IRIE
International Review of Information Ethics Vol. 31 (08/2022)

Authors: Murray, J., Rushby, J., Sanchez, D.

Controlling Smart Technology: A Brief Review of Some Ethical Challenges

Abstract:

This work explores some of the key challenges for incorporating appropriate ethical behavior and respect for social norms in highly-autonomous intelligent machines. Starting with the implications of Azimov's Three Laws of Robotics, we discuss the tradeoffs in human values, as encapsulated in the popular philosophical 'Trolley Problem'. We then examine some of the concerns for how smart systems model their worlds and their interactions with humans. The paper continues with a review of artificial moral agency and *phronesis*, and the techniques being proposed for implementing such agents. It concludes with some notes on recent research directions. This material is primarily an adaptation of work prepared by two of the authors, as part of a report of a series of workshops on machine consciousness, which were held during the summer of 2017.

Agenda:

Author(s):

Dr. John Murray:

- San José State University, San José, California 95192, USA
- ORCID: 0000-0002-6789-7380
- ✉ *jxm@acm.org*

Dr. John Rushby:

- Computer Science Laboratory, SRI International, Menlo Park, California 94025, USA
- ORCID: 0000-0002-6604-9953

Dr. Daniel Sanchez:

- Calidris Partners, Menlo Park, California 94026, USA
- ORCID: 0000-0003-2461-6574

ISSN 1614-1687

Background

In this paper, we draw upon some discussions and findings from a series of Technology and Consciousness Workshops (T&C Workshops), which were hosted by SRI International and undertaken during the summer of 2017. [1] Eight one-week-long workshops were held in various locations, with a total of 50 research specialists and global thought leaders participating in one or more of the workshops. Their disciplines spanned a variety of interests, including neuroscience, robotics and artificial intelligence, computer science, and contemporary physics. Additional perspectives included philosophy of mind, cognitive science, and spiritual and religious traditions.

A broad-brush plenary session opened the workshop series, followed by six in-depth sessions, each of which concentrated on specific research approaches to machine consciousness. Each week-long event consisted of several one-hour specialist presentations and discussions, followed by breakout groups, results sharing, and consensus outputs. The series concluded with a closing plenary event, which synthesised the overall findings from the series and set out guidance for future research directions. The accompanying tables describe the topics and key participants, as well as the set of research questions to be addressed, for each session.

Table 1. T&C Workshop Series: Session topics with key speakers and participants

Workshop Session	**Selected Speakers and Participants**	**Workshop Session**	**Selected Speakers and Participants**
Opening Plenary: Introduction to Consciousness	David Chalmers, Ian Horswill, Antonio Chella, John Sullins, Paul Syverson	**1)** Philosophical Perspectives on Consciousness	Hank Barendregt, Selmer Bringsjord, David Rosenthal, Robin Zebrowski
2) Embodiment and Culture	Alva Noe, Bill Rowe, Susan Kaiser-Greenland, Earth Erowid, John Rushby	**3)** Neuroscience and Cognitive Science	Christof Koch, Mark Bickhard Susan Schneider, Guilio Tononi, Chris Connolly
4) Computation and Logic Approaches	Susan Blackmore, David Gamez, Subhash Kak, David Israel, Natarajan Shankar	**5)** First-Person and Non-Western Perspectives	David Presti, John Murray, Marcia Grabowecky, David Sahner, Damien Williams
6) Artificial Intelligence and Machine Consciousness	Jonathan Moreno, Shannon Vallor, Daniel Sanchez, Ron Rensink, Karen Myers	**Closing Plenary:** Summaries, Synthesis, and Research Directions	Joanna Bryson, Naveen Sundar Govindarajulu, Julia Mossbridge, Owen Holland

Table 2. T&C Workshop Series: Coordinated tasked objectives for each workshop session

Characterizations of consciousness An interdisciplinary dialogue was promoted to achieve a common ground working definition of consciousness across fields, for the purpose of the specific workshop.	***Potential mechanistic underpinnings of consciousness*** Provided with this definition of consciousness, participants could begin exploring the necessary requirements and potential basis for the existence of consciousness.
Metrics of consciousness Can we identify some reasonable, agreed upon, metrics of consciousness that would allow us to assess the presence and level of consciousness in biological and machine agents?	***Perspectives on machine consciousness*** Consider the (speculative) implications of future machine consciousness, particularly for the safety and welfare of inhabitants of future societies.

[1] Williams & Murray: Technology and Consciousness Workshops.

Among the topics that frequently arose were the ethical and moral issues related to the control of highly-autonomous, decision-making machines, which is the principal focus area of this paper.

Much of the current research activity on machine ethics and autonomous systems focuses on near-term technological limitations and challenges, such as biases in face-recognition, job applicant selection, insurance risks, etc. In contrast, in the present work, we explore some issues that are more often related to speculative future systems – those that have achieved greater independence and experience significant levels of near-human sentience or 'Strong AI'.

Some thought experiments in philosophy concern *zombies* – hypothetical entities that replicate the functions of the human brain, but without any underlying consciousness. In this context, we posit that designing and promulgating practical means of ethical control for highly-intelligent machines should be an urgent source of concern for smart technology researchers and developers. This work is adapted from several segments of the final report on the T&C Workshops, in particular Chapter 8, Ethics for Control of Conscious Technology.[2]

Introduction

In the process of designing and deploying highly-autonomous, decision-making machines, it seems prudent to 'bake in' pervasive means of control into such technologies. We control the behavior of humans in society (and possibly pets in a household) by inculcation of ethical and social norms, reinforced by praise and censure, reward and punishment. So, in general, it appears that an ethical framework should be part of all advanced technology and referenced in all its decisions. This requires a built-in notion of *right* and *wrong* and knowledge of the ethical norms and the laws of its environment, together with some way to adjust future behavior by means that resemble praise and censure, or rewards and punishments. The alternative is technology that "does what it does" with no way to curb undesired behavior other than adjusting its algorithms, and no organizing principle for doing so.

Highly-intelligent systems may potentially develop goals and priorities that conflict with those of human society. Thus, they should have overarching constraints built in from the very beginning to forestall this danger. Since we cannot know the particular circumstances that may arise, the constraints need to be general and overarching, rather like Asimov's "Three Laws of Robotics." These three laws are (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) A robot must obey orders given to it by human beings except where such orders would conflict with the First Law (3) A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.[3]

These laws were a plot device in Asimov's science fiction stories, which often concerned the unintended consequences of these plausible-sounding laws, in the process indicating that construction of suitable constraints may be challenging. The question then arises about how to base such constraints, possibly by grounding them on human ethics.[4]

A dissenting opinion, advocating explicit reasoning about hazardous outcomes was provided by Yampolskiy, who proposed a new science of *AI Safety Engineering*.[5] Of course, ethics have been studied and debated for millennia, without achieving consensus. Nonetheless, some broad general principles are known. Ethics are the basic rules by which societies maintain order and cohesion; however, some very successful societies have elements that others find repugnant: for example, Ancient Greece used slaves (Aristotle wrote of "natural slaves") and Ancient Rome had execution as a form of public entertainment. Hence, it seems that the moral foundations of ethics are not universal. Nonetheless, modern "experimental ethics" finds that human moral sense is built on five basic principles that do seem universal: care, fairness, loyalty/ingroup, authority/respect,

[2] Rushby & Sanchez: Technology and Consciousness.

[3] Azimov: I Robot.

[4] Yu et al: Building Ethics into AI.

[5] Yampolskiy: AI Safety Engineering.

and sanctity/purity.[6] What is not universal is preference and weighting among the principles, which behave rather like the five basic senses of taste: different societies and individuals prefer some, and some combinations, to others. For example, western liberals stress fairness while conservatives favor authority.

Trolley Problems

Even if an agreed weighting of the basic principles were built in to advanced technology, it may not be obvious how to apply it. For example, a self-driving car might be confronted by a vehicle crossing against the lights and the choices are to crash into it, likely killing or injuring the occupants of both vehicles, or to swerve onto the sidewalk, likely killing pedestrians. The fairness principle might argue that all lives are equal and utilitarianism might then suggest a decision that minimizes the probable injuries. On the other hand, the care principle might argue that the system has a special responsibility for its own passengers and should seek a solution that minimizes their harm.

Trolley problems are thought experiments used to probe human judgments on these ethical dilemmas.[7] The classic problem posits a runaway street car/trolley that is heading toward a group of five people. You are standing by a switch/point and can throw this to redirect the trolley to a different track where it will hit just one person. Most subjects say it is permissible, indeed preferable, to throw the switch, even though it will injure an innocent who would otherwise be unharmed. However, a variant on the original trolley problem has you and another person standing by the track and suggests that you bring the trolley to a halt, and save the five, by pushing the other person onto the track in front of the trolley. Most subjects will say this is ethically unacceptable, even though it is equivalent to the first case by utilitarian accounting. These examples illustrate the "Doctrine of Double Effect" (DDE), i.e. it is ethically acceptable to cause harm as an unintended (even if predictable) side effect of a (larger) good. The first case satisfies the doctrine, but the second violates the "unintended" condition.

Experimental systems have been developed that can represent and reason about ethical principles such as DDE and these have been applied to trolley problems, including some that involve self-harm (e.g., throwing yourself in front of the trolley) and thereby violate the *unintended* aspect of DDE.[8] [9] It is claimed that fairly sophisticated logical treatments (e.g., intensional logics, counterfactuals, deontic modalities) are needed to represent ethical scenarios, and these might be additional to what is needed for the primary functions of the system (hence, must be introduced explicitly). Other recent work formalizes Kant's categorical imperative (humans must be treated as ends, not as means), which requires a treatment of causality,[10] while another speculates on application of ethics to autonomous cars.[11]

World Models and Communities

There is more to ethical systems than the simple application of ethical rules: the underlying model of the world should have a certain neutrality that may be hard to ensure. For example, a system that interacts with humans may need models of race and gender. Whether these are programmed or learned, they may unwittingly incorporate bias. In order to interact effectively, an artificial theory of mind may need explicitly to construct and consider biased models. So how can we ensure that possibly-biased models do not affect outcomes? Perhaps some judgments should be invariant under different assumptions about self and others: that is, the system should explicitly repeat its calculations under different assumptions as a computational

[6] Haidt: The Righteous Mind.

[7] Jarvis: The Trolley Problem.

[8] Bringsjord: Toward a General Logicist Methodology for Engineering Ethically Correct Robots. .

[9] Govindarajulu et al: On Automating the Doctrine of Double Effect.

[10] Lindner & Bentzen: A Formalizaion of Kant's Second Formulation of the Categorical Imperative.

[11] Kulicki et al: Towards a Formal Ethics for Autonomous Cars.

approximation to Rawls' "Veil of Ignorance".[12] Also, beyond interacting directly with humans, we need to take into account the possibility that a truly rampant technological system could pose many other broad societal hazards. For example, it could damage our infrastructure, undermine our institutions, or our trust in them.

In addition to ethics, highly-intelligent technological systems should also follow the laws and cultural conventions of their community. There is a long history of work on formalizing and reasoning about legal systems.[13] There will surely be circumstances where the law conflicts with some interpretation of ethics, or with the mission objective, so a system constrained by several such overarching frameworks must have a means of resolving conflicts. Individually and in total, these are challenging objectives.

Humans, endowed with an understanding of local ethics and of the law, sometimes make bad judgments, or resolve conflicts among competing ethical principles, in ways that society finds unsatisfactory. Various forms of censure and punishment provide means to correct such errant behavior and it seems that technological systems should also be subject to adjustment and tuning in similar ways. An important question then is what is the "accounting method" that guides such adjustments: is it just some internal measure, or is there some societal score-keeping that has wider significance? In a work commissioned by the US government during WWII, the anthropologist Ruth Benedict proposed a distinction between "guilt cultures" (e.g., the USA) and "shame cultures" (e.g., Japan).[14] This distinction is widely criticized today, but modern reputation systems, as employed for EBay sellers, Uber drivers, etc. can be seen as mechanizing some aspects of shame culture. Indeed, China's Social Credit system[15] extends this technique to the whole society. However, such an approach might usefully provide a framework for societal control of technological systems; the idea being that the technological system should value its reputation and adjust its behavior to maximize this.

Autonomy and Free Will

Being held responsible for our actions and subject to punishment and reward seems to require that we are free to act thus or otherwise. We generally assume that humans have free will, but what about technological systems? And if they do not have free will, can they be subject to the constraints of ethics? Human free will is an immensely difficult subject, which Hume called the most contentious problem in all of metaphysics.[16] It is next to impossible to reconcile the commonsense ("libertarian" or "contra-causal") notion of free will – that our decisions are uncaused causes – with materialism. In a material universe, what happens next is determined by what happened before, so how can such determinism be suspended while I make a decision? Even the probabilistic determinism induced by quantum effects, does not open the door to free will; a random or nondeterministic choice is no more free than a deterministic choice.[17] We are all part of the material world, so our decisions are determined by our current state (or are subject to quantum randomness). As Johnson noted, it only "feels like" I made a free choice: "all theory is against the freedom of the will; all experience is for it".[18] Thus, most philosophers accept only a weaker form of free will ("compatibilism") in which conscious decisions do cause actions, but those decisions are not themselves uncaused causes. That is, nothing prevents our deciding on one thing or the other, but the actual choice is (probabilistically) determined: "we can do what we will, but we cannot will what we will", as Schopenhauer has observed.[19]

Nonetheless, in everyday life we still attribute human acts to free will and we praise or punish accordingly. Philosophers accept this as a useful fiction, as experiments show that subjects primed to see free will as

[12] Rawls: A Theory of Justice.

[13] Von der Lieth Gardner: An AI Approach to Legal Reasoning.

[14] Benedict: The Chrysanthemum and the Sword.

[15] Kobie: The Complicated Truth About China's Social Credit System.

[16] Hume: An Enquiry Concerning Human Understanding

[17] Greenblatt: The Swerve: How the World Became Modern.

[18] Boswell: The Life Of Samuel Johnson LL.D.

[19] Schopenhauer: On the Freedom of the Will.

illusory are more likely to misbehave[20] or to become fatalistic. But, if it is hard to impute free will to humans, it is even harder to impute it to technology which, after all, is just a pile of *Franken-algorithms*.[21] However, as with humans, it is a useful fiction to treat intelligent technology (or any complex system endowed with learning) as if it had free will and hold it responsible for its actions, since its behavior adapts over time as a result of its "life experiences" and rewards and punishment can be a significant input to those adaptations.

Artificial Phronesis and Moral Agents

Our discussion thus far has focused on a rather basic concern – ensuring that advanced technological systems do us no harm. But some such systems might be intended to do positive good – robots to provide assistance and companionship to the elderly, for example. Ethical frameworks to prevent harm might therefore need to be generalized so that technology can enable us to flourish rather than merely survive. This consideration leads us to move beyond simply instilling technology with ethical principles for safety and control purposes, and travel onwards to the larger landscape of *digital phronesis*.[22] Phronesis is a term from Aristotle that refers to ethical wisdom. Several research groups have developed experimental systems for investigating artificial moral agents. These include: the N-Reasons platform, which is used to study ethical reasoning in humans,[23] and the Control Closure tool, which encapsulates ethical protocols.[24]

Another source of inspiration in this area is the Global Workspace Theory (GWT) of consciousness, which combines elements of biology and functionalism.[25] The functionalist aspect offers an architecture for mental activities, where many unconscious mental processes read and write to a global working memory that is selectively attended to, and consciousness corresponds to a *spotlight* of attention on this workspace. The biological aspect associates various elements of brain physiology (e.g., cortical areas, gamma synchrony) in the realization of this architecture. The Learning Intelligent Distribution Agent (LIDA) is based upon GWT.[26]

It is acknowledged that incorporating ethics at multiple levels in highly-complex autonomous systems is by no means a trivial task. In particular, one would likely want to construct formally verified safeguards at the operating system level. Attempting to program in the entire space of ethical quandaries may be infeasible, and attempting to build in general ethical theories is likely just as difficult. It may be that learning artificial phronesis is a reasonable approach, but it would be challenging to implement formally. To go a step further, if an artificial moral agent were to be created, the question arises about how to assess or measure it.

In the context of measuring machine consciousness, theorists have suggested a variety of tests, including; assessing theory of mind (TOM), measuring φ (phi; per Information Integration Theory[27]), extended Turing tests, games, and first person tests. But in the final analysis, if a machine can be designed that is human-like in every material way but without consciousness, shouldn't it at least be owed rights and privileges analogous to that of animals or ecosystems? As humans, at the very least, such a basic duty of care would seem to be our ethical responsibility.

[20] Cave: There’s No Such Thing as Free Will.

[21] Smith: Franken-Algorithms.

[22] Sullins: Artificial Phronesis and the Social Robot.

[23] Danielson: Designing a Machine to Learn About the Ethics of Robotics.

[24] Turilli: Ethical Protocols Design.

[25] Baars: Global Workspace Theory of Consciousness.

[26] Wallach et al: Consciousness and Ethics.

[27] Tononi et al: Integrated Information Theory.

Related Work and Future Directions

In the twentieth century, many academic debates about AI centered upon issues of architecture, semantics, and framing. More recently, discussions about AI consciousness and ethics have come to the fore, and now involve broader participation from other disciplines, including neuroscientists, lawyers, economists, etc.[28] In this context, the SRI Technology & Consciousness Workshop series in 2017 was a significant catalyst for further events on the topic.

Since 2014, the biennial *Robophilosophy* conferences have provided a broad forum for exploring the social and ethical impacts of advanced autonomous technical systems. The 2018 conference, which was held in Vienna, offered some T&C participants the opportunity to continue these discussions, and to promote the topics more widely in this community.[29] *Towards Conscious AI Systems*, a three-day symposium in 2019 at Stanford University that attracted forty-four papers included new work from T&C Workshop participants, as well as added material from other AI and consciousness research teams.[30]

In 2020, a special virtual T&C meeting was held in conjunction with the Euroscience Open Forum event in Trieste.[31] The meeting focused on recent artistic and creative explorations in robotics and autonomous systems, and examined topics like the role of anthropomorphism in AI communications and choreographic improvisations in robot movement.[32]

Recent research by T&C participants envisions a science of consciousness that would enable us to make accurate predictions about the consciousness of humans, animals and machines.[33] Another thread of related research centres upon the need to develop ethical standards for robotic nudging systems. There is an ongoing effort to build practical approaches for designers to use when building robotic systems to the highest ethical standards.[34]

As noted earlier, most current research efforts in machine ethics and autonomous systems are concerned with near-term challenges like biases in face-recognition, insurance risks, etc. In this paper, we have looked beyond this timeframe towards more speculative future systems, a time when machines have achieved greater independence and approach near-human sentience cpapbilities.

Acknowledgements

The authors would like to acknowledge the generous funding support provided by the anonymous sponsors of the SRI Technology & Consciousness Workshop series, as well as the diligent contributions of the participants, upon whose expertise and scholarship this paper is partially based.

[28] Dietrich et al: The AI Wars.

[29] Coeckelbergh, M. et al. Envisioning Robots in Society.

[30] Chella et al: Towards Conscious AI Systems.

[31] ESOF2020: EuroScience Open Forum, Trieste Italy.

[32] Murray & Chella: Creative Expolrations in AI, Robotics, and Autonomous Systems.

[33] Gamez: Human and Machine Consciousness.

[34] Sullins & Dougherty: Ethical Nudging of Users while they Interact with Robots.

References

Asimov, I. Runaround. In: I Robot. Gnome Press. (1950)

Baars, B. J. Global workspace theory of consciousness. Progress in Brain Research, (2005)

Benedict, R. The Chrysanthemum and the Sword. Houghton Mifflin. (1946)

Boswell, J. The Life Of Samuel Johnson, LL.D. Everyman's Library (Knopf), NY. (1811)

Bringsjord, S. et al. Toward a general logicist methodology for engineering ethically correct robots. IEEE Intelligent Systems, 21(4):38–44. (2006)

Cave, S. There's no such thing as free will. The Atlantic. June (2016)

Chella, A. et al. Towards Conscious AI Systems. AAAI Spring Symposium Series, Stanford CA, (2019)

Coeckelbergh, M., Seibt, J., et al. Envisioning Robots in Society – Power, Politics, and Public Space. Proceedings of Robophilosophy, Vienna Austria (2018).

Danielson, P. Designing a machine to learn about the ethics of robotics: The N-Reasons platform. Ethics and information technology, 12(3):251–261. (2010)

Dietrich, E. et al. The AI Wars, 1950–2000, and Their Consequences, Journal of Artificial Intelligence and Consciousness, 9(1): 127-151. (2022). doi.org/10.1142/S2705078521300012

ESOF2020. EuroScience Open Forum, Trieste Italy. September 2020.

Gamez, D. Human and Machine Consciousness. Open Book Publishers. (2018)

Govindarajulu, N. et al. On automating the doctrine of double effect. arXiv. (2017)

Greenblatt, S. The Swerve: How the World Became Modern. WW Norton & Co. (2011)

Haidt, J. The Righteous Mind: Why Good People Are Divided by Politics and Religion. Vintage. (2013)

Hume, D. An Enquiry Concerning Human Understanding. (1748)

*Jarvis Thomson, Judith. The Trolley Problem. Yale Law Journal. **94** (6). (1985)*

Kobie, N. The complicated truth about China's social credit system. Wired UK. June (2019)

Kulicki, P. et al. Towards a formal ethics for autonomous cars. ibid. (DEON). (2018)

Lindner, F. & Bentzen, M. A formalizaion of Kant's second formulation of the Categorical Imperative. In 14th Intl Conf on Deontic Logic and Normative Systems (DEON). (2018)

Murray, J. & Chella, A. Creative Explorations in AI, Robotics, and Autonomous Systems. ESOF. (2020)

Rawls, J. A Theory of Justice. Belknap Press/Harvard University Press. (1971)

Rushby, J. and Sanchez, D. Technology and Consciousness, SRI International. (2018)

Schopenhauer, A. On the Freedom of the Will. Oxford: Basil Blackwell. (1839)

Smith, A. Franken-algorithms. The Guardian, Aug 29 2018.

Sullins, J. Artificial phronesis and the social robot. In Seibt, J., Norskov, M., & Andersen, S. S. (eds), What Social Robots Can and Should Do, pages 37–39. IOS Press. (2016)

Sullins, J & Dougherty, S. Ethical Nudging of Users While They Interact with Robots. In Norkov, M. et al (eds), Culturally Sustainable Social Robotics: Proceedings of Robophilosophy, Aarhus Denmark (2020)

Tononi, G. et al. Integrated information theory: From consciousness to its physical substrate. Nature Reviews Neuroscience, 17(7):450–461. (2016)

Turilli, M. Ethical protocols design. Ethics and Information Technology, 9(1):49–62. (2007)

Von der Lieth Gardner, A. An AI Approach to Legal Reasoning. MIT Press. (1987)

Wallach, W., Allen, C., and Franklin, S. Consciousness and ethics: Artificially conscious moral agents. International Journal of Machine Consciousness, 3(01):177–192. (2011)

Williams, D. and Murray, J. Technology and consciousness workshops: An introductory overview. Journal of Artificial Intelligence and Consciousness, 7(01): 133-140. (2020) doi.org/10.1142/S2705078520710010

Yampolskiy, R. V. Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In Philosophy and Theory of Artificial Intelligence, volume 5 of Studies in Applied Philosophy, Epistemology and Rational Ethics, pages 389–396. Springer. (2013)

Yu, H. et al. Building ethics into artificial intelligence. In Proc 27th Intl Joint Conf on Artificial Intelligence (IJCAI-18), Stockholm, Sweden. (2018)