

could also be inclusive with respect to the communities that will contribute to this emerging body of knowledge, opening special issues to the multiple communities identified above. Importantly, such an OA platform should also expose data, analysis methods and workflows to assure reproducibility and other modes of data reuse. Achieving these aims in an international context that satisfies the requirements recently introduced by the European Union's General Data Protection Regulation is already an urgent undertaking.

How Can the Field Advance and Mature Along these Lines?

While some of the above-mentioned efforts could be facilitated by small organized groups of investigators who may obtain funds for national or international research coordination, our experience suggests that developing an organization infrastructure with these goals will produce larger dividends. Independently, we argue that these considerations should not be left to the initiative or discretion of funding-agency personnel. In a nutshell, program directors and administrators at private foundations or governmental agencies have several competing interests and demands on funding, some of which may not fully capture the importance of international coordination and collaboration, or may not sufficiently emphasize industry collaborations. They may therefore be less effective in devising potential strategic plans of actions as compared with groups of experts in these areas. Rather, an internationally oriented organization (or advisory board; see [6]) that can maintain effective connections with public and private funding bodies would be more effective in advancing such efforts. Such an organization would also be in position to identify and advocate for research areas suitable for larger-scale research coordination both in terms of paradigmatic and conceptual synthesis, and basic research. Such efforts may well surpass the usual scale of funded

research projects and could involve international coordination among multiple agencies.

Cognitive Neuroscience can capitalize on such emerging opportunities to improve its science and increase its relevance. However, such efforts would require community-level coordination that is not yet in place.

Disclaimer Statement

No prior or current NSF personnel were consulted with, nor commented on, any version of this manuscript. The opinions expressed are the authors' alone and do not reflect in any way the position of the NSF.

¹Center for Mind/Brain Sciences (CIMEC), University of Trento, Rovereto, TN, Italy

²Department of Psychology, The University of Chicago, Chicago, IL, USA

*Correspondence: uri.hasson@unitn.it (U. Hasson).
<https://doi.org/10.1016/j.tics.2019.02.007>

© 2019 Elsevier Ltd. All rights reserved.


References

- Waytowich, N. *et al.* (2018) Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials. *J. Neural Eng.* 15, 066031
- Angrick, M. *et al.* (2019) Speech synthesis from ECoG using densely connected 3D convolutional neural networks. *J. Neural Eng.* Published online 4 March 2019. <http://dx.doi.org/10.1088/1741-2552/ab0c59> Published online 4 March 2019
- Barsalou, L.W. (2017) What does semantic tiling of the cortex tell us about semantics? *Neuropsychologia*, 105, 18–38
- Cichy, R.M. and Kaiser, D. (2019) Deep neural networks as scientific models. *Trends Cogn. Sci.* 23, 305–317
- Wang, Y. *et al.* (2016) Real-time full correlation matrix analysis of fMRI data. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 1242–1251, IEEE
- Gunsalus, C.K. *et al.* (2019) Overdue: a US advisory board for research integrity. *Nature*, 566, 173–175
- Kragel, J.E. *et al.* (2015) Neural activity in the medial temporal lobe reveals the fidelity of mental time travel. *J. Neurosci.* 35, 2914–2926
- Nunez, M.D. *et al.* (2017) How attention influences perceptual decision making: single-trial EEG correlates of drift-diffusion model parameters. *J. Math. Psychol.* 76, 117–130
- Turner, B.M. *et al.* (2017) Approaches to analysis in model-based cognitive neuroscience. *J. Math. Psychol.* 76, 65–79
- Nelson, M.J. *et al.* (2017) Neurophysiological dynamics of phrase-structure building during sentence processing. *Proc. Natl. Acad. Sci. U. S. A.* 114, E3669–E3678

- Broderick, M.P. *et al.* (2018) Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.* 28, 803–809
- Avesani, P. *et al.* (2019) The open diffusion data derivatives, brain data upcycling via integrated publishing of derivatives and reproducible open cloud services. *PsyArXiv* Published online 21 December 2018. <http://dx.doi.org/10.31234/osf.io/y82n7> Published online 21 December 2018
- Gorgolewski, K. *et al.* (2017) OpenNeuro—a free online platform for sharing and analysis of neuroimaging data. In *Proceedings from the 23rd Annual Meeting of the Organization for Human Brain Mapping (OHBM)*, Vancouver, Canada, p. 1677
- Kennedy, D.N. *et al.* (2019) Everything matters: the repro-prim perspective on reproducible neuroimaging. *Front. Neuroinform.* 13, 1

Science & Society

Holding Robots Responsible: The Elements of Machine Morality

Yochanan E. Bigman ^{1,*}, Adam Waytz,² Ron Alterovitz,³ and Kurt Gray¹

As robots become more autonomous, people will see them as more responsible for wrongdoing. Moral psychology suggests that judgments of robot responsibility will hinge on perceived situational awareness, intentionality, and free will, plus human likeness and the robot's capacity for harm. We also consider questions of robot rights and moral decision-making.

Advances in robotics mean that humans already share roads, skies, and hospitals with autonomous machines. Soon, it will become commonplace for cars to autonomously maneuver across highways, military drones to autonomously select missile trajectories, and medical robots to autonomously seek out and remove tumors. The actions of these autonomous machines can spell life and death for humans [1], such as when self-driving vehicles kill pedestrians.

When robots harm humans, how will we understand their moral responsibility?

Morality and Autonomy

Philosophy, law, and modern cognitive science all reveal that judgments of human moral responsibility hinge on autonomy [2,3]. This explains why children, who appear less autonomous than adults, are held less responsible for wrongdoing. Autonomy is also likely crucial in judgments of robot moral responsibility [4,5]. The reason people ponder and debate the ethical implications of drones and self-driving cars (but not tractors or blenders) is because these machines can act autonomously.

Admittedly, today's robots have limited autonomy, but it is an expressed goal of roboticists to develop fully autonomous robots: machine systems that can act without human input [6]. As robots become more autonomous, their potential for moral responsibility will only grow. Even as roboticists create robots with more 'objective' autonomy, we note that 'subjective' autonomy may be more important: work in cognitive science suggests that autonomy and moral responsibility are more matters of perception than objective truth [3].

Perceiving the Minds of Robots

For programmers and developers, autonomy is understood as a robot's ability to operate in dynamic real-world environments for extended periods of time without external human control [6]. However, for everyday people, autonomy is more likely tied to a robot's mental capacities. Some may balk at the idea that robots have (or will have) any human-like mental capacities, but people also long balked at the idea that animals had minds, and now think of them as having rich inner lives.

Of course, animals are flesh and blood, whereas machines are silicon and circuits, but research emphasizes that minds are always matters of perception [3,7]. The 'problem of other minds' means that the

thoughts and feelings of others are ultimately inaccessible, and so we are left to perceive them based upon context, cues, and cultural assumptions. Importantly, people do ascribe to machines at least some ability to think, plan, remember, and exert self-control [7,8], and as when judging humans, people make sense of the morality of robots based upon these ascriptions of mind [8].

How people see mind, that is, 'mind perception', predicts moral judgments [3], but mind perception is not monolithic: there are many mental abilities [8], some of which (e.g., the ability to plan ahead) are more relevant to autonomy and moral judgment than others (e.g., the ability to feel thirsty). Cognitive science has outlined these autonomy-relevant abilities as they concern humans, but only a subset of these are likely important for making sense of morality in autonomous machines. Here, we outline one subset of robot 'mental' abilities that likely seem relevant to autonomy (and therefore moral judgment).

Autonomous Elements Tied to Robot Morality

Situation Awareness

For observers to perceive a person as morally responsible for wrongdoing, that person must seem to be aware of the moral concerns inherent in the situation [9]. For example, a young child unaware of the danger of guns will not be held responsible for shooting someone. For a robot to be held responsible for causing harm, it will likely need to be seen as aware that its actions are indeed harmful. Although today's robots cannot appreciate the depths of others' suffering, they can at least understand some situational aspects. For example, robots can understand whether stimuli belong to protected categories, such as civilians for military drones, pedestrians for autonomous cars, and healthy organs for medical robots. People already ascribe some of this 'meaning-lite' understanding to machines [7], and we suggest

that greater ascriptions of situational awareness will increase perceptions of robot responsibility.

Intentionality

Agents are seen as more responsible for intentional actions than for unintentional actions, often because people infer a desire or a reason behind intentional acts [10]. Although people are unlikely to perceive robots as capable of desire, they do see robots as capable of intentionality, that is, holding a belief that an action will have a certain outcome [7]. This perception is consistent with robots' ability to evaluate multiple response options in the service of achieving a goal [11]. We suggest that the more people see robots as intentional agents, being able to understand and select their own goals, the more they will be ascribed moral responsibility.

Free Will

The ability to freely act, or to 'do otherwise' [2], is a cornerstone of lay judgments of moral responsibility [2]. Although robots are not seen as possessing a rich human-like free will, they are ascribed the ability to independently implement actions [7]. Consistent with this ascription, today's robots can independently execute action programs [11]; however, this independence is relatively constrained. The behavior of robots is predictable given the transparency of their (human-given) programming, and predictability undermines perceptions of free will [2]. Technological advances (e.g., deep neural networks) will likely render the minds of machines less transparent to both programmers and perceivers, thereby elevating perceptions of unpredictability. We suggest that as robotic minds become more opaque, people will see robots as possessing more free will, and ascribe them more moral responsibility.

Human Likeness

People perceive the mind of machines based on their abilities and behaviors,

but also on their appearance. The more human-like a machine looks, the more people perceive it as having a mind, a phenomenon called anthropomorphism [12]. Individuals vary in their tendency to anthropomorphize, but people consistently perceive more mind, and therefore more moral responsibility, in machines that look and act like humans [13]. We suggest that having human-like bodies, human-like voices, and human-like faces will all cause people to attribute more moral responsibility to machines.

Potential Harm

Even with powerful computational abilities, today's robots are limited in their ability to act upon the world. As technology advances, these increased capacities (e.g., the ability to walk, shoot, operate, and drive) will allow robots to cause more damage to humans. Studies reveal that observing damage and suffering lead people to search for an intentional agent to hold responsible for that damage [14]. If people cannot find another person to hold responsible, they will seek other agents, including corporations and gods [14], and infer the capacity for intention. This link between suffering and intention means that the more robots cause damage, the more they will seem to possess intentionality, and thus (as we outline above) lead to increased perceptions of moral responsibility. We therefore suggest that causing harm can amplify both perceptions of mind and judgments of moral responsibility.

Concluding Remarks and Future Implications

The future of robotics holds considerable promise, but it is also important to consider what today's semi-autonomous machines might mean for moral judgment. As Box 1 explores, even robots with some perceived mind can help shield their human creators and owners (e.g., corporations and governments) from

Box 1. Machines Can Shield Humans from Responsibility

When people harm others, they often try to avoid responsibility by pointing fingers elsewhere. Soldiers who commit heinous acts invoke the mantra that they were 'just following orders' from superior officers. Conversely, superior officers shirk responsibility by claiming that they did not actually pull the trigger. These excuses can work because perceived responsibility is often a zero-sum game. The more we assign responsibility to the proximate agent (the entity who physically perpetrated the harm), the less we assign responsibility to the distal agent (the entity who directed the harm), and vice versa [3].

As robots spread through society, they will more frequently become the proximal agent in harm-doing: collateral damage will be caused by drones and accidents will be caused by self-driving cars. Although humans will remain the distal agents who program and direct these machines, the more that people can point fingers at their autonomous robots, the less they will be held accountable for wrongdoing, a fact that corporations and governments could leverage to escape responsibility for misdeeds. Increasing autonomy for robots could mean increasing absolution for their owners.

responsibility. Today's machines are also capable of making some kind of moral decisions, and Box 2 explores whether people actually want machines to make these basic decisions.

Although we focus here on moral responsibility, we note that people might also see sophisticated machines as worthy of moral rights. While some might find the idea of robots rights to be ridiculous, the American Society for the Prevention of Cruelty to Robots and a 2017 European Union report both argue for extending some moral protections to machines. Debates about whether to recognize the personhood of robots often revolve around its impact on humanity (i.e., expanding the moral circle to machines may better protect other people), but also involves questions about whether robots

possess the appropriate mind required for rights. Although autonomy is important for judgments of moral responsibility, discussions of moral rights typically focus on the ability to feel. It is an open question whether robots will ever be capable of feeling love or pain, and relatedly, whether people will ever perceive these abilities in machines.

Whether we are considering questions of moral responsibility or rights, issues of robot morality may currently seem like science fiction. However, we suggest that now, while machines and our intuitions about them are still in flux, is the best time to systematically explore questions of robot morality. By understanding how human minds make sense of morality, and how people perceive the minds of machines, we can help society think more

Box 2. Do We Want Machines Making Moral Decisions?

Many discuss how robots should make moral decisions [1], but it is worth asking whether they should make moral decisions in the first place. For example, some argue that autonomous military robots (e.g., drones) should never independently make decisions about human life and death. However, others argue in favor of these autonomous military robots, suggesting that they could be programmed to follow the rules of war better than humans.

Putting these ethical debates in perspective, research reveals that people are reluctant to have machines make any moral decisions, whether in the military, the law, driving, or medicine [8]. One reason for people's aversion to machines making moral decisions is that they see robots as lacking a full human mind [7,8]. Without the full human ability to think and feel, we do not see robots as qualified to make decisions about human lives.

This aversion to machine moral decision-making has seem quite robust [8], but it may fade as the perceived mental capacities of machines advance [15]. As the autonomy of machines rises, people may become more comfortable with robots making moral decisions, although people may eventually wonder whether the goals of machines align with their own.

clearly about the impending rise of robots and help roboticists understand how their creations are likely to be received.

Acknowledgments

We thank Bertram Malle, Ilan Finkelstein, Michael Clamann, and an anonymous reviewer for comments on a draft of this paper. This work has been supported by the National Science Foundation award SPRF-1714298) to Y.E.B. by the National Science Foundation awards IIS-1149965 and CCF-1533844 to R.A., and a grant from the Charles Koch Foundation to K.G.

¹Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA

²Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA

³Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA

*Correspondence: ybigman@email.unc.edu (Y.E.E).

<https://doi.org/10.1016/j.tics.2019.02.008>

© 2019 Elsevier Ltd. All rights reserved.



References

- Awad, E. et al. (2018) The moral machine experiment. *Nature*, 563, 59–64
- Shariff, A.F. et al. (2014) Free will and punishment: a mechanistic view of human nature reduces retribution. *Psychol. Sci.* 25, 1563–1570
- Wegner, D.M. and Gray, K. (2017) *The Mind Club*, Viking
- Kim, T. and Hinds, P. (2006) Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *ROMAN 2006: The 15th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 80–85, IEEE
- van der Woerd, S. and Haselager, P. (2017) When robots appear to have a mind: the human perception of machine agency and responsibility. *New Ideas Psychol.* <http://dx.doi.org/10.1016/j.newideapsych.2017.11.001>
- Bekey, G.A. (2005) *Autonomous Robots: From Biological Inspiration to Implementation and Control*, The MIT Press
- Weisman, K. et al. (2017) Rethinking people's conceptions of mental life. *Proc. Natl. Acad. Sci. U. S. A.* 114, 11374–11379
- Bigman, Y.E. and Gray, K. (2018) People are averse to machines making moral decisions. *Cognition*, 181, 21–34
- Kissinger-Knox, A. et al. (2018) Does non-moral ignorance exculpate? Situational awareness and attributions of blame and forgiveness. *Acta Anal.* 33, 161–179
- Monroe, A.E. and Malle, B.F. (2017) Two paths to blame: intentionality directs moral information processing along two distinct tracks. *J. Exp. Psychol. Gen.* 146, 23–33
- Dudek, G. and Jenkin, M. (2010) *Computational Principles of Mobile Robotics*, Cambridge University Press
- de Visser, E.J. et al. (2016) Almost human: anthropomorphism increases trust resilience in cognitive agents. *J. Exp. Psychol. Appl.* 22, 331–349
- Waytz, A. et al. (2014) The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. *J. Exp. Soc. Psychol.* 52, 113–117
- Gray, K. et al. (2014) The myth of harmless wrongs in moral cognition: automatic dyadic completion from sin to suffering. *J. Exp. Psychol. Gen.* 143, 1600–1615
- Malle, B.F. et al. (2019) AI in the sky: how people morally evaluate human and machine decisions in a lethal strike dilemma. In *Robots and Well-Being* (Aldinhas Ferreira, I., Silva Sequeira, J., Virk, G.S., Kadar, E.E. and Tokhi, O., eds), Springer

Letter

Does Explaining Social Behavior Require Multiple Memory Systems?

Pieter Van Dessel ^{1,*}
Bertram Gawronski,² and
Jan De Houwer¹

Amodio [1] argues that social cognition research has for many decades relied on imprecise dual-process models that build on questionable assumptions about how people learn and represent information. He presents an alternative framework for explaining social behavior as the product of multiple dissociable memory systems, based on the idea that cognitive neuroscience has revealed evidence for the existence of separate systems underlying distinct forms of learning and memory.

Although we applaud Amodio's attempt to build bridges between social cognition, learning psychology, and neuroscience, we believe that his interactive memory systems model rests on shaky grounds. In our view, the most significant limitation is the idea that behavioral dissociations provide strong evidence for multiple memory systems with functionally distinct learning mechanisms. A major problem with this idea is that behavioral dissociations can arise from processes during the retrieval and use of stored information, which does not

require any assumptions about distinct memory systems or distinct forms of learning. For example, in contrast to Amodio's argument that double dissociations between implicit evaluative bias and implicit stereotypical bias in the prediction of different forms of discriminatory behavior provide evidence for distinct memory systems [2], the observed dissociation may simply indicate that people retrieve and use different kinds of information when faced with different kinds of behavioral decisions (e.g., how close to sit next to a stranger vs. whom to choose as a partner for a trivia task). Such differences in the retrieval and use of stored information do not imply that different types of information (e.g., evaluative vs. stereotypical) are stored in distinct memory systems.

The same concern applies to dissociations involving neural structures. For example, in instrumental learning tasks, Parkinson's disease patients with striatal dysfunction have been found to verbally report the correct reward contingencies without making reward-congruent choices, whereas patients with hippocampal lesions show the reversed impairment [3]. Amodio interprets such findings as evidence for independent representations of conceptual and instrumental knowledge arising from distinct forms of learning [1]. However, such dissociations can also arise from differences in retrieval processes drawing upon a single memory system. In line with this concern, it has been argued that dissociations in the behavior of Parkinson's disease and hippocampal lesion patients reflect differences in the expression of a single type of representation in two tasks that require different ways of retrieving these representations [4]. Theoretical ambiguities like these have led to increased skepticism about the idea that cognitive