Online Ethics Center
FOR ENGINEERING AND SCIENCE

Robot Morality and Artificial Moral Agency Bibliography

Author(s)

Kelly Laas

Year

2020

Description

A bibliography looking at issues of robot morality and artificial intelligence and moral agency.

Body

Websites

International Committee for Robot Arms Control **https://www.icrac.net/**
This organization brings together experts in the area of robotics technology, artificial intelligence, robot ethics, government relations and international security who are concerned about the pressing dangers military robots pose to peace and international security and to civilians in war.

Moral Machines Blog **http://moralmachines.blogspot.com/**
Written by Wendell Wallach (author of 2012 book, Moral Machines*) and Colin Allen, this blog looks at the theory and development of artificial moral agents and*

computational ethics.

Books

Arkin: Robert. (2009). Governing lethal behavior in autonomous robots. Boca rattan: CRC Press.
Drawing from his work with the U.S. Army Research Office, DARPA, and other defence contractors, the author explores how to produce an artificial conscience in robots that may allow them to perform more ethically than humans in the battlefield. The author looks at the philosophical basis, motivation, theory and design recommendations for the implementation of artificial moral agency.

Krishman, Armin. (2009). Killer robots: Legality and ethicality of autonomous weapons. Burlington, V.T.: Ashgate.
Military robots and other, potentially autonomous robotic systems such as unmanned combat air vehicles (UCAVs) and unmanned ground vehicles (UGVs) could soon be introduced to the battlefield. Although the current technological issues will no doubt be overcome, the greatest obstacles to automated weapons on the battlefield are likely to be legal and ethical concerns. Armin Krishnan explores the technological, legal and ethical issues connected to combat robotics, examining both the opportunities and limitations of autonomous weapons. He also proposes solutions to the future regulation of military robotics through international law.

Lin, Patrick, Keith Abney, and George A. Bekey. (2011). Robot Ethics: the ethical and social implications of robotics. MIT Press.
Starting with an overview of ethical issues raised by the growing use of robots in our lives and relevant ethical theories, this book looks at the possibility of programming robot ethics to the ethical use of military robots of ware to issues of privacy and liability in the use of robots. The authors also look at the implications of using robots as sexual partners, caregivers, and servants. Finally, the authors explore if robots should be given rights or moral consideration.

Lin, Patrick, Keith Abreny, and Ryan Jenkins. Robot Ethics 2.0. New York: Oxford University Press.
Gathering together both old and new voices in the debate about robot ethics, this

volume seeks to explore the interdisciplinary and international discussion surrounding this issue. It focuses on the case study of autonomous cars as a way to look at diverse issues from liability to psychology that this subject touches.

Singer, P.W. 2009. Wired for war: The robotics revolution and conflict in the 21st century. New York: Penguin.
This book looks at some of the major changes military technology how these developments will change not only how wars are fought, but also the politics, economics, law and ethics that surround war itself.

Tzafestas, Spyros G. 2016. Robotics: A Navigating Overview. New York: Springer.
This volume explores the ethical questions that arise in the development, creation and use of robots that are capable of semiautonomous or autonomous decision making and human-like action. It examines how ethical and moral theories can and must be applied to address the complex and critical issues of the application of these intelligent robots in society.

Wallach, Wendell. (2009). Moral machines. Teaching robots right from wrong. New York, Oxford University Press.
Examines the challenge of building artificial moral agents, and argues that even if a fully moral machine is a long way off, it is necessary to start building a kind of functional morality that allows artificial moral agents to have some basic ethical sensitivity.

Journal Article

Allen, Colin, Iva Smit, and Wendell Wallach. (2015). Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology*. 7(3): 149-155. doi: 10.1007/s10676-006-0004-4.
A principal goal of the discipline of artificial morality is to design artificial agents to act as if they are moral agents. Intermediate goals of artificial morality are directed at building into AI systems sensitivity to the values, ethics, and legality of activities. The goal of this paper is to discuss strategies for implementing artificial morality and the differing criteria for success that are appropriate to different strategies.

Arkin, Ronald C. (2009). Ethical robots in warfare. *IEEE Technology and Society Magazine*. 28(1): 30-33. doi: 10.1109/MTS.2009.931858.
This article argues that robots not only can be better than soldiers in conducting warfare in certain circumstances, but they also can be more humane in the battlefield than humans. As robots can be built that do not exhibit fear, anger, frustration, or revenge, and that ultimately (and the key word here is ultimately) behave in a more humane manner than even human beings in these harsh circumstances and severe duress. People have not evolved to function in these conditions, but robots can be engineered to function well in them.

Asaro, Peter. (2009). Modelling the moral user. I*EEE Technology & Society Magazine.* 28(1): 20-24. doi: 10.1109/MTS.2009.931863.
Discusses the ethical design and regulation of autonomous lethal robots amid global concerns, interests, and justifications in the U.S.

Ashrafian, Hutan. (2015). AIonAI: Humanitarian Law of Artificial Intelligence and Robotics. *Science and Engineering Ethics.* 31(1) 29-40. doi: 10.1007/s11948-013-9513-9.
This paper focuses on ethical issues concerning the moral nature of robot-robot interactions. A new robotic law is proposed and termed AIonAI or artificial intelligence-on-artificial intelligence. This law tackles the overlooked area where future artificial intelligences will likely interact amongst themselves, potentially leading to exploitation. As such, they would benefit from adopting a universal law of rights to recognise inherent dignity and the inalienable rights of artificial intelligences. Such a consideration can help prevent exploitation and abuse of rational and sentient beings, but would also importantly reflect on our moral code of ethics and the humanity of our civilisation.

Borenstein, Jason and Ron Arkin. (2014). Robotic Nudges: The ethics of engineering a more socially just human being. *Science and Engineering Et hics*. doi: 10.1007/s11948-015-9636-2.

In this paper, the authors discuss whether companion robots should be permitted to “nudge” their human users in the direction of being “more ethical”. The authors use Rawlsian principles of justice to illustrate how robots might nurture “socially just” tendencies in their human counterparts. Designing technological artifacts in such a way to influence human behavior is already well-established but merely because the practice is commonplace does not necessarily resolve the ethical issues associated

with its implementation.

Coeckelbergh, Mark. (2009). Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society.* 24(2): 181-189. doi: 10.1007/s00146-009-0208-3.

This article seeks to answer a host of questions about non-human artificial entities such as robots and intelligent information systems. Sometimes they are called 'artificial agents'. But are they agents at all? And if so, should they be considered as moral agents and be held morally responsible? They do things to us in various ways, and what happens can be and has to be discussed in terms of right and wrong, good or bad. But does that make them agents or moral agents? And who is responsible for the consequences of their actions? The designer? The user? The robot?

Coeckelbergh, Mark. (2010). Moral Appearances, Emotions, Robots, and Human Morality. *Ethics and Information Technology.* 12(3): 235-241. doi: 10.1007/s10676-010-9221-y.

Though it is unlikely that we can ever build fully 'moral robots', as morality depends on emotions, we might nevertheless be able to build quasi-moral robots that can learn to create the appearance of emotions and the appearance of being fully moral. The article looks at how current robots do not meet standard necessary conditions for having emotions, and how it is unlikely that we can ever establish whether robots satisfy these conditions. The author then looks how this way of drawing robots into our social-moral world is less problematic than it might first seem, since human morality also relies on such appearances.

Davies, S. (2009). It's war – but not as we know it [autonomous military robotics] *Engineering & Technology.* 4(9): 40-43.

Intelligent machines deployed on battlefields around the world-from mobile grenade launchers to rocket-firing drones-can already identify and lock onto targets without human help. Currently, a human hand is always behind this technology, but this could change in the near future. To the extent that military robots can considerably reduce unethical conduct on the battlefield- greatly reducing human and political costs - there is a compelling reason to pursue their development as well as to study their capacity to act ethically.

Davenport, David. (2014). Moral Mechanisms. *Philosophy and Technology.* 27(1):47-60. doi: 10.1007/s13347-013-0147-2.
As highly intelligent autonomous robots are gradually introduced into the home and workplace, ensuring public safety becomes extremely important. Given that such machines will learn from interactions with their environment, standard safety engineering methodologies may not be applicable. Instead, we need to ensure that the machines themselves know right from wrong; we need moral mechanisms.

DeBaets, Amy Michelle. (2014). Can a Robot Pursue the Good? Exploring Artificial Moral Agency. *Journal of Evolution and Technology,* 24(3): 76-86.
The author explores an understanding of the potential moral agency of robots, arguing that the key characteristics of physical embodiment, adaptive learning, empathy in action, and a teleology toward the good are the primary necessary components for a machine to become a moral agent.

Hew, Patrick Chisan. (2014). Artificial Moral Agents are Infeasible with Foreseeable Technologies. *Ethics and Information Technology*. 16(3): 197-206.
For an **artificial** agent to be morally praiseworthy, its rules for behaviour and the mechanisms for supplying those rules must not be supplied entirely by external humans. Such systems are a substantial departure from current technologies and theory, and are a low prospect. With foreseeable technologies, an **artificial** agent will carry zero responsibility for its behavior and humans will retain full responsibility.

Howard, Ayanna and Jason Borenstein. 2017. "The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequality. Science and Engineering Ethics. Online first. doi: 10.1007/s11948-017-9975-2.
Recently, there has been an upsurge of attention focused on bias and it’s impact on specialized artificial intelligence (AI) applications. Allegations of racism and sexism have permeated the conversation as stories surface about search engines delivering job postings for well-paying technical jobs to men and not women, or providing arrest mugshots when keywords such as ‘‘black teenagers’’ are entered. Learning algorithms are evolving; they are often created from parsing through large datasets of online information while having truth labels bestowed on them by crowd-sourced masses. These specialized AI algorithms have been liberated from the minds of researchers and startups, and released onto the public. Yet intelligent though they

may be, these algorithms maintain some of the same biases that permeate society. They find patterns within datasets that reflect implicit biases and, in so doing, emphasize and reinforce these biases as global truth. This paper describes specific examples of how bias has infused itself into current AI and robotic systems, and how it may affect the future design of such systems. More specifically, we draw attention to how bias may affect the functioning of (1) a robot peacekeeper, (2) a self-driving car, and (3) a medical robot. We conclude with an overview of measures that could be taken to mitigate or halt bias from permeating robotic technology.

Johnson, Aaron M and Sidney Axinn. (2013). The Morality of Autonomous Robots. *Journal of Military Ethics*. 12(2) 129-141. doi: 10.1080/15027570.2013.818399.

This article focuses on the question, should the decision to take a human life be relinquished to a machine? The authors argue no, and offer several reasons for banning autonomous robots for use in lethal operations.

Johnson, Deborah G. (2017). Reframing AI Discourse." *Minds and Machines.* 27: 575-590. doi: 10.1007/s11023-017-9417-6.

A critically important ethical issue facing the AI research community is how AI research and AI products can be responsibly conceptualised and presented to the public. A good deal of fear and concern about uncontrollable AI is now being displayed in public discourse. Public understanding of AI is being shaped in a way that may ultimately impede AI research. The public discourse as well as discourse among AI researchers leads to at least two problems: a confusion about the notion of ‘autonomy’ that induces people to attribute to machines something comparable to human autonomy, and a ‘sociotechnical blindness’ that hides the essential role played by humans at every stage of the design and deployment of an AI system. Here our purpose is to develop and use a language with the aim to reframe the discourse in AI and shed light on the real issues in the discipline.

Johnson, Deborah G and Keith W. Miller. (2008). Un-making Artificial Moral Agents. *Ethics and Information Technology*. 10(2-3): 123-133. doi: 10.1007/s10676-008-9174-6.

Floridi and Sanders, seminal work, “On the morality of artificial agents” has catalyzed attention around the moral status of computer systems that perform tasks for humans, effectively acting as “artificial agents.” In this paper the authors argue that the move to distinguish levels of abstraction is far from decisive on this issue. They also argue that adopting certain levels of abstraction out of context can be

dangerous when the level of abstraction obscures the humans who constitute computer systems. They frame the debate as a struggle over the meaning and significance of computer systems that behave independently, and not as a debate about the 'true' status of autonomous systems. They argue that while levels of abstraction are useful for particular purposes, when it comes to agency and responsibility, computer systems should be conceptualized and identified in ways that keep them tethered to the humans who create and deploy them.

Nagenborg, Michael, Rafael Capurro, Jutta Weber and Christoph Pingel. (2009). Ethical regulations on robotics in Europe. *AI & Society*. 22(3):349-366. doi: 10.1007/s00146-007-0153-y.

There are only a few ethical regulations that deal explicitly with robots, in contrast to a vast number of regulations, which may be applied. This article focuses on ethical issues with regard to "responsibility and autonomous robots", "machines as a replacement for humans", and "tele-presence". Examining examples from health care, the military, and entertainment, the authors demonstrate that there are legal challenges with regard to these issues.

Pearson, Yvette, and Jason Borenstein. (2012). Creating "Companions" for Children: The Ethics of Designing Esthetic Features for Robots. *AI & Society*. doi: 10.1007/s00146-012-0431-1.

Taking the term ''companion'' in a broad sense to include robot caregivers, playmates, assistive devices, and toys, the authors examine ethical issues that emerge from designing companion robots for children.

Santoro, Matteo, Dante Marino, and Guglielmo Tamburrini. (2008). Learning robots interacting with humans: From epistemic risk to responsibility. *AI & Society*. 22(2): 301-314. doi: 10.1007/s00146-007-0155-9.

Discusses the theoretical and practical limitations in humans' ability to predict and control the behavior of learning robots in their interactions with humans, and the responsibility we have for harm caused by learning robot actions.

Sparrow, Robert. 2009. Building a better warbot: ethical issues in the design of unmanned systems for military applications. *Science and Engineering Ethics.* 15(2): 169-187.

Unmanned systems in military applications will often play a role in determining the success or failure of combat missions and thus in determining who lives and dies in

times of war. Designers of UMS must therefore consider ethical, as well as operational, requirements and limits when developing UMS.

Sparrow, Robert. (2007). Killer robots. *Journal of Applied Philosophy.* 24(1): 62-77. doi: 10.1111/j.1468-5930.2007.00346.x.
Unmanned systems in military applications will often play a role in determining the success or failure of combat missions and thus in determining who lives and dies in times of war. Designers of UMS must therefore consider ethical, as well as operational, requirements and limits when developing UMS. The author groups the ethical issues involved in UMS design under two broad headings, Building Safe Systems and Designing for the Law of Armed Conflict, and identifies and discusses a number of issues under each of these headings.

Tonkins, Ryan. (2009). A challenge for machine ethics. *Minds & Machines*. 19(3): 421-438. doi: 10.1007/s11023-009-9159-1.
This paper articulates a pressing challenge for Machine Ethics: To identify an ethical framework that is both implementable into machines and whose tenets permit the creation of such AMAs in the first place. Without consistency between ethics and engineering, the resulting AMAs would not be genuine ethical robots, and hence the discipline of Machine Ethics would be a failure in this regard. Here this challenge is articulated through a critical analysis of the development of Kantian AMAs, as one of the leading contenders for being the ethic that can be implemented into machines. In the end, however, the development of Kantian artificial moral machines is found to be anti-Kantian. The upshot of all this is that machine ethicists need to look elsewhere for an ethic to implement into their machines.

Tonkins, Ryan. (2012). Out of Character: On the Creation of Virtuous Machines. *Ethics and Information Technology.* 14(2): 137-149. doi: 10.1007/s10676-012-9290-1.
The emerging field of machine ethics is concerned with creating autonomous artificial moral agents that perform ethically significant actions out in the world. Scholars such as Wallach and Allen have argued that a virtue-based moral framework is a promising tool for meeting this end. The author argues that even if we could program autonomous machines to follow a virtue-based moral framework, there are certain pressing ethical issues that need to be taken into account, prior to the implementation and development stages. He discusses whether the creation of virtuous autonomous machines is morally permitted by the central tenets of virtue ethics and finds that creation of such machines violates certain tenets of virtue

ethics, and hence that the creation and use of those machines is impermissible.

Wallach, Wendell. (2008). Implementing moral decision-making facilities in computers and robots. *AI & Society*. 22(4): 463-475. doi: 10.1007/s00146-007-0093-6.
The challenge of designing computer systems and robots with the ability to make moral judgments is stepping out of science fiction and moving into the laboratory. The subject has been designated by several names, including machine ethics, machine morality, artificial morality, or computational morality. Most references to the challenge elucidate one facet or another of what is a very rich topic.

Wendell, Wallach, and Colin Allen. (2013). Framing Robot Arms Control. *Ethics and Information Technology*. 15(2): 125-135. doi: 10.1007/s10676-012-9303-0.
This article draws from previous work done by the authors on autonomy and ethics for robots and applies it to military robots and robot arms control. The authors conclude with a proposal for a first step towards limiting the deployment of autonomous weapons capable of initiating lethal force.

Contributor(s)

Jason Borenstein

Rights

Use of Materials on the OEC

Resource Type

Bibliography

Parent Collection

OEC Bibliographies

Topics

Artificial Intelligence and Robotics
Controversies

Discipline(s)

Computer Sciences

Computer, Math, and Physical Sciences