The ethics of risk: Ethical analysis in an uncertain world.* Springer. ISBN: 1-137-33365-0

- *Road vehicle automation.* 2014. New York: Springer. ISBN: 978-3-319-05989-1

- Nick Bostrom. 2014. *Superintelligence: Paths, Dangers, Strategies.* 1 edition. Oxford: Oxford University Press, September 3. ISBN: 978-0-19-967811-2

- Michael Anderson and Susan Leigh Anderson. 2011. *Machine ethics.* Cambridge University Press. ISBN: 1-139-49774-X

- Peter Danielson. 1992. *Artificial Morality: Virtuous Robots for Virtual Games.* Routledge

- Joseph C. Pitt. 1999. *Thinking About Technology: Foundations of the Philosophy of Technology.* New York: Seven Bridges Pr Llc, August 1. ISBN: 978-1-889119-12-0

- J. H. Fetzer. 2001. *Computers and Cognition: Why Minds are not Machines.* Studies in Cognitive Systems. Springer Netherlands. ISBN: 978-1-4020-0243-4, accessed July 5, 2018. `//www.springer.com/gp/book/9781402002434`

7 Syllabi

- Boyuan Chen. 2018. "Columbia: AI Safety, Ethics, and Policy." Accessed August 1. `http://www.cs.columbia.edu/~dechant/aisafety18/syllabus.html`
  `http://www.cs.columbia.edu/~dechant/aisafety18/index.html`

- Griswold. 2018. "Harvard Law School: The Ethics and Governance of Artificial Intelligence." Accessed August 1. `https://h2o.law.harvard.edu/playlists/53282`
  `https://h2o.law.harvard.edu/playlists/53282`

- Moritz Hardt. 2018. "UC Berkeley: Fairness in Machine Learning." Accessed August 1. `https://fairmlclass.github.io/`
  `https://fairmlclass.github.io/`

- Joi Ito. 2018. *MIT Media Lab: The Ethics and Governance of Artificial Intelligence*
  `https://www.media.mit.edu/courses/the-ethics-and-governance-of-artificial-intelligence/`

- Jerry Kaplan. 2018. "Stanford: Artificial Intelligence - Philosophy, Ethics, and Impact." Accessed August 1. `http://web.stanford.edu/class/cs122/`
  `http://web.stanford.edu/class/cs122/`

- Jan Leike. 2018. "80000hours: AI Safety Syllabus." 80,000 Hours. Accessed August 1. `https://80000hours.org/ai-safety-syllabus/`
  `https://80000hours.org/ai-safety-syllabus/`

- Daniel Schwartz et al. 2018. "Columbia: Computers and Society." Accessed August 1. `https://www.cs.columbia.edu/~smb/classes/s18/lectures.html`
  `https://www.cs.columbia.edu/~smb/classes/s18/index.html`

- Dr Mark Sprevak. n.d. "University of Edinburgh: Course aims and objectives": 11
  `http://www.drps.ed.ac.uk/17-18/dpt/cxphil10167.htm`

- Yulia Tsvetkov and Alan W Black. 2018. "CMU: Computational Ethics for NLP." accessed August 1. `http://demo.clab.cs.cmu.edu/ethical_nlp/#readings`
  `http://demo.clab.cs.cmu.edu/ethical_nlp/#summary`

- “UC Berkeley: Center for Human-Compatible AI.” 2018. Accessed August 2. `http://humancompatible.ai/bibliography`
`http://humancompatible.ai/bibliography`

Literature Cited

Allen, Colin. 2002. “Calculated morality: Ethical computing in the limit.” *Cognitive, emotive and ethical aspects of decision making and human action* 1:19–23.

Allen, Colin, Iva Smit, and Wendell Wallach. 2005. “Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches.” *Ethics and Information Technology* 7, no. 3 (September 1): 149–155. ISSN: 1388-1957, 1572-8439, accessed August 5, 2018. doi:10.1007/s10676-006-0004-4. http://link.springer.com/article/10.1007/s10676-006-0004-4.

Allen, Colin, Gary Varner, and Jason Zinser. 2000. “Prolegomena to any future artificial moral agent.” *Journal of Experimental & Theoretical Artificial Intelligence* 12, no. 3 (July 1): 251–261. ISSN: 0952-813X, accessed June 27, 2018. doi:10.1080/09528130050111428. https://doi.org/10.1080/09528130050111428.

Anderson, Michael, and Susan Leigh Anderson. 2007. “Machine Ethics: Creating an Ethical Intelligent Agent.” *AI Magazine* 28, no. 4 (December 15): 15. ISSN: 2371-9621, accessed August 5, 2018. https://www.aaai.org/ojs/index.php/aimagazine/article/view/2065.

———. 2011. *Machine ethics.* Cambridge University Press. ISBN: 1-139-49774-X.

Asaro, Peter M. 2006. “What Should We Want From a Robot Ethic.” *International Review of Information Ethics* 6 (12): 9–16.

Asaro, Peter M. n.d. “Robots and Responsibility from a Legal Perspective”: 5.

Bechtel, William. 1985. “Attributing Responsibility to Computer Systems.” *Metaphilosophy* 16 (4): 296–306. ISSN: 1467-9973, accessed July 5, 2018. doi:10.1111/j.1467-9973.1985.tb00176.x. http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9973.1985.tb00176.x.

Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. 2016a. “The social dilemma of autonomous vehicles.” *Science* 352, no. 6293 (June 24): 1573–1576. ISSN: 0036-8075, 1095-9203, accessed June 17, 2018. doi:10.1126/science.aaf2654. arXiv: 1510.03346. http://arxiv.org/abs/1510.03346.

Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. 2016b. "The social dilemma of autonomous vehicles." *Science* 352, no. 6293 (June 24): 1573–1576. ISSN: 0036-8075, 1095-9203, accessed August 6, 2018. doi:10.1126/science.aaf2654. http://science.sciencemag.org/content/352/6293/1573.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies.* 1 edition. Oxford: Oxford University Press, September 3. ISBN: 978-0-19-967811-2.

Bringsjord, Selmer. 2008. "Ethical robots: the future can heed us." *AI & SOCIETY* 22, no. 4 (April 1): 539–550. ISSN: 0951-5666, 1435-5655, accessed August 13, 2018. doi:10.1007/s00146-007-0090-9. http://link.springer.com/article/10.1007/s00146-007-0090-9.

Bringsjord, Selmer, and Atriya Sen. 2016. "On Creative Self-Driving Cars: Hire the Computational Logicians, Fast." *Applied Artificial Intelligence* 30, no. 8 (September 13): 758–786. ISSN: 0883-9514, accessed August 6, 2018. doi:10.1080/08839514.2016.1229906. https://doi.org/10.1080/08839514.2016.1229906.

Brustoloni, Jose C. 1991. *Autonomous Agents: Characterization and Requirements.*

Chen, Boyuan. 2018. "Columbia: AI Safety, Ethics, and Policy." Accessed August 1. http://www.cs.columbia.edu/~dechant/aisafety18/syllabus.html.

Coeckelbergh, Mark. 2010. "Robot rights? Towards a social-relational justification of moral consideration." *Ethics and Information Technology* 12, no. 3 (September): 209–221. ISSN: 1388-1957, 1572-8439, accessed August 4, 2018. doi:10.1007/s10676-010-9235-5. http://link.springer.com/10.1007/s10676-010-9235-5.

———. 2014. "The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics." *Philosophy and Technology* 27 (1): 61–77.

Conitzer, Vincent, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. 2017. "Moral Decision Making Frameworks for Artificial Intelligence." In *AAAI,* 4831–4835.

Danielson, Peter. 1992. *Artificial Morality: Virtuous Robots for Virtual Games.* Routledge.

Danks, David, and Alex John London. 2017. "Algorithmic Bias in Autonomous Systems," 4691–4697. International Joint Conferences on Artificial Intelligence Organization, August. ISBN: 978-0-9992411-0-3, accessed August 6, 2018. doi:10.24963/ijcai.2017/654. https://www.ijcai.org/proceedings/2017/654.

Dietrich, Eric. 2001. "Homo sapiens 2.0: why we should build the better robots of our nature." *Journal of Experimental & Theoretical Artificial Intelligence* 13, no. 4 (October): 323–328. ISSN: 0952-813X, 1362-3079, accessed August 13, 2018. doi:10.1080/09528130110100289. http://www.tandfonline.com/doi/abs/10.1080/09528130110100289.

Fetzer, J. H. 2001. *Computers and Cognition: Why Minds are not Machines.* Studies in Cognitive Systems. Springer Netherlands. ISBN: 978-1-4020-0243-4, accessed July 5, 2018. //www.springer.com/gp/book/9781402002434.

Floridi, Luciano, and J. W. Sanders. 2004. "On the Morality of Artificial Agents." *Minds and Machines* 14, no. 3 (August 1): 349–379. ISSN: 0924-6495, 1572-8641, accessed June 27, 2018. doi:10.1023/B:MIND.0000035461.63578.9d. http://link.springer.com/article/10.1023/B:MIND.0000035461.63578.9d.

Floridi, Luciano, and Jeff W. Sanders. 2001. "Artificial evil and the foundation of computer ethics." *Ethics and Information Technology* 3 (1): 55–66.

Franklin, Stan, and Art Graesser. 1996. "Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents." In *International Workshop on Agent Theories, Architectures, and Languages,* 21–35. Springer.

Friedman, Batya. 1995. ""It's the computer's fault": reasoning about computers as moral agents," 226–227. ACM Press. ISBN: 978-0-89791-755-1, accessed August 13, 2018. doi:10.1145/223355.223537. http://portal.acm.org/citation.cfm?doid=223355.223537.

Gerdes, J. Christian, and Sarah M. Thornton. 2016. "Implementable Ethics for Autonomous Vehicles." In *Autonomous Driving,* 87–102. Springer, Berlin, Heidelberg. ISBN: 978-3-662-48845-4 978-3-662-48847-8, accessed June 17, 2018. doi:10.1007/978-3-662-48847-8_5. https://link.springer.com/chapter/10.1007/978-3-662-48847-8_5.

Gless, Sabine, Emily Silverman, and Thomas Weigend. 2016. "If Robots Cause Harm, Who Is to Blame: Self-Driving Cars and Criminal Liability." *New Criminal Law Review* 19:412–436. Accessed August 5, 2018. https://heinonline.org/HOL/P?h=hein.journals/bufcr19&i=418.

Gogoll, Jan, and Julian F. Müller. 2017. "Autonomous Cars: In Favor of a Mandatory Ethics Setting." *Science and Engineering Ethics* 23 (3): 681–700.

Griswold. 2018. "Harvard Law School: The Ethics and Governance of Artificial Intelligence." Accessed August 1. https://h2o.law.harvard.edu/playlists/53282.

Grodzinsky, F. S., K. Miller, and M. J. Wolf. 2012. "Moral Responsibility for Computing Artifacts: "the Rules" and Issues of Trust." *SIGCAS Comput. Soc.* 42, no. 2 (December): 15–25. ISSN: 0095-2737, accessed August 11, 2018. doi:10.1145/2422509.2422511. http://doi.acm.org/10.1145/2422509.2422511.

Guarini, M. 2006. "Particularism and the Classification and Reclassification of Moral Cases." *IEEE Intelligent Systems* 21, no. 4 (July): 22–28. ISSN: 1541-1672. doi:10.1109/MIS.2006.76.

Hancock, P. A. 2014. "Automation: how much is too much?" *Ergonomics* 57, no. 3 (March 4): 449–454. ISSN: 0014-0139, accessed August 6, 2018. doi:10.1080/00140139.2013.816375. https://doi.org/10.1080/00140139.2013.816375.

Hansson, S. 2013. *The ethics of risk: Ethical analysis in an uncertain world.* Springer. ISBN: 1-137-33365-0.

Hardt, Moritz. 2018. "UC Berkeley: Fairness in Machine Learning." Accessed August 1. https://fairmlclass.github.io/.

Hayes, Patrick J., Kenneth M. Ford, and Neil Agnew. 1994. "On babies and bathwater: A cautionary tale." *AI magazine* 15 (4): 15.

Hellström, Thomas. 2013. "On the moral responsibility of military robots." *Ethics and Information Technology* 15, no. 2 (June): 99–107. ISSN: 1388-1957, 1572-8439, accessed May 25, 2018. doi:10.1007/s10676-012-9301-2. http://link.springer.com/10.1007/s10676-012-9301-2.

Hevelke, Alexander, and Julian Nida-Rümelin. 2015. "Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis." *Science and Engineering Ethics* 21 (3): 619–630.

Hoven, Jeroen Van Den. 1997. "Computer Ethics and Moral Methodology." *Metaphilosophy* 28, no. 3 (July 1): 234–248. ISSN: 1467-9973, accessed August 5, 2018. doi:10.1111/1467-9973.00053. http://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9973.00053.

Irrgang, Bernhard. 2006. "Ethical Acts in Robotics." *Ubiquity* 2006 (September): 1:2–1:16. ISSN: 1530-2180, accessed August 13, 2018. doi:10.1145/1164069.1164071. http://doi.acm.org/10.1145/1164069.1164071.

Ito, Joi. 2018. *MIT Media Lab: The Ethics and Governance of Artificial Intelligence.*

Johnson, Deborah G. 2006. "Computer systems: Moral entities but not moral agents." *Ethics and Information Technology* 8, no. 4 (December 5): 195–204. ISSN: 1388-1957, 1572-8439, accessed May 25, 2018. doi:10.1007/s10676-006-9111-5. http://link.springer.com/10.1007/s10676-006-9111-5.

Johnson, Deborah G., and Thomas M. Powers. 2001. "Computers as Surrogate Agents." In *Information Technology and Moral Philosophy,* edited by Jeroen van den Hoven and John Weckert, 251–269. Cambridge: Cambridge University Press. ISBN: 978-0-511-49872-5 978-0-521-85549-5 978-0-521-67161-3, accessed July 5, 2018. doi:10.1017/CBO9780511498725.014. https://www.cambridge.org/core/product/identifier/CBO9780511498725A020/type/book_part.

Kaplan, Jerry. 2018. "Stanford: Artificial Intelligence - Philosophy, Ethics, and Impact." Accessed August 1. http://web.stanford.edu/class/cs122/.

Kiss, George. 1990. "Autonomous Agents, AI and Chaos Theory."

Legg, Shane, and Marcus Hutter. 2007. "A Collection of Definitions of Intelligence." *arXiv:0706.3639 [cs]* (June 25). Accessed August 5, 2018. arXiv: 0706.3639. http://arxiv.org/abs/0706.3639.

Leike, Jan. 2018. "80000hours: AI Safety Syllabus." 80,000 Hours. Accessed August 1. https://80000hours.org/ai-safety-syllabus/.

Li, Jamy, Xuan Zhao, Mu-Jung Cho, Wendy Ju, and Bertram Malle. 2016. "From Trolley to Autonomous Vehicle: Perceptions of Responsibility and Moral Norms in Traffic Accidents with Self-Driving Cars." April 5. doi:10.4271/2016-01-0164.

Liu, Hin-Yan. 2012. “Categorization and legality of autonomous and remote weapons systems.” *International Review of the Red Cross* 94, no. 886 (June): 627–652. ISSN: 18163831, accessed August 5, 2018. doi:10.1017/S181638311300012X. http://proxy.lib.sfu.ca/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=88366770&site=ehost-live.

Matthias, Andreas. 2004. “The responsibility gap: Ascribing responsibility for the actions of learning automata.” *Ethics and Information Technology* 6, no. 3 (September 1): 175–183. ISSN: 1388-1957, 1572-8439, accessed June 27, 2018. doi:10.1007/s10676-004-3422-1. http://link.springer.com/article/10.1007/s10676-004-3422-1.

Mikhail, John. 2007. “Universal moral grammar: theory, evidence and the future.” *Trends in Cognitive Sciences* 11, no. 4 (April): 143–152. ISSN: 13646613, accessed August 5, 2018. doi:10.1016/j.tics.2006.12.007. http://linkinghub.elsevier.com/retrieve/pii/S1364661307000496.

Mladenovic, Milos N., and Tristram McPherson. 2016. “Engineering Social Justice into Traffic Control for Self-Driving Vehicles?” *Science and Engineering Ethics* 22, no. 4 (August 1): 1131–1149. ISSN: 1353-3452, 1471-5546, accessed June 17, 2018. doi:10.1007/s11948-015-9690-9. http://link.springer.com/article/10.1007/s11948-015-9690-9.

Moberg, Dennis, and David F. Caldwell. 2007. “An Exploratory Investigation of the Effect of Ethical Culture in Activating Moral Imagination.” *Journal of Business Ethics* 73, no. 2 (June 1): 193–204. ISSN: 0167-4544, 1573-0697, accessed August 6, 2018. doi:10.1007/s10551-006-9190-6. http://link.springer.com/article/10.1007/s10551-006-9190-6.

Moon, Youngme, and Clifford Nass. 1998. “Are computers scapegoats? Attributions of responsibility in human–computer interaction.” *International Journal of Human-Computer Studies* 49, no. 1 (July): 79–94. ISSN: 10715819, accessed August 13, 2018. doi:10.1006/ijhc.1998.0199. http://linkinghub.elsevier.com/retrieve/pii/S1071581998901999.

Moor, J. H. 2006. “The Nature, Importance, and Difficulty of Machine Ethics.” *IEEE Intelligent Systems* 21, no. 4 (July): 18–21. ISSN: 1541-1672. doi:10.1109/MIS.2006.80.

Moore, J. H. 2001. “The Turing Test: Past, Present and Future.” *Minds and Machines* 11 (1).

Nass, Clifford, and Youngme Moon. 2018. "Machines and Mindlessness: Social Responses to Computers." *Journal of Social Issues* 56 (1): 81–103. ISSN: 1540-4560, accessed June 27. doi:10.1111/0022-4537.00153. http://spssi.onlinelibrary.wiley.com/doi/abs/10.1111/0022-4537.00153.

Nyholm, Sven, and Jilles Smids. 2016. "The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?" *Ethical Theory and Moral Practice* 19 (5): 1275–1289.

———. n.d. "Automated Cars Meet Human Drivers: Responsible Human-Robot Coordination and The Ethics of Mixed Traffic." *Ethics and Information Technology.*

Parasuraman, R., T. B. Sheridan, and C. D. Wickens. 2000. "A model for types and levels of human interaction with automation." *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30, no. 3 (May): 286–297. ISSN: 1083-4427. doi:10.1109/3468.844354.

Picard, R W. 1997. "Affective Computing": 16.

Picard, Rosalind W. 2003. "Affective computing: challenges." *International Journal of Human-Computer Studies* 59 (1): 55–64.

Pitt, Joseph C. 1999. *Thinking About Technology: Foundations of the Philosophy of Technology.* New York: Seven Bridges Pr Llc, August 1. ISBN: 978-1-889119-12-0.

Road vehicle automation. 2014. New York: Springer. ISBN: 978-3-319-05989-1.

Schwartz, Daniel, Joshua Zweig, Joanne Kim, and Ikya Jupudy. 2018. "Columbia: Computers and Society." Accessed August 1. https://www.cs.columbia.edu/~smb/classes/s18/lectures.html.

Solum, Lawrence B. 1992. "Legal Personhood for Artificial Intelligences." *North Carolina Law Review* 70:1231.

Sparrow, Robert. 2007. "Killer Robots." *Journal of Applied Philosophy* 24 (1): 62–77. ISSN: 1468-5930, accessed May 25, 2018. doi:10.1111/j.1468-5930.2007.00346.x. http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-5930.2007.00346.x.

Sprevak, Dr Mark. n.d. "University of Edinburgh: Course aims and objectives": 11.

Stahl, Bernd Carsten. 2004. "Information, Ethics, and Computers: The Problem of Autonomous Moral Agents." *Minds and Machines* 14, no. 1 (February 1): 67–83. ISSN: 0924-6495, 1572-8641, accessed June 27, 2018. doi:10.1023/B:MIND.0000005136.61217.93. http://link.springer.com/article/10.1023/B:MIND.0000005136.61217.93.

———. 2006. "Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency." *Ethics and Information Technology* 8, no. 4 (December 5): 205–213. ISSN: 1388-1957, 1572-8439, accessed August 4, 2018. doi:10.1007/s10676-006-9112-4. http://link.springer.com/10.1007/s10676-006-9112-4.

Steels, Luc. 1995. "When are robots intelligent autonomous agents?" ISSN: 10.1016/0921-8890(95)00011-4, accessed August 5, 2018. https://digital.csic.es/handle/10261/127979.

Stieb, James A. 2008. "A Critique of Positive Responsibility in Computing." *Science and Engineering Ethics* 14, no. 2 (June): 219–233. ISSN: 1353-3452, 1471-5546, accessed August 4, 2018. doi:10.1007/s11948-008-9067-4. http://link.springer.com/10.1007/s11948-008-9067-4.

Sullins, John P. 2006. "When is a robot a moral agent?"

Torrance, Steve. 2014. "Artificial Consciousness and Artificial Ethics: Between Realism and Social Relationism." *Philosophy & Technology* 27, no. 1 (March 1): 9–29. ISSN: 2210-5433, 2210-5441, accessed August 5, 2018. doi:10.1007/s13347-013-0136-5. http://link.springer.com/article/10.1007/s13347-013-0136-5.

Tsvetkov, Yulia, and Alan W Black. 2018. "CMU: Computational Ethics for NLP." Accessed August 1. http://demo.clab.cs.cmu.edu/ethical_nlp/#readings.

"UC Berkeley: Center for Human-Compatible AI." 2018. Accessed August 2. http://humancompatible.ai/bibliography.

Wallach, Wendell. 2015. *A Dangerous Master: How to Keep Technology from Slipping Beyond Our Control.* New York: Basic Books, June 2. ISBN: 978-0-465-05862-4.

Wallach, Wendell, and Colin Allen. 2010. *Moral Machines: Teaching Robots Right from Wrong.* 1 edition. Oxford: Oxford University Press, June 3. ISBN: 978-0-19-973797-0.

Wallach, Wendell, and WW Associates. n.d. "Artificial Morality: Bounded Rationality, Bounded Morality and Emotions": 5.

"When Hal Kills, Who's to Blame? Computer Ethics - Tufts Digital Library." 2018. Accessed June 27. https://dl.tufts.edu/catalog/tufts:ddennett-1997.00011.

Whitby, Blay. 2008. "Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents." *Interacting with Computers* 20, no. 3 (May): 326–333. ISSN: 09535438, accessed August 4, 2018. doi:10.1016/j.intcom.2008.02.002. https://academic.oup.com/iwc/article-lookup/doi/10.1016/j.intcom.2008.02.002.

Yuan, Quan, Yan Gao, and Yibing Li. 2016. "Suppose Future Traffic Accidents Based on Development of Self-driving Vehicles." In *Man-Machine-Environment System Engineering,* 253–261. Lecture Notes in Electrical Engineering. Springer, Singapore, October 21. ISBN: 978-981-10-2322-4 978-981-10-2323-1, accessed August 5, 2018. doi:10.1007/978-981-10-2323-1_28. http://link.springer.com/chapter/10.1007/978-981-10-2323-1_28.