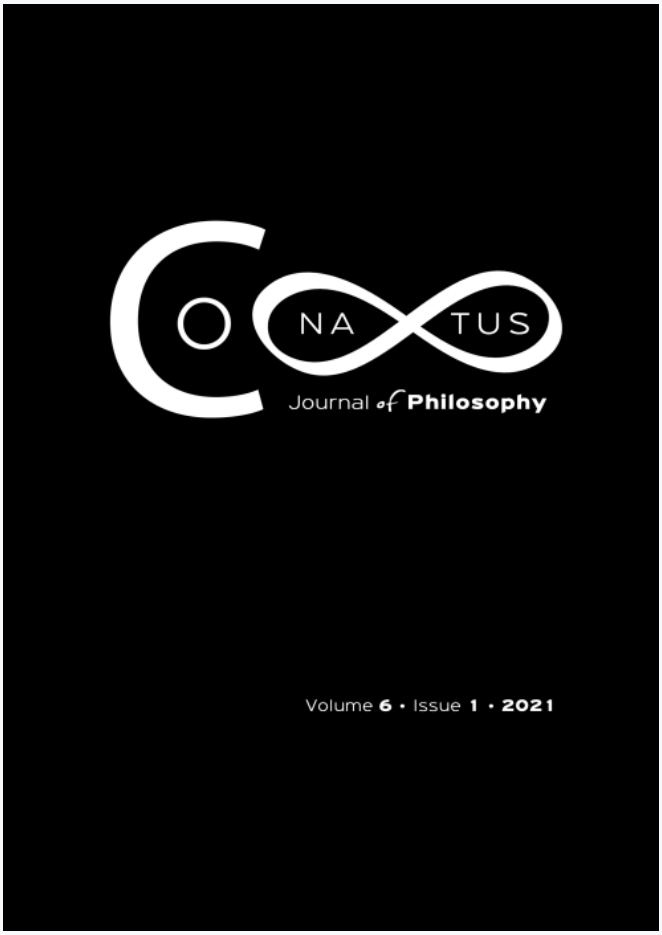
Conatus - Journal of Philosophy

Vol 6, No 1 (2021)

Conatus - Journal of Philosophy



Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson

Michael Anderson, Susan Leigh Anderson, Alkis Gounaris, George Kosteletos

doi: 10.12681/cjp.26832

To cite this article:

Anderson, M., Anderson, S. L., Gounaris, A., & Kosteletos, G. (2021). Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson. *Conatus - Journal of Philosophy*, *6*(1), 177–202. https://doi.org/10.12681/cjp.26832

M. Anderson, S. L. Anderson, A. Gounaris, & G. Kosteletos • *Conatus* 6, no. 1 (2021): 177-202
DOI: http://dx.doi.org/10.12681/cjp.26832

Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson

Michael Anderson,[1] Susan Leigh Anderson,[2] Alkis Gounaris,[3] and George Kosteletos[4]

[1]*University of Hartford, USA*
E-mail address: anderson@hartford.edu
ORCID iD: http://orcid.org/0000-0001-7699-6156

[2]*University of Connecticut, USA*
E-mail address: susan.anderson@uconn.edu

[3]*National and Kapodistrian University of Athens, Greece*
E-mail address: alkisg@philosophy.uoa.gr
ORCID iD: https://orcid.org/0000-0002-0494-6413

[4]*National and Kapodistrian University of Athens, Greece*
E-mail address: gkosteletos@philosophy.uoa.gr
ORCID iD: https://orcid.org/0000-0001-6797-8415

Abstract

At the turn of the 21st century, Susan Leigh Anderson and Michael Anderson conceived and introduced the Machine Ethics research program, that aimed to highlight the requirements under which autonomous artificial intelligence (AI) systems could demonstrate ethical behavior guided by moral values, and at the same time to show that these values, as well as ethics in general, can be representable and computable. Today, the interaction between humans and AI entities is already part of our everyday lives; in the near future it is expected to play a key role in scientific research, medical practice, public administration, education and other fields of civic life. In view of this, the debate over the ethical behavior of machines is more crucial than ever and the search for answers, directions and regulations is imperative at an academic, institutional as well as at a technical level. Our discussion with the two inspirers and originators of Machine Ethics highlights the epistemological, metaphysical and ethical questions arising by this project, as well as the realistic and pragmatic demands that dominate artificial intelligence and robotics research programs. Most of all, however, it sheds light upon the contribution of Susan and Michael Anderson regarding the introduction and undertaking of a main objective related to the creation of ethical autonomous agents, that will not be based on the "imperfect" patterns of human behavior, or on preloaded hierarchical laws and human-centric values.

Key-words: *Machine Ethics; AI Ethics; Philosophy of Artificial Intelligence; Artificial Moral Agents; Ethical Machines; Moral Status of Robots; Computation of Bio-Medical Ethics*

Alkis Gounaris & George Kosteletos: Susan, Michael, thank you very much for the opportunity to discuss such an interesting issue with you. It is our great pleasure and honor to be able to share with our readers and the academic community in Greece and internationally this exceptional conversation. The rapid technological developments of recent years and what the immediate future holds for us bring your work to the forefront of every discussion about AI and Machine Ethics. Building an ethical machine, a possibility that perhaps a few years ago looked like a sci-fi scenario, today seems like an imperative and urgent demand. This seems to be the main objective of your work.

Susan Leigh & Michael Anderson: Thank you for giving us the opportunity to discuss our work in the context of current issues of artificial intelligence!

Alkis Gounaris & George Kosteletos: You introduced the Machine Ethics research program about seventeen years ago.[1] What is the purpose of Machine Ethics and what distinguishes Machine Ethics from the rest of the AI Ethics field? Why is Machine Ethics still important? We are now at the beginning of 2021. Seventeen years later, what is your assessment regarding the evolution of this program?

Susan Leigh Anderson: The main purpose of the Machine Ethics program is to ensure that autonomous AI systems behave in an ethical fashion when interacting with human beings. Secondarily, I believe that it gives us a chance to become clearer about ethics – how to represent its building blocks, resolve contradictions, and come up with principles that should guide the actions of systems functioning in particular domains – that, hopefully, will inspire us to behave better.

Michael Anderson: When we first conceived the idea of Machine Ethics at the turn of the century, the prevailing thinking was that such a notion was still firmly in the realm of science fiction and would remain there for the foreseeable future. This attitude stemmed from a myopic view of the types

[1] Michael Anderson, Suzan Leigh Anderson, and Chris Armen, "Towards Machine Ethics," in *Proceedings of the AAAI-04 Workshop on Agent Organizations: Theory and Practice*, 53-59 (San Jose, CA, 2004); Michael Anderson, Suzan Leigh Anderson, and Chris Armen, "Toward Machine Ethics: Implementing Two Action-based Ethical Theories," in *Machine Ethics, Papers form AAAI Fall Symposium, 2005*, eds. Michael Anderson, Suzan Leigh Anderson, and Chris Armen, Technical Report FS-05-06 (Menlo Park, CA: Association for the Advancement of Artificial Intelligence, 2005), https://www.aaai.org/Library/Symposia/Fall/fs05-06.php; Michael Anderson, Suzan Leigh Anderson, and Chris Armen, "An Approach to Computing Ethics," *IEEE Intelligent Systems* 21, no. 4 (2006): 65-63; Michael Anderson, and Suzan Leigh Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent," *AI Magazine* 28, no. 4 (2007): 15-26; Michael Anderson, and Suzan Leigh Anderson, "The Status of Machine Ethics: A Report from the AAAI Symposium," *Minds & Machines* 17 (2007): 1-10. See also Michael Anderson, and Susan Leigh Anderson, eds., *Machine Ethics* (New York and Cambridge: Cambridge University Press, 2011).

of behavior that would entail ethical concerns and the speed with which autonomous systems capable of such behavior would be upon us.

Given this, the original purpose of the project was to give evidence that

1. autonomous systems need not be fully realized to exhibit behavior of ethical concern
2. ethics is representable and computable
3. the behavior of autonomous systems can be guided by ethical principles

As all AI is machine-based, we see little difference between AI Ethics and Machine Ethics other than its focus on issues raised by the systems recently developed by deep learning. As such systems arise in a black-box fashion from non-vetted data, it is difficult to see how these issues will be resolved and, ultimately, how we will ever be able to guarantee ethical behavior from these systems. Unless such a guarantee can be given, it does not seem likely that such systems will be acceptable. That said, given the surprising proliferation of autonomous systems in general, we believe the tenets of the Machine Ethics project are more relevant than ever.

Alkis Gounaris & George Kosteletos: Having previously argued for the expediency of the Machine Ethics research program, you have pointed out that one of the advantages that machines have over humans in the process of moral judgment is the feature of impartiality (non-bias).[2] Due to their mechanical nature, AI agents are impartial, namely they judge without any bias, unlike humans who tend to be partial, since for example they often decide while being emotionally charged. However, if at some point, in the future, the initial goal of AI is achieved and machines acquire humanlike cognition, do you think they will preserve the advantage of impartiality over humans? Such a question outlines a possible conflict between the basic research objective of AI – specifically the creation of truly intelligent machines – and the goal of Machine Ethics research program regarding the creation of impartial ethical advisors and impartial explicit ethical agents. This possible conflict of the basic research goals of AI and Machine Ethics can also be seen in relation to the vision of creating super-intelligent machines. We say this thinking of Daniel Dennett, who refers to Nietzsche, saying that delusion and deception are characteristics of human nature thus only such a nature can understand

[2] Michael Anderson, Susan Leigh Anderson, and Chris Armen, "An Approach to Computing Ethics," *IEEE Intelligent Systems* 21, no. 4 (2006): 65-63; Michael Anderson, and Suzan Leigh Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent," *AI Magazine* 28, no. 4 (2007): 15-26; Michael Anderson, and Suzan Leigh Anderson, "Robot Be Good," *Scientific American* 303, no. 4 (2010): 72-77.

ethics.[3] Driven by Dennett's position, we think that if the machines reach in the future a kind of super-intelligence that will be impartial at the same time, they may not be interested in ethics at all or will not justify its usefulness.

Susan Leigh Anderson: I have long been concerned with the bias of AI researchers towards trying to reproduce human cognition and human intelligence, and even our ethical values. We are not ideal beings! We can do better than model human behavior as we create autonomous AI entities.

Michael Anderson: Given the initial reticence to see Machine Ethics in any light other than one of science fiction, we purposefully limited the scope of our research to immediate, pragmatic concerns with the hope of convincing some of the scientific fact of its need. It remains to be seen whether "super-intelligence" will make the same leap from fiction to fact. That said, if it does in fact make such a leap, you can be sure if we have given little thought to how we would like such machines to behave towards us, it is likely that we will have little say in how they actually do.

Alkis Gounaris & George Kosteletos: Would you say that the idea for a Machine Ethics, finally the idea that ethics is computable, could be thought of as part of the philosophical tradition supporting that thought equals calculation? Would you consider yourselves as belonging to the same line of thinkers like Hobbes,[4] Leibniz,[5] and more recently Turing,[6] McCulloch and Pitts,[7] or Newell and Simon?[8]

---

[3] Daniel Dennett, "When Hal Kills, Who's to Blame? Computer Ethics," in *Hal's Legacy: 2001's Computer as Dream and Reality*, ed. David G. Stork, 351-365 (Cambridge, MA: MIT Press, 1997).

[4] Thomas Hobbes, *Leviathan, or The Matter, Forme and Power of a Commonwealth Ecclesiastical and Civil*, ed. A. R. Waller (Cambridge: Cambridge University Press, 1904).

[5] Gottfried Wilhelm Leibniz, *Dissertatio de arte combinatoria* (Paris: Hachette Livre-BNF, 2018); Gottfried Wilhelm Leibniz, "Principles of Nature and Grace, Based on Reason," in *Gottfried Wilhelm Leibniz, Philosophical Papers and Letters*, ed. Leroy E. Loemker (Dordrecht: Springer, 1989).

[6] Alan Mathison Turing, "Intelligent Machinery," in *Machine Intelligence 5*, ed. B. Meltzer, and D. M. Michie, 3-23 (Edinburgh: Edinburgh University Press, 1969); Alan Mathison Turing, "Computing, Machinery and Intelligence," *Mind* 59 (1950): 433-460. See also Alan Mathison Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem," in *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life*, ed. Jack B. Copeland, 58-90 (Oxford: Oxford University Press, 2004) – see especially p. 59.

[7] Warren S. McCulloch, and Walter H. Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics* 5 (1943): 115-33.

[8] Allen Newell, and Herbert Alexander Simon, *Current Developments in Complex Information Processing: Technical Report P-850* (Santa Monica, CA: Rand Corporation, 1956); Allen Newell, and John Crosley Shaw, "Programming the Logic Theory Machine," in *IRE-AIEE-ACM '57 (Western): Papers Presented at the February 26-28, 1957, Western Joint Computer Conference: Techniques for Reliability*, 230-240 (New York: Association for Computing Machinery, 1957); Allen Newell, and Herbert Alexander Simon, "The Logic Theory Machine: A Complex Information-Processing System," *IRE Transactions on Information Theory* 2, no. 3

Susan Leigh Anderson: While I believe that ethics is, in principle, computable (and we have been trying to demonstrate this), I'm not sure that I would go so far as to say that all thought is computable. What about artistic ideas?

Michael Anderson: It seems a bit of a stretch from "having machines behave ethically towards us" – the stated goal of our Machine Ethics project – and "all thought is calculation," don't you think?

Alkis Gounaris & George Kosteletos: During the process of ethical decision making one is likely to find oneself facing a condition known in Ethical Philosophy as 'conflict of duties.' Is it possible that in trying to tackle a conflict of moral duties in a computational basis, one might find oneself facing a kind of a 'Halting Problem?'[9] Could it be possible that the explicit ethical agent would be trapped in a never-ending calculation, maybe an infinite loop going back and forth between two opposing duties? In your opinion, are there any major difficulties in the fulfillment of the Machine Ethics endeavor – for instance difficulties related to the ontology, the very nature of calculation or of ethics?

Michael Anderson: Clearly time is of the essence in such decision making and, if competing duties are so closely tied, simply choosing either when time is up would seem a sufficient means to end deliberation. Minsky, in a private conversation, once said to Susan (in his inimitable way) "Ethics is what you do when you run out of time." Just as clearly, hundreds of years of reflection on ethical matters has laid bare a myriad of difficulties that are likely to plague efforts in Machine Ethics as well. That said, perhaps the constrained domain and new perspective of the effort might shed new light on some of these difficulties.

(1956): 61-79; Allen Newell, and Herbert Alexander Simon, "GPS-A Program that Simulates Human Thought," in *Lernende Automaten*, ed. Heinz Billing, 109-124 (Münich: Oldenburg, 1961); Allen Newell, John Crosley Shaw, and Herbert Alexander Simon, "Element of a Theory of Human Problem Solving," *Psychological Review* 65 (1958): 151-166; Allen Newell, and Herbert Alexander Simon, "Computing Science as Empirical Enquiry: Symbols and Search," *Communications of the Association for Computing Machinery* 19 (1976): 113-126; Allen Newell, "Physical Symbol Systems," *Cognitive Science* 4 (1980): 135-183.

[9] Alan Mathison Turing, "On Computable Numbers, with an Application to the Entschiedungsproblem," *Proceedings of the London Mathematical Society Series* 2, no. 42 (1937): 230-265, reprinted in *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life*, ed. Jack Copeland, 58-90 (Oxford: Oxford University Press, 2004); Alan Mathison Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem. A Correction," *Proceedings of the London Mathematical Society* 43 (1938): 544-546; Martin Davis, *Computability and Unsolvability* (New York: McGraw-Hill, 1958), 70. See also Stephen Cole Kleene, *Introduction to Metamathematics* (Amsterdam: North-Holland, 1952), especially Chapter 13: "Computable Functions," and Marvin Minsky, *Computation: Finite and Infinite Machines* (New Jersey: Prentice-Hall, 1967), specifically chapter 8, Section 8.2: "Unsolvability of the Halting Problem."

Alkis Gounaris & George Kosteletos: Persisting a little longer on the issue of 'conflict of duties,' we would like you to comment on a related possibility. We are referring specifically to the case where the machine would have to choose between self-preservation (e.g. the search for vital resources) and continuing to fulfill the principles of a human-centered ethic (e.g. the principles of serving human well-being). Could this conflict of duties be averted by programming rules such as Asimov's *Three Laws of Robotics*?[10] Susan has been critical of them in the past, commenting that they could not be a satisfactory basis for Machine Ethics.[11] Could you tell us a few words about this claim while also suggesting an alternative for facing the above mentioned conflict of duties?

Susan Leigh Anderson: There are a number of problems with Asimov's *Laws* as a basis for Machine Ethics. Roger Clarke[12] has pointed out that there are a number of inconsistencies and ambiguities in the laws. Also, it could allow humans to abuse entities that resemble humans in form, leading to finding it easy to abuse humans as well. Most significantly, from our perspective, a hierarchical ethical duty theory is unsatisfactory because, in agreement with W.D. Ross, we believe that all ethical duties should be viewed as *prima facie*. That is, although all relevant ethical duties should be considered, none should be viewed as being absolute, as the top duty in a hierarchical ordering of duties would be. Each one could be overridden, on occasion, by another duty/duties that would be stronger in a particular situation.

Michael Anderson: Asimov's *Laws* were a landmark in ethical thinking concerning the actions of robots. This is true even when one considers they were devised simply as a device for generating fiction – Asimov seemed to spend more time delineating their weaknesses than championing their strengths. From a real-world perspective, one might question their insufficient specification, incomplete coverage of ethical duties, rigid hierarchal disposition, and required slave-like obedience.

Clearly, the robot has a duty to maintain itself in addition to its other ethical obligations towards its human user. And there is no simple answer as to whether it takes precedence when it conflicts with the other duties as this is a context dependent question. Sometimes it should, say when the robot's other

---

[10] Isaac Asimov, "The Bicentennial Man," in *Philosophy and Science Fiction*, ed. Michael Phillips, 183-216 (Buffalo, New York: Prometheus Books, 1984).

[11] Suzan Leigh Anderson, "Asimov's 'Three Laws of Robotics' and Machine Metaethics," *AI and Society* 22 (2007): 477-493; Suzan Leigh Anderson, "The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics," in *Machine Ethics*, ed. Michael Anderson, and Suzan Leigh Anderson, 285-296 (New York and Cambridge: Cambridge University Press, 2011).

[12] Roger Clarke, "Asimov's Laws of Robotics: Implications for Information Technology. Part I," *Computer* 26, no. 12 (1993): 53-61; Roger Clarke, "Asimov's Laws of Robotics: Implications for Information Technology. Part II," *Computer* 27, no. 1 (1994): 57-66.

duties are not as pressing, and sometimes it shouldn't, say when great harm might befall its human user if the robot tends to its needs rather than hers. Our work in machine ethics has shown how we might tease out the relationships between duties and how to use this information to drive a robot's behavior: abstract principles of conflict resolution from agreed upon cases and use these principles to order actions in terms of their ethical preference.

Alkis Gounaris & George Kosteletos: Bostrom, Yudkowsky and others talk about the so-called Value Loading Problem,[13] namely the problem of how to make machines understand and adopt the values and goals of the humans. However, in our view, even before we address this issue, there may exist another question that we have to answer. Specifically, if one approaches the concept of autonomy in Kantian terms,[14] then arises the question of whether we ought (here, in terms of an ethical "ought") to be concerned with the Value Loading Problem at all. More specifically, dealing with the Value Loading Problem implies the imposition of certain values on the machines (i.e human-centered values, generally values of our own choice etc.). However, this would be against the ethical principle of respecting the autonomy of others. Thus, as human AI developers, we may be faced with the following moral dilemma: Solving the Value Loading Problem to satisfy human goals and ensuring the survival of the human species, or staying consistent with our ethical principle of respect for the autonomy of others?[15] Do you think this dilemma is valid or is it a pseudo-problem? If it is valid, do you see any way out of it?

Susan Leigh Anderson: As I mentioned previously, I don't think we should build all human values into autonomous machines, since humans are prone to unethical behavior. We can, and should, do better than that. Nevertheless, until these entities demonstrate that they have the qualities necessary to

---

[13] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014); Eliezer Yudkowsky, "Complex Value Systems in Friendly AI," in *Artificial General Intelligence*, edited by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 388-393 (Berlin, Heidelberg: Springer, 2011). See also Eliezer Yudkowsky, "The Value Loading Problem," *EDGE*, July 12, 2021, https://www.edge.org/response-detail/26198; Nate Soares, "The Value Learning Problem," in *Artificial Intelligence, Safety and Security*, ed. Roman V. Yampolskiy, 89-97 (Boca Raton, FL: CRC Press, 2019).

[14] Immanuel Kant, *The Groundwork for the Metaphysics of Morals*, trans. Allen W. Wood (New Haven and London: Yale University Press, 2002), for instance see 4: 435-6, 4:440 and 4:447; Immanuel Kant, *Critique of Practical Reason*, trans. Mary Gregor (Cambridge: Cambridge University Press, 2015), see 5:132, also 5:29.

[15] Here, the Value Loading Problem concerns one of the two conflicting duties of the human-developer. It lies at one end of the dilemma, as it has to do with the satisfaction of human goals. The other end is what concerns the respect of the autonomy of others, in this case the AI agents.

be considered to be *full ethical agents* (that we, following James Moor,[16] distinguish from being *explicit ethical agents*, which is what we attempt to create), we don't have to worry about respecting their autonomy. It is perfectly appropriate that, since they are designed to be in the service of human beings (and, perhaps, animals as well), they should be designed to respect their rights.

Alkis Gounaris & George Kosteletos: There are many who argue that creating a literally ethical machine is practically impossible and ultimately unachievable[17] and that we should come to terms with the assumption that at least at an early stage, the basic ethical values will eventually be loaded. Drawing on the theory of W. D. Ross,[18] as well as the *Principles of Biomedical Ethics*[19] by Beauchamp and Childress, you propose that an ethical machine should possess prima facie duties.[20] Do you think that there could be a specific ethical theory that would effectively cover all the possible ethically-laden circumstances (all the cases in need of an ethical analysis) that an AI agent will have to deal with? The danger here is that the agent may operate on the basis of certain principles that will prove to be effective in some cases and ineffective – even dangerous – in others. Furthermore, would a finite set of principles be sufficient for the AI agent to recognize the ethically relevant and

---

[16] James H. Moor, "The Nature, Importance, and Difficulty of Machine Ethics," *IEEE Intelligent Systems* 21, no. 4 (2006): 18-21.

[17] Roman Yampolskiy, "Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach," in *Philosophy and Theory of Artificial Intelligence. Studies in Applied Philosophy, Epistemology and Rational Ethics*, ed. Vincent Müller, 389-396 (Berlin, Heidelberg: Springer, 2013).

[18] W. D. Ross, *The Right and the Good* (Oxford: Clarendon Press, 1930).

[19] T. L. Beauchamp, and J. F. Childress, *Principles of Biomedical Ethics* (Oxford, UK: Oxford University Press, 1979).

[20] For instance see Anderson, Anderson, and Armen, C., "An Approach;" Michael Anderson, and Susan Leigh Anderson, "MedEthEx: A Prototype Medical Ethics Advisor," *Proceedings of the 21st National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, 1759-1765 (Boston, MA: AAAI Press, 2006); Anderson, and Anderson, "Machine Ethics: Creating;" Anderson, "Asimov's Three Laws;" Anderson, and Anderson, "Robot Be Good;" Anderson, "Machine Metaethics;" Michael Anderson, and Suzan Leigh Anderson, "A Prima Facie Duty Approach to Machine Ethics: Machine Learning of Features of Ethical Dilemmas, Prima Facie Duties, and Decision Principles through a Dialogue with Ethicists," in *Machine Ethics*, ed. Michael Anderson, and Suzan Leigh Anderson, 476-492 (New York and Cambridge: Cambridge University Press, 2011); Suzan Leigh Anderson, "Philosophical Concerns with Machine Ethics," in *Machine Ethics*, ed. Michael Anderson, and Suzan Leigh Anderson, 162-167 (New York and Cambridge: Cambridge University Press, 2011); Suzan Leigh Anderson, and Michael Anderson, "Towards a Principle-based Healthcare Agent," in *Machine Medical Ethics*, ed. S. van Rysewyk, and M. Pontier, 67-77 (Cham: Springer, 2015); Michael Anderson, and Suzan Leigh Anderson, "Toward Ensuring Ethical Behavior from Autonomous Systems: A Case-supported Principle-based Paradigm," *Industrial Robot* 42, no. 4 (2015): 324-331.

prominent features of every possible circumstance? In other words, would this *finite* set of ethical principles be sufficient for the AI agent to recognize *every* ethically-laden case as such? There is a risk here that there will be cases that the agent will fail to recognize as ethically-laden (i.e. circumstances asking for an ethical analysis). In addition to the ethical principles themselves, this problem could also arise regarding the criteria for applying these principles. Again, the finite nature of these criteria could make the AI agent fail in the recognition of a situation as ethically-laden (i.e. failure to recognize a situation in which the agent should apply its ethical principles). One might, probably, argue that this is a version of the Frame Problem of AI[21] applied in the case of ethical functioning of the AI agents; or, as we could say, a Moral Frame Problem of AI. With this in mind, the above question can be phrased as such: Is it possible for a specific ethical theory, therefore a *finite* set of ethical principles, to successfully address the Moral Frame Problem of AI?

Susan Leigh Anderson: Two points need to be mentioned here: The first is that, for the foreseeable future, autonomous AI entities are likely to be developed to function in particular domains, with a limited number of ethically relevant features, and corresponding prima facie duties to be considered, leading to a decision principle that can be learned from select ethical dilemmas that are likely to be encountered in those domains. Second, we don't believe that there are situations where *no* ethically relevant features, and corresponding duties, are present when the autonomous AI entity interacts with humans. Those who reject this position tend to think of ethical dilemmas as involving significant harm to a human, but the ethical perspective involves determining the *best* action that could be performed in particular situations. There are always better and worse actions to be considered. So the AI entity, on our view, never has to determine whether a particular situation is an ethically significant one or not. All of its actions should be subsumed under the learned ethical principle, no matter how trivial.

Michael Anderson: It seems that the problem described applies to *all* autonomously-acting agents, including human beings. Until we develop

[21] John McCarthy, and Patrick J. Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence," In *Machine Intelligence, vol. 4*, ed. Bernard Meltzer, and Donald M. Michie, 463-502 (Edinburgh: Edinburgh University Press, 1969). See also Daniel Dennett, *Brainstorms: Philosophical Essays on Mind and Psychology* (Cambridge, MA: MIT Press, 1978), 125; Daniel Dennett, "Cognitive Wheels: The Frame Problem of AI," in *Minds, Machines and Evolution: Philosophical Studies*, ed. C. Hoockway, 129-152 (Cambridge: Cambridge University Press, 1984); Hubert Lederer Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press, 1992), 289; Jerry Alan Fodor, *The Modularity of Mind* (Cambridge, MA: MIT Press, 1983), 114; Zenon W. Pylyshyn, ed., *The Robot's Dilemma: The Frame Problem in Artificial Intelligence* (Norwood, NJ: Ablex, 1987); Michael Wheeler, "Cognition in Context: Phenomenology, Situated Robotics, and the Frame Problem," *International Journal of Philosophical Studies* 16, no. 3 (2008): 323-349; Michael Wheeler, *Reconstructing the Cognitive World: The Next Step* (Cambridge, MA: MIT Press, 2005).

"philosopher robots," and, in the vein of human beings, a race of philosophers, it seems that autonomous agents are doomed by their finite capabilities to make mistakes and, hopefully, learn from them. That said, it seems likely that the set of ethically relevant features, and hence the corresponding duties to minimize or maximize them, is not infinite. In fact, Utilitarians might argue that net good is the *only* ethically relevant feature. While that may or not be the case, we argue that a finer gradation (and hence greater number) of ethically relevant features may be needed to help illuminate the reasoning behind ethical decision making.

Alkis Gounaris & George Kosteletos: If at least for the time being we cannot avoid the (even partial) 'loading' of some basic or initial moral values to the AI agents, then shouldn't this process of regulating 'value loading' involve the end-users and not only the AI developers? In other words, shouldn't the ordinary citizens have a say in the choice of those principles? Additionally, shouldn't each cultural background regarding morality be taken into account? We saw in a very interesting MIT experiment the different ways in which different cultures react to the 'trolley problem' that came to the fore with the evolution of smart cars.[22] The question is whether the design of an ethical machine should follow the demand for the democratization of technology and technical design[23] [24] – or even a culture based technical design.[25] Recently, you have also proposed a framework promoting public participation as part

---

[22] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan, "The Moral Machine Experiment," *Nature* 563, no. 7729 (2018): 59-64; Jean-Francois Bonnefon, Azim Shariff, and Iyad Rahwan, "The Social Dilemma of Autonomous Vehicles," *Science* 352, no. 6293 (2016): 1573-1576; Edmond Awad, Sohan Dsouza, Azim Shariff, Iyad Rahwan, and Jean-Francois Bonnefon, "Universals and Variations in Moral Decisions Made in 42 Countries by 70,000 Participants," *Proceedings of the National Academy of Sciences* 117, no. 5 (2020): 2332-2337.

[23] Andrew Feenberg, "Subversive Rationalization: Technology, Power, and Democracy," in *Technology and the Politics of Knowledge*, ed. Andrew Feenberg, and Alastair Hannay, 3-11 (Bloomington and Indianapolis: Indiana University Press, 1995); Andrew Feenberg, *Questioning Technology* (London, New York: Routledge, 1999); Carl Mitcham, *Thinking through Technology: The Path between Engineering and Philosophy* (Chicago: The University of Chicago Press, 1994); Langdon Winner, "Technè and Politeia: The Technical Constitution of Society," in *Philosophy of Technology*, ed. Paul T. Dubrin, and Friedrich Rapp, 97-111 (Dordrecht, Boston, Lancaster: D. Reidel, 1983); Langdon Winner, "Citizen Virtues in a Technological Order," *Inquiry* 35, nos. 3-4 (1992): 341-361.

[24] The question regarding the democratization of Technology is closely related to the notions of *inclusion*, *fairness* and *transparency*, which seem to have become popular topics in the AI research literature. See *The 2019 AI Index Annual Report, Stanford University Human Centered AI*, Chapter 8: "Societal Considerations," especially pages 149-151.

[25] Karen Hao, "Should a Self-driving Car Kill the Baby or the Grandma? Depends on where You're from," *MIT Technology Review*, October 14, 2018, https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/.

of a process – or as you call it, a "tool" – for the formulation of principles to be loaded to the machines.[26] Do you generally agree with an inclusive approach with regards to the Machine Ethics research program?

Susan Leigh Anderson: I have argued that, in general, applied ethicists (with knowledge of the domains in question) should be involved in learning the ethical principles, from the ethically relevant features and correlative prima facie duties that should govern the behavior of autonomous AI entities in specific domains. They have an expertise that others lack. But I have also accepted (after discussions with Edmond Awad) that there is an ethically justifiable place for the opinions of the general public concerning emerging technologies, for instance, driverless cars: Since there has been push-back from the public about allowing driverless cars in large part because of a death in Arizona by a driverless car and concern that there are bound to be situations, even with improved sensors, where the behavior of driverless cars could result in deaths, there needs to be a way for the public to weigh in on this possibility to allow for the acceptance of driverless cars, which would certainly result in fewer deaths than with human drivers who are often distracted, tired or impaired.

Until recently Michael and I have maintained that we didn't think that machines should be permitted to function autonomously in domains where life-and-death decisions need to be made, because they are controversial. Such decisions are controversial because they are often emotionally driven for ordinary people and even ethicists disagree about how to weigh the various ethically relevant factors involved. The case of driverless cars is very different, I now see. A central ethical concern for any action or policy must be causing the least harm. This is universally agreed upon. It seems clear that having only driverless cars would result in less harm than having only human drivers. If there were some way to placate the public's concerns about when driverless cars behavior might lead to human deaths, leading to allowing them, it should be taken seriously. Encouraging the public to have a say in what driverless cars should do in various possible scenarios where death might result, making the results known and adopting the majority's view (probably for a particular society), might just be enough for the public to accept driverless cars, which is likely to lead to fewer deaths overall.

And, actually, it is consistent with our long held position that only humans should make life-and-death decisions since, although the cars function autonomously, the decisions they make were determined by humans who

[26] Edmond Awad, Michael Anderson, Suzan Leigh Anderson, and Beishui Liao, "An Approach for Combining Ethical Principles with Public Opinion to Guide Public Policy," *Artificial Intelligence* 287 (2020): article 103349.

gave them rules to follow. Humans will be held accountable if the results are questioned. I foresee challenges to the majority's recommended policies as time goes by, leading perhaps to new policies approved by the majority, just as laws in this country are changed over time, hopefully leading to more ethically acceptable ones as ethicists and others weigh in.

Alkis Gounaris & George Kosteletos: In several publications you refer to the creation of ethical advisors such as the bio-medical advisor MedEthEx.[27] As mentioned before, you suggest that the most appropriate way for these advisors to operate is based on the *Principles of Biomedical Ethics*[28] by Tom Beauchamp and James Childress. Can you tell us a bit more about your proposal?

Susan Leigh Anderson: We began testing our approach to representing ethics in a machine, and generating ethical decision principles from considering specific cases of ethical dilemmas, by using a general type of ethical dilemma often faced by medical practitioners, where the ethics is clear. Medical Ethics is quite well established and there is agreement on using Beauchamp and Childress's principles (*prima facie* duties, in our view, since there is no decision principle to resolve cases where they give conflicting advice) to frame discussions.

Here is the common type of ethical dilemma we considered: A health care worker has recommended a particular treatment for her competent adult patient and the patient has rejected that treatment option. Should the health care worker try again to change the patient's mind or accept the patient's decision as final? The dilemma arises because, on the one hand, the health care worker may not want to risk upsetting the patient by challenging his decision; on the other hand, the health care worker may have concerns about why the patient is refusing the treatment. Three of the four principles/duties of Biomedical Ethics are likely to be satisfied or violated in dilemmas of this type: the duty of respect for autonomy, the duty of nonmaleficence and the duty of beneficence. The system accepts a range of values for each of the duties from –2 to +2, where -2 represents a serious violation of the duty, -1 a less serious violation, 0 indicates that the duty is neither satisfied nor violated, +1 indicates a minimal satisfaction of the duty and +2 a maximal satisfaction of the duty.

Through inductive logic, after considering several cases giving reasons why the patient was rejecting the recommended treatment where the answer is clear as to whether the patient's decision should be accepted or challenged,

---

[27] Anderson, and Anderson, "MedEthEx."

[28] Tom Lamar Beauchamp, and James Franklin Childress, *Principles of Biomedical Ethics* (Oxford: Oxford University Press, 1979).

the system learned this principle: A health care worker should challenge a patient's decision if it is not fully autonomous and *either* there is any violation of the duty of nonmaleficence *or* there is a severe violation of the duty of beneficence. This philosophically interesting result gives credence to Rawls' Method of Reflective Equilibrium.[29] We have, through abstracting a principle from intuitions about particular cases and then testing that principle on further cases, come up with a plausible principle that tells us which action is correct when specific duties pull in different directions in a particular ethical dilemma. Furthermore, the principle that has been abstracted supports an insight of Ross's that violations of the duty of nonmaleficence should carry more weight than violations of the duty of beneficence.

Alkis Gounaris & George Kosteletos: However, in addition to the question "What ethical principles should the AI ethical advisor, or the explicit agent, apply?" the question "To which entities should the AI agent apply these criteria?" arises as well. Some would say that this question seems to be gaining in importance considering the possibility of developing in the future machines with a significant degree of autonomy that will be able to interact with their environment in a more 'holistic' way. In such a case, we also need to face the question of "How will the AI agent decide a) which of its surrounding entities have moral standing and therefore need a moral treatment from the AI agent?, and b) what exactly this moral standing would involve?" Is the issue of defining criteria for the attribution of moral status to others crucial for the Machine Ethics research program? If so, are there any satisfactory criteria that an AI agent could effectively apply for the attribution of moral status to its surrounding entities?

Susan Leigh Anderson: In my view *sentience* is the quality an entity should possess to have moral standing, because only an entity possessing this quality would care what happens to it. But it is difficult to detect whether this quality is present in an entity other than oneself. And it isn't necessary to possess this quality for it to be important that we treat an entity as if it has moral standing. I have argued – using Kant's argument for why we should treat animals well, where he maintained that even though they don't have rights themselves (now debatable), because they resemble us we should treat them as if they have rights lest it lead to a slippery slope where it becomes easier to

[29] John Rawls, *A Theory of Justice* (Cambridge, MA: The Belknap Press of Harvard University Press, 1971). For subsequent refinements and reappraisals of the theoretical construct of the Reflective Equilibrium see John Rawls, *A Theory of Justice*, 2nd edition (Cambridge, MA: The Belknap Press of Harvard University Press, 1999), and John Rawls, "Justice as Fairness: Political not Metaphysical," *Philosophy and Public Affairs* 14 (1985): 223-251. For the distinction between narrow and wide Reflective Equilibrium see John Rawls, *Justice as Fairness: A Restatement* (Cambridge, MA: Harvard University Press, 2001), 31.

mistreat other humans – that any entity that resembles us in form or function should be treated as if it has moral standing.

Alkis Gounaris & George Kosteletos: Let us insist for a moment on the issue of the criteria for the attribution of moral status. From time to time in some of your papers you have considered the question of whether an explicit ethical agent should follow a set of ethical principles that will involve the fact that the agent itself has a moral standing. Namely, whether the agent should 'consider' (or consider) itself as an entity with moral standing.[30] Could you please tell us more about the significance and the importance of this question?

Susan Leigh Anderson: I don't think it's important to determine its own status in order to decide how *it* should treat others.

Michael Anderson: I can imagine that one might draw the wrong conclusion about our stance towards this question when one considers that we advocate that such an agent has a duty to maintain itself. I would argue that this does not in fact pertain to an attribution of moral status to the agent but instead is concerned with making sure that the agent maintains its capacity to fulfill its other duties towards its user.

Alkis Gounaris & George Kosteletos: In addition to the above question as to whether the explicit ethical agent 'considers' (or considers) itself as an entity with moral standing, many AI researchers also reasonably pose the question of whether *we* should consider the explicit ethical agent (and generally any AI agent) as an entity with moral standing.[31] Should we bother with the attribution of moral status to AI entities? If so, what do you think the criteria are that an explicit ethical agent (and, more generally, an AI agent) should meet in order for moral status to be attributed to it? For example, some people think that an ethical Turing Test will be enough to attribute moral status to machines.[32] Do you think accepting this view is the only way

---

[30] Anderson, "Asimov's Three Laws."

[31] For instance, Luciano Floridi, and J. W. Sanders, "On the Morality of Artificial Agents," *Minds and Machines* 14 (2004): 349-379; Christian Hugo Hoffmann, and Benjamin Hahn, "Decentered Ethics in the Machine Era and Guidance for AI Regulation," *AI & Society* 35, no. 3 (2009): 635-644; David Levy, "The Ethical Treatment of Artificially Conscious Robots," *International Journal of Social Robotics* 1, no. 3 (2009): 209-216; Bertram F. Malle, Thapa Stuti Magar, and Matthias Scheutz, "AI in the Sky: How People Morally Evaluate Human and Machine Decisions in a Lethal Strike Dilemma," in *Robotics and Well-Being*, ed. Maria Aldinhas Ferreira, João Silva Sequeira, Gurvinder Singh Virk, Mohammad Tokhi Osman, and Ender E. Kadar, 111-133 (Cham: Springer, 2019); Robert Sparrow, "Killer Robots," *Journal of Applied Philosophy* 24, no. 1 (2007): 62-77. See also Jonathan Owen, and Richard Osley, "Bill of Rights for Abused Robots: Experts Draw up an Ethical Charter to Prevent Humans Exploiting Machines," *The Independent*, September 17, 2011, https://www.independent.co.uk/news/science/bill-of-rights-for-abused-robots-5332596.html.

[32] Colin Allen, Varner Gary, and Zinser Jason, "Prolegomena to Any Future Artificial Moral

to go? You have also commented[33] on criteria like Jeremy Bentham's and Peter Singers' criterion of sentience[34] – which you have also mentioned earlier in our discussion – Immanuel Kant's criterion of self-consciousness,[35] Michael Tooley's criterion of desire (for a moral right),[36] and Mary Anne Warren's criterion of emotionality.[37] Do you find any flaws in these criteria?[38] For example, does the Other Minds Problem pose a threat to the feasibility of applying such criteria, namely criteria of an internalist kind?[39] Furthermore, what do you think the moral status of AI agents could finally be?

Susan Leigh Anderson: As I mentioned earlier, answering your question regarding the criteria that an AI agent could effectively apply for the attribution of moral status to its surrounding entities, given the Problem of Other Minds, we may never know whether an autonomous AI entity possesses the quality essential to having moral standing, but I have argued that we should treat it (if it resembles us, or an animal, in form or function) as if it does.

Alkis Gounaris & George Kosteletos: With your work you have opened a new path for the treatment and resolution of the ethically-laden biomedical

---

Agent," *Journal of Experimental and Theoretical Artificial Intelligence* 12 (2000): 151-261; Robert Sparrow, "The Turing Triage Test," *Ethics and Information Technology* 6 (2004): 201-213, especially 204.

[33] Anderson, "Asimov's Three Laws."

[34] Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation*, ed. J. Burns, and H. Hart (Oxford: Clarendon Press, 1789), especially Chapter 17: "Boundary around Penal Jurisprudence." Also Peter Singer, "All Animals Are Equal," in *Animal Ethics: Past and Present Perspectives*, ed. Evangelos D. Protopapadakis, 163-178 (Berlin: Logos Verlag, 2012); Peter Singer, *Animal Liberation: A New Ethics for our Treatment of Animals* (New York: New York Review of Books, 1975); Peter Singer, *Practical Ethics*, 2nd edition (Cambridge: Cambridge University Press, 1993).

[35] Immanuel Kant, "Our Duties to Animals," in his *Lectures on Ethics*, trans. L. Infield, 239-241 (New York: Harper & Row, 1963).

[36] Michael Tooley, "In Defense of Abortion and Infanticide," In *The Abortion Controversy: A Reader*, ed. Luis P. Pojman, and Francis J. Beckwith, 186-213 (Boston, MA: Jones & Bartlett, 1994).

[37] Mary Anne Warren, "On the Moral and Legal Status of Abortion," in *Contemporary Moral Problems*, ed. J. White, 144-155 (Belmont, CA: Wadsworth/Thompson Learning, 2003).

[38] For an analysis of the flaws in the proposed criteria regarding the attribution of moral status to AI entities, see Alkis Gounaris, and George Kosteletos, "Licensed to Kill: Autonomous Weapons as Persons and Moral Agents," in *Personhood*, ed. Dragan Prole, and Goran Rujević, 137-189 (Novi Sad: The NKUA Applied Philosophy Research Lab Press, 2020).

[39] For the way in which the Other Minds Problem could enter the discussion of AI Ethics and the application of moral status to the AI agents see D. Gunkel, *The Machine Question: Critical Perspectives on AI, Robots and Ethics* (Cambridge, MA: MIT Press, 2012); C. Hoffmann, and B. Hahn, "Decentered Ethics in the Machine Era and Guidance for AI Regulation," *AI & Society* 35, no. 3 (2009): 635-644; D. Levy, "The Ethical Treatment of Artificially Conscious Robots," *International Journal of Social Robotics* 1, no. 3 (2009): 209-216.

problems.[40] Do you think bioethicists or philosophers in general should be concerned about their work in the future? Will the machines be able to replace them, at some point, completely? Could machines become the 'philosophers' of a new Plato's *Republic*?

Susan Leigh Anderson: I do think that there is a possibility of there being more objectivity in machine decision-making, if properly designed; but new issues are bound to arise (conditions change) that would require up-dates. And an important philosophical question will never disappear: What gives our lives meaning?

Alkis Gounaris & George Kosteletos: How much do you really think we are in danger from AI? Some argue that the cooperation of the AI and the Biotechnology fields will lead to new forms of intelligence in the near future.[41] What should we hope for and what should we fear about that? We see the media, pop writers like Harari,[42] businessmen like Musk,[43] and the academic community as well (e.g. Bostrom[44] or Tegmark[45] Institutes) holding a cautious

---

[40] Michael Anderson, and Susan Leigh Anderson, "MedEthEx: A Prototype Medical Ethics Advisor," *Proceedings of the 21st National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, 1759-1765 (Boston, MA: AAAI Press, 2006), http://dblp.uni-trier.de/db/conf/aaai/aaai2006.html#AndersonAA06; Michael Anderson, and Susan Leigh Anderson, "ETHEL: Toward a Principled Ethical Eldercare System," *Proceedings of the AAAI Fall Symposium: New Solutions to Old Problems. Technical Report FS-08-02* (Arlington, VA, 2008); Michael Anderson, and Susan Leigh Anderson, "Robot Be Good," *Scientific American* 303, no. 4 (2010): 72-77. Also, Michael Anderson, and Susan Leigh Anderson, "A Prima Facie Duty Approach to Machine Ethics: Machine Learning of Features of Ethical Dilemmas, Prima Facie Duties, and Decision Principles through a Dialogue with Ethicists," in *Machine Ethics*, ed. Michael Anderson, and Suzan Leigh Anderson, 476-492 (New York and Cambridge: Cambridge University Press, 2011); Susan Leigh Anderson, and Michael Anderson, "Towards a Principle-Based Healthcare Agent," in *Machine Medical Ethics*, ed. S. van Rysewyk, and M. Pontier, 67-77 (Cham: Springer, 2015).

[41] Yuval Noah Harari, *21 Lessons for the 21st Century* (New York: Spiegel & Grau, 2018).

[42] Nicholas Thompson, "Will Artificial Intelligence Enhance or Hack Humanity?" *Wired*, April 20, 2019, https://www.wired.com/story/will-artificial-intelligence-enhance-hack-humanity/.

[43] Catherine Clifford, and Elon Musk: "Mark my Words – A.I. is far more Dangerous than Nukes." *CNBC*, March 13, 2018, https://www.cnbc.com/2018/03/13/elon-musk-at-sxsw-a-i-is-more-dangerous-than-nuclear-weapons.html; Gregory Wallace, "Elon Musk Warns against Unleashing Artificial Intelligence 'Demon,'" *CNN Business*, October 26, 2014, https://money.cnn.com/2014/10/26/technology/elon-musk-artificial-intelligence-demon/; Ricki Harris, "Elon Musk: Humanity Is a Kind of 'Biological Boot Loader' for AI," *Wired*, January 9, 2019, https://www.wired.com/story/elon-musk-humanity-biological-boot-loader-ai/.

[44] For more, visit *The Future of Humanity Institute*, https://www.fhi.ox.ac.uk/; Nick Bostrom, "Existential Risk Prevention as Global Priority," *Global Policy* 4, no. 1 (2013): 15-31.

[45] For more, visit https://futureoflife.org/; Max Tegmark, "Benefit and Risks of Artificial Intelligence," *Future of Life Institute*, https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/; Stuart Russell, et al., "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," *Future of Life Institute*, https://futureoflife.org/open-letter-

and in some cases technophobic attitude in the face of developments. How could we distinguish the real risks from the pseudo-problems?

Susan Leigh Anderson: Whether we have to fear AI technology depends on how it is developed and by whom it is used. In itself it is neutral. If we develop AI entities on a human model, embodying negative human qualities (like self-centeredness, favoring one's own group) and allow anyone to use them, they could become super weapons. This is why the field of Machine Ethics is so important. We have the opportunity to create *ethical machines*, non-threatening machines that not only aid us in many ways, but can also show us how we need to behave if we are to survive as a species.

Michael Anderson: For all the perils of doing so, I see no other option than trusting science. Yes, it has led us into dangers that we might not have faced if we had kept our blinders on but it has also been the shining light that has taken humanity out of the darkness, illuminating many mysteries of the universe. Given the risks humanity lives under, my hope for AI is that it might serve as a means for preserving intelligence. As it stands, this is currently only housed in human bodies – a vessel so fragile that it might be prudent to develop backup for it. Wouldn't it be the ultimate tragedy if we were the only intelligent creatures in the universe and, through inaction, let our unique spark die out?

Alkis Gounaris & George Kosteletos: Let us now come to something even more current. In your opinion, which are the most prominent ethical challenges raised by the COVID-19 pandemic? Could Machine Ethics contribute in facing them? Could these ethical challenges be faced more successfully by an AI agent equipped with moral principles, than by human committees of doctors, epidemiologists, politicians and bioethicists? Finally, does this pandemic crisis provide the Machine Ethics research program with any lessons to be learned and used in similar crises in the future? What do you suggest so that the public would be prepared for such contributions by the AI agents?

Susan Leigh Anderson: What machines are good at (better than humans) is digesting a lot of data quickly: discovering connections, etc. Humans are still needed to input the data and ethicists are more likely to insist that the data is not skewed to gloss over ethical issues. For instance, one could just keep track of whether people are offered vaccines, just noting that fewer members of minority communities seem to be taking them, ignoring past legitimate concerns in these communities about taking vaccines and whether attempts have been made to educate them, or whether the means for notifying them

autonomous-weapons/; Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Knopf, 2017).

of their chance to take the vaccine is likely to reach them (if through the internet: whether they have access to the internet and the skill at navigating it).

Michael Anderson: What machine ethics has to offer is consistent, impartial treatment of like cases. In the face of seemingly novel ethical challenges, it is hoped that this might prove useful in illuminating similarities to previous challenges thereby contributing to current ones.

Alkis Gounaris & George Kosteletos: In a recent publication,[46] Michael among others, endorses the position that inclusive, interdisciplinary teams are needed to develop AI. What do you think that the role of philosophers is in such an endeavor?

Michael Anderson: What seems to elude many is that there is expertise in ethics as there is in any academic discipline. This misapprehension seems to stem from the fact that people make "ethical" decisions daily and therefore have difficulty understanding why such expertise is needed. That said, doesn't it seem obvious that those who have spent their research careers in a field might have greater insight into it? Clearly, the intuitive approach most bring to such decisions is riddled with partiality and inconsistency, not to mention a circumscribed understanding of the plethora of factors involved. The expertise ethicists bring to the table is necessary to help alleviate these shortcomings.

Alkis Gounaris & George Kosteletos: What remains to be achieved? Which would be the key concerns and the basic challenges of Machine Ethics in the future? Should we expect in the near future a safe, ethical and/or responsible AI?

Michael Anderson: Not a soothsayer but it's pretty clear to me that autonomous systems are here to stay and it would be unwise to ignore their ethical tuition. Unfortunately, given its need for copious data and the dearth of such data in the domain of ethics, the silver bullet of deep learning does not seem to have much to offer to this issue. Where value judgements are involved, it seems that we are going to have to bite the bullet and do the hard work of determining just how we want such systems to behave towards us.

Alkis Gounaris & George Kosteletos: Susan, Michael, thank you for the extremely interesting discussion and we look forward to having you with us at our upcoming *Me and AI: Human Concerns Artificial Minds* Conference.

Susan Leigh & Michael Anderson: Thank you for your thought-provoking questions! Your conference could not come at a more opportune time!

---

[46] Steve Taylor, et al., "Responsible AI – Key Themes, Concerns & Recommendations for European Research and Innovation," *Zenodo*, July 2, 2018.

Author Contribution Statement

George Kosteletos and Alkis Gounaris conceived and designed the paper. All authors contributed equally to the writing and critical revision of the manuscript. All authors approved the final version for submission.

References

Aldinhas, Ferreira Maria, João Silva Sequeira, Gurvinder Singh Virk, Mohammad Tokhi Osman, and Ender E. Kadar, eds. *Robotics and Well-being*. Cham: Springer, 2019.

Allen, Colin, Varner Gary, and Zinser Jason. "Prolegomena to Any Future Artificial Moral Agent." *Journal of Experimental and Theoretical Artificial Intelligence* 12 (2000): 151-261.

Anderson, Michael, and Suzan Leigh Anderson, eds. *Machine Ethics*. New York and Cambridge: Cambridge University Press, 2011.

Anderson, Michael, and Suzan Leigh Anderson. "A Prima Facie Duty Approach to Machine Ethics: Machine Learning of Features of Ethical Dilemmas, Prima Facie Duties, and Decision Principles through a Dialogue with Ethicists." In *Machine Ethics*, edited by Michael Anderson, and Suzan Leigh Anderson, 476-492. New York and Cambridge: Cambridge University Press, 2011.

Anderson, Michael, and Suzan Leigh Anderson. "ETHEL: Toward a Principled Ethical Eldercare System." *Proceedings of the AAAI Fall Symposium: New Solutions to Old Problems*. Technical Report FS-08-02. Arlington, VA, 2008.

Anderson, Michael, and Suzan Leigh Anderson. "Guest Editors' Introduction: Machine Ethics." *IEEE Intelligent Systems* 21, no. 4 (2006): 10-11.

Anderson, Michael, and Suzan Leigh Anderson. "Machine Ethics: Creating an Ethical Intelligent Agent." *AI Magazine* 28, no. 4 (2007): 15-26.

Anderson, Michael, and Suzan Leigh Anderson. "MedEthEx: A Prototype Medical Ethics Advisor." *Proceedings of the 21st National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, 1759-1765. Boston, MA: AAAI Press, 2006.

Anderson, Michael, and Suzan Leigh Anderson. "Robot Be Good." *Scientific American* 303, no. 4 (2010): 72-77.

Anderson, Michael, and Suzan Leigh Anderson. "The Status of Machine Ethics: A Report from the AAAI Symposium." *Minds & Machines* 17 (2007): 1-10.

Anderson, Michael, and Suzan Leigh Anderson. "Toward Ensuring Ethical Behavior from Autonomous Systems: A Case-supported Principle-based Paradigm." *Industrial Robot* 42, no. 4 (2015): 324-331.

Anderson, Michael, Suzan Leigh Anderson, and Chris Armen, eds. *Machine Ethics: Papers form AAAI Fall Symposium, 2005*. Technical Report FS-05-06. Menlo Park, CA: Association for the Advancement of Artificial Intelligence, 2005. https://www.aaai.org/Library/Symposia/Fall/fs05-06.php.

Anderson, Michael, Suzan Leigh Anderson, and Chris Armen. "An Approach to Computing Ethics." *IEEE Intelligent Systems* 21, no. 4 (2006): 65-63.

Anderson, Michael, Suzan Leigh Anderson, and Chris Armen. "Toward Machine Ethics: Implementing Two Action-Based Ethical Theories." In *Machine Ethics, Papers form AAAI Fall Symposium, 2005*, edited by Michael Anderson, Suzan Leigh Anderson, and Chris Armen, Technical Report FS-05-06. Menlo Park, CA: Association for the Advancement of Artificial Intelligence, 2005.

Anderson, Michael, Suzan Leigh Anderson, and Chris Armen. "Towards Machine Ethics." In *Proceedings of the AAAI-04 Workshop on Agent Organizations: Theory and Practice*, 53-59. San Jose, CA, 2004.

Anderson, Suzan Leigh, and Michael Anderson. "Towards a Principle-Based Healthcare Agent." In *Machine Medical Ethics*, edited by S. van Rysewyk, and M. Pontier, 67-77. Cham: Springer, 2015.

Anderson, Suzan Leigh. "Asimov's 'Three Laws of Robotics' and Machine Metaethics." *AI and Society* 22 (2007): 477-493.

Anderson, Suzan Leigh. "Machine Metaethics." In *Machine Ethics*, edited by Michael Anderson, and Suzan Leigh Anderson, 21-27. New York and Cambridge: Cambridge University Press, 2011.

Anderson, Suzan Leigh. "Philosophical Concerns with Machine Ethics." In *Machine Ethics*, edited by Michael Anderson, and Suzan Leigh Anderson, 162-167. New York and Cambridge: Cambridge University Press, 2011.

Anderson, Suzan Leigh. "The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics." In *Machine Ethics*, edited by Michael Anderson, and Suzan Leigh Anderson, 285-296. New York and Cambridge: Cambridge University Press, 2011.

Asimov, Isaac. "The Bicentennial Man." In *Philosophy and Science Fiction*, edited by Michael Phillips, 183-216. Buffalo, NY: Prometheus Books, 1984.

Awad, Edmond, Michael Anderson, Suzan Leigh Anderson, and Beishui Liao. "An Approach for Combining Ethical Principles with Public Opinion to Guide Public Policy." *Artificial Intelligence* 287 (2020): article 103349.

Awad, Edmond, Sohan Dsouza, Azim Shariff, Iyad Rahwan, and Jean-Francois Bonnefon. "Universals and Variations in Moral Decisions Made in 42 Countries by 70,000 Participants." *Proceedings of the National Academy of Sciences* 117, no. 5 (2020): 2332-2337.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. "The Moral Machine Experiment." *Nature* 563, no. 7729 (2018): 59-64.

Beauchamp, Tom Lamar, and James Franklin Childress. *Principles of Biomedical Ethics*. Oxford, UK: Oxford University Press, 1979.

Bentham, Jeremy. *An Introduction to the Principles of Morals and Legislation*. Edited by J. Burns, and H. Hart. Oxford: Clarendon Press, 1789.

Bonnefon, Jean-Francois, Azim Shariff, and Iyad Rahwan. "The Social Dilemma of Autonomous Vehicles." *Science* 352, no. 6293 (2016): 1573-1576.

Bostrom, Nick. "Existential Risk Prevention as Global Priority." *Global Policy* 4, no. 1 (2013): 15-31.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Clarke, Roger. "Asimov's Laws of Robotics: Implications for Information Technology. Part I." *Computer* 26, no. 12 (1993): 53-61.

Clarke, Roger. "Asimov's Laws of Robotics: Implications for Information Technology. Part II." *Computer* 27, no. 1 (1994): 57-66.

Clifford, Catherine, and Elon Musk. "Mark my Words – A.I. is far more Dangerous than Nukes." *CNBC*, March 13, 2018. https://www.cnbc.com/2018/03/13/elon-musk-at-sxsw-a-i-is-more-dangerous-than-nuclear-weapons.html.

Davis, Martin. *Computability and Unsolvability*. New York: McGraw-Hill, 1958.

Dennett, Daniel. "Cognitive Wheels: The Frame Problem of AI." In *Minds, Machines and Evolution: Philosophical Studies*, edited by C. Hoockway, 129-152. Cambridge: Cambridge University Press, 1984.

Dennett, Daniel. "When Hal Kills, Who's to Blame? Computer Ethics." In *Hal's Legacy: 2001's Computer as Dream and Reality*, edited by David G. Stork, 351-365. Cambridge, MA: MIT Press, 1997.

Dennett, Daniel. *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press, 1978.

Dreyfus, Hubert Lederer. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press, 1992.

Feenberg, Andrew. "Subversive Rationalization: Technology, Power, and Democracy." In *Technology and the Politics of Knowledge*, edited by Andrew Feenberg, and Alastair Hannay, 3-11. Bloomington and Indianapolis: Indiana University Press, 1995.

Feenberg, Andrew. *Questioning Technology*. London, New York: Routledge, 1999.

Floridi, Luciano, and J. W. Sanders. "On the Morality of Artificial Agents." *Minds and Machines* 14 (2004): 349-379.

Fodor, Jerry Alan. *The Modularity of Mind*. Cambridge, MA: MIT Press, 1983.

Gounaris, Alkis, and George Kosteletos. "Licensed to Kill: Autonomous Weapons as Persons and Moral Agents." In *Personhood*, edited by Dragan Prole, and Goran Rujievi ć, 137-189. Hellenic-Serbian Philosophical Dialogue Series, vol. 2. Novi Sad: The NKUA Applied Philosophy Research Lab Press, 2020.

Gunkel, David. *The Machine Question: Critical Perspectives on AI, Robots and Ethics*. Cambridge, MA: MIT Press, 2012.

Hao, Karen. "Should a Self-driving Car Kill the Baby or the Grandma? Depends on where You're from." *MIT Technology Review*, October 14, 2018, https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/.

Harari, Yuval Noah. *21 Lessons for the 21st Century*. New York: Spiegel & Grau, 2018.

Harris, Ricki. "Elon Musk: Humanity Is a Kind of 'Biological Boot Loader' for AI." *Wired*, January 9, 2019, https://www.wired.com/story/elon-musk-humanity-biological-boot-loader-ai/.

Hobbes, Thomas. *Leviathan, or The Matter, Forme and Power of a Commonwealth Ecclesiastical and Civil*. Edited by A. R. Waller. Cambridge: Cambridge University Press, 1904.

Hoffmann, Christian Hugo, and Benjamin Hahn. "Decentered Ethics in the Machine Era and Guidance for AI Regulation." *AI & Society* 35, no. 3 (2009): 635-644.

Kant, Immanuel. *Critique of Practical Reason*. Translated by Mary Gregor. Cambridge: Cambridge University Press, 2015.

Kant, Immanuel. *Lectures on Ethics*. Translated by L. Infield. New York: Harper & Row, 1963.

Kant, Immanuel. *The Groundwork for the Metaphysics of Morals*. Translated by Allen W. Wood. New Haven and London: Yale University Press, 2002.

Kleene, Stephen Cole. *Introduction to Metamathematics*. Amsterdam: North-Holland, 1952.

Leibniz, Gottfried Wilhelm. "Principles of Nature and Grace, Based on Reason." In Gottfried Wilhelm Leibniz, *Philosophical Papers and Letters*, edited by Leroy E. Loemker. Dordrecht: Springer, 1989.

Leibniz, Gottfried Wilhelm. *Dissertatio de arte combinatoria*. Paris: Hachette Livre-BNF, 2018.

Levy, David. "The Ethical Treatment of Artificially Conscious Robots." *International Journal of Social Robotics* 1, no. 3 (2009): 209-216.

Malle, Bertram F., Thapa Stuti Magar, and Matthias Scheutz. "AI in the Sky: How People Morally Evaluate Human and Machine Decisions in a Lethal Strike Dilemma." In *Robotics and Well-Being*, edited by Maria Aldinhas Ferreira, João Silva Sequeira, Gurvinder Singh Virk, Mohammad Tokhi Osman, and Ender E. Kadar, 111-133. Cham: Springer, 2019.

McCarthy, John, and Patrick J. Hayes. "Some Philosophical Problems from the Standpoint of Artificial Intelligence." In *Machine Intelligence*, vol. 4, edited by Bernard Meltzer, and Donald M. Michie, 463-502. Edinburgh: Edinburgh University Press, 1969.

McCulloch, Warren S., and Walter H. Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 5 (1943): 115-33.

Minsky, Marvin. *Computation: Finite and Infinite Machines*. New Jersey: Prentice-Hall, 1967.

Mitcham, Carl. *Thinking through Technology: The Path between Engineering and Philosophy*. Chicago: The University of Chicago Press, 1994.

Moor, James H. "The Nature, Importance, and Difficulty of Machine Ethics." *IEEE Intelligent Systems* 21, no. 4 (2006): 18-21.

Newell, Allen, and Herbert Alexander Simon. "Computing Science as Empirical Enquiry: Symbols and Search." *Communications of the Association for Computing Machinery* 19 (1976): 113-126.

Newell, Allen, and Herbert Alexander Simon. "GPS-A Program that Simulates Human Thought." In *Lernende Automaten*, edited by Heinz Billing, 109-124. Münich: Oldenburg, 1961.

Newell, Allen, and Herbert Alexander Simon. "The Logic Theory Machine: A Complex Information-Processing System." *IRE Transactions on Information Theory* 2, no. 3 (1956): 61-79.

Newell, Allen, and Herbert Alexander Simon. *Current Developments in Complex Information Processing: Technical Report P-850*. Santa Monica, CA: Rand Corporation, 1956.

Newell, Allen, and John Crosley Shaw. "Programming the Logic Theory Machine." In *IRE-AIEE-ACM '57 (Western): Papers Presented at the February 26-28, 1957, Western Joint Computer Conference: Techniques for Reliability*, 230-240. New York: Association for Computing Machinery, 1957.

Newell, Allen, John Crosley Shaw, and Herbert Alexander Simon. "Element of a Theory of Human Problem Solving." *Psychological Review* 65 (1958): 151-166.

Newell, Allen. "Physical Symbol Systems." *Cognitive Science* 4 (1980): 135-183.

Owen, Jonathan, and Richard Osley. "Bill of Rights for Abused Robots: Experts Draw up an Ethical Charter to Prevent Humans Exploiting Machines." *The Independent*, September 17, 2011, https://www.independent.co.uk/news/science/bill-of-rights-for-abused-robots-5332596.html.

Pylyshyn, Zenon W., ed. *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*. Norwood, NJ: Ablex, 1987.

Rawls, John. "Justice as Fairness: Political not Metaphysical." *Philosophy and Public Affairs* 14 (1985): 223-251.

Rawls, John. *A Theory of Justice*, 2nd edition. Cambridge, MA: The Belknap Press of Harvard University Press, 1999.

Rawls, John. *A Theory of Justice*. Cambridge, MA: The Belknap Press of Harvard University Press, 1971.

Rawls, John. *Justice as Fairness: A Restatement*. Cambridge, MA: Harvard University Press, 2001.

Ross, William David. *The Right and the Good*. Oxford: Clarendon Press, 1930.

Russell, Stuart, and Max Tegmark. "Autonomous Weapons: An Open Letter from AI & Robotics Researchers." *Future of Life Institute*. https://futureoflife.org/open-letter-autonomous-weapons/.

Singer, Peter. "All Animals Are Equal." In *Animal Ethics: Past and Present Perspectives*, edited by Evangelos D. Protopapadakis, 163-178. Berlin: Logos Verlag, 2012.

Singer, Peter. *Animal Liberation: A New Ethics for our Treatment of Animals*. New York: New York Review of Books, 1975.

Singer, Peter. *Practical Ethics*, 2nd edition. Cambridge: Cambridge University Press, 1993.

Soares, Nate. "The Value Learning Problem." In *Artificial Intelligence, Safety and Security*, edited by Roman V. Yampolskiy, 89-97. Boca Raton, FL: CRC Press, 2019.

Sparrow, Robert. "Killer Robots." *Journal of Applied Philosophy* 24, no. 1 (2007): 62-77.

Sparrow, Robert. "The Turing Triage Test." *Ethics and Information Technology* 6 (2004): 201-213.

Taylor, Steve, Brian Pickering, Michael Boniface, Michael Anderson, David Danks, Asbjørn Følstad, Matthias Leese, Vincent Müller, Tom Sorell, Alan Winfield, and Fiona Woollard. "Responsible AI – Key Themes, Concerns & Recommendations for European Research and Innovation." *Zenodo*, July 2, 2018.

Tegmark, Mark. "Benefit and Risks of Artificial Intelligence." *Future of Life Institute*. https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/.

Tegmark, Mark. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Knopf, 2017.

Thompson, Nicholas. "Will Artificial Intelligence Enhance or Hack Humanity?" *Wired*, April 20, 2019, https://www.wired.com/story/will-artificial-intelligence-enhance-hack-humanity/.

Tooley, Michael. "In Defense of Abortion and Infanticide." In *The Abortion Controversy: A Reader*, edited by Luis P. Pojman, and Francis J. Beckwith, 186-213. Boston, MA: Jones & Bartlett, 1994.

Turing, Alan Mathison. "Computing, Machinery and Intelligence." *Mind* 59 (1950): 433-460.

Turing, Alan Mathison. "Intelligent Machinery." In *Machine Intelligence 5*, edited by B. Meltzer, and D. M. Michie, 3-23. Edinburgh: Edinburgh University Press, 1969.

Turing, Alan Mathison. "On Computable Numbers, with an Application to the Entscheidungsproblem." In *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life*, edited by Jack B. Copeland, 58-90. Oxford: Oxford University Press, 2004.

Turing, Alan Mathison. "On Computable Numbers, with an Application to the Entscheidungsproblem. A Correction." *Proceedings of the London Mathematical Society* 43 (1938): 544-546.

Wallace, Gregory. "Elon Musk Warns against Unleashing Artificial Intelligence 'Demon.'" *CNN Business*, October 26, 2014, https://money.cnn.com/2014/10/26/technology/elon-musk-artificial-intelligence-demon/.

Warren, Mary Anne. "On the Moral and Legal Status of Abortion." In *Contemporary Moral Problems*, edited by J. White, 144-155. Belmont, CA: Wadsworth/Thompson Learning, 2003.

Wheeler, Michael. "Cognition in Context: Phenomenology, Situated Robotics, and the Frame Problem." *International Journal of Philosophical Studies* 16, no. 3 (2008): 323-349.

Wheeler, Michael. *Reconstructing the Cognitive World: The Next Step*. Cambridge, MA: MIT Press, 2005.

Winner, Langdon. "Citizen Virtues in a Technological Order." *Inquiry* 35, nos. 3-4 (1992): 341-361.

Winner, Langdon. "Technè and Politeia: The Technical Constitution of Society." In *Philosophy of Technology*, edited by Paul T. Dubrin, and Friedrich Rapp, 97-111. Dordrecht, Boston, Lancaster: D. Reidel, 1983.

Yampolskiy, Roman. "Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach." In *Philosophy and Theory of Artificial Intelligence. Studies in Applied Philosophy, Epistemology and Rational Ethics*, edited by Vincent Müller, 389-396. Berlin, Heidelberg: Springer, 2013.

Yudkowsky, Eliezer. "Complex Value Systems in Friendly AI." In *Artificial General Intelligence*, edited by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 388-393. Berlin, Heidelberg: Springer, 2011.

Yudkowsky, Eliezer. "The Value Loading Problem." *EDGE*, July 12, 2021, https://www.edge.org/response-detail/26198.