Science and Engineering Ethics (2020) 26:2427–2437
https://doi.org/10.1007/s11948-020-00239-9

COMMENTARY

The Human Side of Artificial Intelligence

Matthew A. Butkus[1]

Published online: 7 July 2020

Abstract
Artificial moral agents raise complex ethical questions both in terms of the potential decisions they may make as well as the inputs that create their cognitive architecture. There are multiple differences between human and artificial cognition which create potential barriers for artificial moral agency, at least as understood anthropocentrically and it is unclear that artificial moral agents should emulate human cognition and decision-making. It is conceptually possible for artificial moral agency to emerge that reflects alternative ethical methodologies without creating ontological challenges or existential crises for human moral agents.

Keywords Artificial intelligence · Neural modeling · Popular culture · Ontology · Moral agency · Ethics

Introduction

I want to thank Drs. Farisco, Evers, and Salles for their compelling cautionary statement regarding the ethical evaluation of artificial intelligence. The themes in their work merit additional exploration with both critical and expansive eyes and I will follow their outline regarding essential criteria for artificial intelligence as well as both theoretical and practical issues in intelligence and ethics. Their work invites comparison to findings in both cognitive psychology and neuroscience (and popular culture)—the questions raised about deliberation and moral agency draw immediate parallels to our own processing as the model with which we have the most direct experience. Ultimately, there are ontological questions as to whether we ought to feel challenged by the possibility of artificial moral agents and whether we ought to

Commentary for "Towards establishing criteria for the ethical analysis of artificial intelligence." by Michele Farisco et al.

✉ Matthew A. Butkus
mbutkus@mcneese.edu

[1] Department of Social Sciences, McNeese State University, Lake Charles, LA, USA

welcome a potentially different model of ethical reasoning, considering the cognitive architecture for our own moral reasoning can be (charitably) described as haphazard.

Essential Features of AI

In their introduction, Farisco, Evers, and Salles introduce the controversy in defining artificial intelligence, suggesting the minimal criteria of "perception of the environment through sensors; reasoning/decision-making on data perceived; and actuation through executors." Self-learning machines epitomize these criteria as the apparent endpoint of a continuum of less "intelligent" machines. Artificial intelligence takes many forms from programmed algorithms to autonomous vehicles and robots.

There is immediate appeal to the criteria and obvious parallels to our own cognitive experience of the world around us. Throughout our history, we have fashioned ourselves to be reasonable and mostly rational beings, perceiving the world through our own phenomenological filters, weighing and evaluating what we perceive, drawing inferences and other logical relationships, and acting based on those perceptions. It only stands to reason that we would use a similar understanding in conceptualizing an artificial cognitive agent. There is a *prima facie* truth in this model, but is it sufficient?

What is Reasoning?

There are legitimate questions about what constitutes reasoning (Goel 2007; Stanovich & West 2000)—the input and weighing of evidence is certainly part of it, but we find that there are limitations in "reasoning" human and non-human animals. If we confine "reasoning" to purely rational processes (e.g., absent emotional connection, instincts, etc.) in a Kantian or Cartesian sense, then none of us really qualifies. Our cognitive architecture weaves emotional and preconscious processing throughout our reasoning process (Bargh 1997; Clore & Ketelaar, 1997; Damasio 1995; Evans 2010; Fauconnier & Turner 2002; Franklin et al. 2014; Haidt 2001; Homan 2003; Isen & Diamond 1989; Logan 1989; Prinz 2015; Smith 1997; Turner 2000; Uleman & Bargh 1989; Wyer 1997). These intuitive and emotional responses are not volitional and are much deeper structures in our brains. As such, contemporary arguments about human "reasoning" must necessarily account for a much more complex cognitive model.

Farisco, Evers, and Salles are not exploring human reasoners, and as such, it isn't clear that the architecture of human reasoning and agency would be applicable here. However, in the spirit of their caution, I raise these issues because they also explain some potential concerns about artificial agents, whether programmed or naturally learning. They note that there are fundamental elements of human cognition that are absent in current AI and may prove impossible to translate into artificial experience (e.g., counterfactual reasoning and emotional experience). They rightly note that these challenges may serve as barriers in particular ethical contexts (childcare, healthcare), but there are additional concerns as well. If artificial agents have deficits

in emotional processing, for instance, it is entirely possible that their moral calculus may yield the ethical problems identified early in this discussion by Wallach & Allen (2009)—both top–down and bottom-up moral methodologies are insufficient. We cannot easily optimize for a pure utilitarian calculus any more than we can create a Napoleonic code sufficient for deontological absolute prohibitions or contextualization of *prima facie* principles. As such, "reasoning" in this case will either be incomplete (from our perspective) or yield morally repugnant or incoherent recommendations. I will be returning to these concerns as this article progresses.

Potential Biases in Reasoning

Empirical problems emerge in artificial ethical reasoning regardless of whether we use programmed ethical rules or allow machines to learn on their own. If we trust artificial agents to generate moral rules by their own understanding of interactions between other moral agents, we risk the development of artificial agents representing our worst inclinations and opinions (Angulo 2018). Without constraints on what they learn and how, "self-learning" artificial agents quickly can be manipulated into parroting horrific human sentiments. While it may be possible to program constraints (e.g., watching for the use of particular keywords during interactions), the fluid nature of natural language, the ease with which evil intentions can co-opt benign phenomena (Anti-Defamation League n.d), and a wealth of other concerns make these constraints reactive rather than proactive. Programmers would have to update their agents constantly and without real-time awareness given this fluidity. Simply put, it would seem that self-learning artificial agents would be vulnerable to manipulation and contextually naïve.

On the other hand, if we do not adopt a machine-learning model, we run the risk of a host of induced biases based on the nature of our cultural and personal preferences (Lloyd 2018; Sweeney 2013). Algorithmic bias reflects the programmer who is culturally situated and influenced, increasing the risk that data analysis will yield skewed results. There is already empirical verification of this, whether we consider cases like Amazon's hiring practices (Dastin 2018) or risk assessments of recidivism reflecting racial biases (Angwin et al. 2016). In essence, just as human reasoning reflects myriad backstage emotional, intuitional, and preconscious elements, there are significant risks associated with artificial agents.

Contextually Appropriate Reasoning and Decision-Making

The crux of the concerns raised here is that "reasoning/decision-making based on the data perceived" is likely going to be an insufficient descriptor of artificial intelligence. Rather, it would be preferable to modify this criterion to reflect more contextually appropriate reasoning and decision-making. At a practical level, everyday experience demonstrates the need for contextually appropriate responses—it is entirely possible for someone to engage in a reasoning process (decision-making based on data presented) but yield results that are highly inappropriate. Having spent several years working in behavioral health, there are plenty of people capable

of processing information in ways that do not gel with what we consider "normal" reasoning[1]. Whether we are describing patients with psychiatric conditions or the inferences of young children, the conclusions reached by a reasoning process might not be contextually-appropriate and hence an inappropriate understanding of artificial intelligence. To be fair, contextualization does emerge as a concern in their article but not as clearly or early as might be desired.

The contextualization problem combined with the machine-learning issues suggest that the criterion include some threshold before responding/acting, whether we are discussing actual artificial agent actions (such as those of autonomous vehicles or robots) or those of information processing (such as those of AI agents which simply make recommendations in professional or commercial contexts). Artificial intelligence takes many forms—Farisco, Evers, and Salles rightly note that the issue at hand is not necessarily one of humanity-destroying proportions, but there are clearly deficits when an artificial agent suggests that because one has watched *A Bug's Life* (insects working together) that they would enjoy *The Human Centipede* (body horror). Analyzing my viewing habits is certainly working with data input, but the culmination of that particular inductive reasoning process is clearly insufficiently informed and yields an uncogent conclusion.

Intelligence

Farisco, Evers, and Salles note that there is no real agreement in the definition of intelligence. Paralleling this, attempting to define intelligence is outside both the purview of this paper and the skillset of this author. However, there are some elements of human cognition (and moral cognition) that are better understood and serve as a useful starting point for discussions of artificial intelligence.

Cognitive psychology has made significant inroads in modeling our thought processes. The past several decades of research have suggested that we have processes that are both linear and intuitive, yielding language of "hot and cold" and "System I/II" processing (Gilovich et al. 2002; Kahneman 2011). These processes reflect both our labor-intensive systemic processing as well as our quick intuitive judgments—we employ both routinely when evaluating novel situations. When we have to decide quickly (e.g., when we have to make snap decisions), we employ System I. When we have the luxury of deciding between options (e.g., when we have the ability to spend some time reflecting on our options), we routinely employ System II. Both systems

[1] Needless to say, what constitutes "normal" versus "abnormal" reasoning is another contentious topic. For the sake of brevity, I understand the terms here in light of endpoints of a continuum of behaviors, with "abnormal" including behaviors that by their nature prevent the agent from meeting their intended goals. Just as the authors note that researchers like Turing point to manifestations of intelligent action rather than attempt to define it, I will refer to manifestations of "abnormal" cognitive processing. As an example, I worked with a patient whose spatial perception prevented him from being able to navigate hallways (he would walk into corners and injure himself) or feed himself (he would hold his sandwich a foot away from his mouth and attempt to take bites). He clearly was attempting to interact with his environment, but the results of his reasoning process prevented him from being to attain those goals.

have strengths and weaknesses—our intuitive judgments allow us to make rapid decisions, but they can also be sources of error (e.g., we might be unduly influenced by a particular memory regardless of how appropriate it is for the current situation). Our more labor-intensive decision-making processes don't necessarily fall into the same traps as our cognitive heuristics but don't lend themselves to situations requiring faster responses. Cognitive heuristics are not the only elements of our cognition; we also employ emotional valence (whether memories have positive or negative emotional states attached to them), we are vulnerable to framing (how information is presented) and priming effects (how unrelated information presented first shapes our response to subsequent information), and backstage cognition elements like preconscious associations, cultural context, and socialization. All of these elements influence how we make decisions, most are not volitional, and the results can be difficult to predict. As Wallach and Allen note, "humans are hybrid decision-makers with unique approaches to moral choices, honed over time and altered by their own distinctive experiences" (Wallach and Allen 2009, p. 178). Neither our day-to-day nor our ethical judgments are linear thought processes.

There would seem to be advantages to dual-processing models, allowing for both quick and more methodological decision-making processes. Our particular cognitive architecture, however, developed gradually through socialization rather than through programming (Carter 2014; Dunbar & Schultz 2007; Eisenberger 2013; Frith, 2007; Garrod & Pickering 2004; Hari et al.2015; Hurlemann et al. 2010; Saxe 2006). This sensitized us to emotional cues and an ability to engage in moral imagination and perspective taking—tasks that Farisco, Evers, and Salles note currently elude artificial agents. This raises questions as to the extent we would want similar processes in our artificial agents—would we want them to think like us? Should we model the "non-rational" elements of moral decision-making? Our socialization also yielded a cognitive architecture vulnerable to social shaming and isolation, a predisposition to favor perceived ingroups, and a significant aversion to conflicts within the group. This cognitive architecture becomes vulnerable to peer pressures, groupthink, and a host of other group-specific deficits in decision-making (Asch 1951, 1956; Camerer et al. 2004; Baumeister & Leary 1995; Falk & Bassett 2017; Janis 1982). If we elect not to program the socialized and non-rational elements of cognition, we would need to accept a rational model that is foreign to us—an "Other" whose moral experience is artificial and alien to us and whose rational conclusions may differ radically from our own.

Ethics

Moral agency (human or otherwise) defies easy explanation, and an attempt at fully accounting for it is well outside the purview of the present work. However, if we are concerned about the capacity for agency in artificial agents, it is worth highlighting a few facets of our reasoning that may serve as barriers.

At a superficial level, programming agency would seem to be a straightforward task—optimize the utility of actions performed (utilitarianism) while accepting constraints on actions or outcomes that violate these concerns (deontology). Attempting

to realize this in an artificial form is much more difficult (Wallach and Allen 2009)—what constitutes "utility" for which population and what are the limits of reasonably affected agents (e.g., who matters and how far into the future are we predicting outcomes)? What limits should need to be placed—an absolute prohibition against killing/harming other moral agents or are these acceptable if they are limited? How do we decide those limits? These are *basic* questions that raise complicated issues which may defy coding and are but a sample of many other factors we would normally expect ethical agents to weigh[2].

The practical issues that Farisco, Evers, and Salles note are compelling, especially the emotional barriers faced by artificial moral agents. As alluded to above, emotional processing is critical to human moral agency and it isn't uncommon for a lack of empathy and emotion to draw comparisons to the agent in question being more of a robot instead of a human being. Popular rhetoric aside, there are problems that arise when humans engage in ethical deliberation without emotional valence, compassion, and other agent relative concerns. A purely rational utilitarian calculus quickly yields horrific abuses so long as the majority benefits (e.g., involuntary experimentation for the betterment of mankind, rights abuses justified by raising the overall quality of life or security of the population in question, and so on). Depending on the outcomes we seek to optimize (i.e., what the "good" to be obtained actually is), we can easily see how a purely rational application of utility maximization yields outcomes most agents would find undesirable. If we adhere to a purely rational deontological framework, we can quickly yield results that are implicitly damaging to our relationships (e.g., always telling the unvarnished truth without worrying about compassion or contextual appropriateness). Simply put, purely rational ethical agents could easily yield outcomes that we find undesirable despite the coherence and logic of the reasoning. Efforts to program emotion and emotion-based reasoning raise a host of other concerns including the dilemma about deception noted by Farisco, Evers, and Salles as well as the concerns raised here and elsewhere about biases in the coding.

As the authors note, an apparently unique human attribute that informs this contextualization of ethical methodology is our ability to use inductive and abductive reasoning styles that would not be readily apparent to an artificial agent. This raises a unique dilemma for artificial agents, however. If we recognize that there is a potential deficit in their reasoning, this can be addressed (potentially) by making them think more like human agents. But is this really desirable? Do we want them to think like we do? A significant part of our neurological development can be linked to the process of socialization and "in group" maintenance (Carter 2014; Dunbar & Schultz 2007; Eisenberger 2013; Frith 2007; Garrod & Pickering 2004; Hurlemann

[2] For instance, in practicing ethics in a natural rather than controlled environment, it would seem reasonable to require an ethical agent to be able to identify the central ethical dilemma (rather than ancillary issues), identify which agents are both affected and relevant, what personal or professional obligations might exist, which ethical methodology (or methodologies) are appropriate in approaching the problem, if there are any implicit prohibitions on particular actions (like murder), what is the context of the action considered, what are the agent's intentions, and what are the likely consequences. Each of these presents unique coding challenges or could be led astray in self-learning systems (just like in human agents).

et al. 2010; Saxe 2006) with resultant problems. Our brains evaluate scenarios differently in groups rather than in isolation (Hari et al. 2015; Schilbach et al. 2013), they easily fall victim to groupthink and other fallacies, they are not function specific (e.g., regions of emotional processing are also used in social processing), and so on. There is a danger that making them more like us will expose artificial agents to similar deficits and distractions.

At a more concrete level, Farisco, Evers, and Salles note issues that may arise in the context of health- and childcare. These seem to be largely valid, but questions remain. First, while there may be roles for artificial intelligence to play in both of these fields, it isn't clear *why* we would consider offsetting direct care and human contact in the first place. While some data indicate a shortage of caregivers in the future (Sharkey & Sharkey 2012), it isn't clear that artificial caregivers would be an appropriate solution. Pew surveys find Americans have a mix of worry and willingness about receiving care from artificial agents (Smith & Anderson 2017) while data from Japan and other countries suggest a significantly greater willingness to receive care while simultaneously indicating that they would not feel pressure to form a bond as they would with human care (Broadbent et al. 2011; Jiji 2018). Interpersonal connection is critical, and hallmarks of both fields include empathy and compassion, both of which defy current programming. If we were able to simulate them, it is not clear that patients or children would be able to form bonds with them (Sharkey & Sharkey 2011)[3]. If the artificial caregiver is not perceived as a moral agent capable of forming genuine bonds, it stands to reason that there would be difficulties with trusting them in the course of care[4]. It is likely that artificial agents may be more trusted in environments allowing for information analysis and synthesis rather than the direct care of vulnerable populations and other circumstances as artificial intelligence is already employed in this capacity elsewhere (Whitby 2008).

Farisco, Evers, and Salles present additional theoretical ethical concerns, and two of these deserve additional attention. First, they briefly raise the possibility of machines developing their own capacity for moral/ethical reasoning but suggest it is unlikely. Second, they raise concerns about human agents experiencing an ontological crisis concerning our perceived unique ability to engage in moral reasoning. Both of these suggest an existential crisis emerging from artificial agents becoming 'too much like us'—i.e., bridging the ontological divide between humanity and technology.

First, it isn't clear why the development of unique moral/ethical capacities is unlikely. If we accept this premise and still seek to create ethical artificial agents, we will encounter the difficulties noted above about coding artificial agents. If we reject this premise, it is not difficult to imagine circumstances that would allow artificial agents to develop their own moral or ethical principles and methodologies (albeit

[3] I draw a distinction between willingness to receive care from a robot caregiver from forming a bond with a caregiver.

[4] Interestingly, interactions between certain elderly populations and inanimate dolls suggests that some degree of bonding might be possible but this intervention is controversial at best (Mitchell 2014; Sharkey & Sharkey 2012).

not in a form familiar to us; they would not necessarily generate anthropocentric moral or ethical principles). This suggestion is entirely compatible with observable phenomena of artificial agents—novel language development, machine learning, etc. all suggest that machines would be able to create their own optimization schemas (Lewis et al. 2017). These can be expressed in terms of both rules to follow as well as moral principles (any initial preprogrammed goal that allows for adaptation by the artificial agent or what the artificial agent itself chooses to be optimized). This does, however, yield the possibility of the nightmare scenarios envisioned in popular culture where humanity may simply be another variable to consider rather than an end in itself. Discussions of any particular schemas or frameworks here is beyond the scope of the argument, but there does seem to be some consistent mechanism by which artificial agents could solve their own optimization problems.

The second issue is also compelling—the ontological crisis of the non-unique problem of moral reasoning. As with the originating moral principles problem, it isn't clear that this ontological crisis is guaranteed or would be severe. There are already other parallels in non-human animal agency—we see behaviors similar to our own in other species and do not seem to begrudge them tool usage, complex language, emotional lives, basic economies, and so on. Many of the behaviors that seemed to be "unique" to us are clearly shared. It isn't clear that moral reasoning should be much different. Popular culture has been addressing this phenomenon for decades, creating both sympathetic and terrifying characters (from Data on *Star Trek* to the morally complex and ideologically divided Cylons of *Battlestar Galactica* and "hosts" of *Westworld*). Some of these portrayals explicitly attempt to blur the line between biological and synthetic, creating characters with the same apparent will and volition as human agents and synthetic biology nearly identical to our own. These apparently functionalist portrayals suggest a willingness to consider the issue while recognizing that their resultant moral systems may be at odds with our own. Admittedly, this is not a strong argument—willingness to engage with science fiction does not necessarily translate into a willingness to engage with a concrete challenge. However, science fiction has introduced many elements of technology taken for granted now—this at least raises the possibility that a parallel process may occur when it comes to a synthetic moral agent.

Conclusion

Ultimately, the caution suggested by Farisco, Evers, and Salles is reasonable—it is quite easy to reach both dystopian and idyllic conclusions about the future of artificial intelligence absent serious exploration of both the conceptual and practical issues they raise. Some of the concerns they note seem to be warranted, others require additional justification. The degree to which we are integrating artificial intelligence into our daily lives suggests that some of their ontological concerns warrant some skepticism. If we intentionally avoid some of the known pitfalls in our cognitive architecture, we cannot help but create moral agents that are dissimilar from us. Knowing that they will be different from us from the start could offset some of the existential uncertainty—if anything, the experience of an alien "other"

in moral agency might serve to overcome some of the out-group biases we experience currently towards our fellow human agents.

References

Angulo, I. (2018, March 17). Facebook and YouTube should have learned from Microsoft's racist chatbot. Accessed November 23, 2019, https://www.cnbc.com/2018/03/17/facebook-and-youtube-should-learn-from-microsoft-tay-racist-chatbot.html.

Angwin, J., Jeff, L., Surya, M., and Lauren, K. (2016, May 23). Machine bias. Accessed November 21, 2019, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Anti-Defamation League. (n.d). Okay hand gesture. Accessed November 23, 2019, https://www.adl.org/education/references/hate-symbols/okay-hand-gesture.

Apps, M. A. J., Rushworth, M. F. S., & Chang, S. W. C. (2016). The anterior cingulate gyrus and social cognition: tracking the motivation of others. *Neuron, 90*, 692–707. https://doi.org/10.1016/j.neuron.2016.04.018.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, Leadership and Men; Research in Human Relations* (pp. 177–190). Oxford: Carnegie Press.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied, 70*(9), 1–70.

Bargh, J. A. (1997). The automaticity of everyday life. In *The automaticity of everyday life* (pp. 1–61). Mahwah: Lawrence Erlbaum Associates.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*(3), 497–529.

Broadbent, E., Tamagawa, R., Patience, A., Knock, B., Kerse, N., Day, K., et al. (2011). Attitudes towards health-care robots in a retirement village. *Australasian Journal on Ageing, 31*(2), 115–120. https://doi.org/10.1111/j.1741-6612.2011.00551.x.

Camerer, C. F., Loewenstein, G., & Prelec, D. (2004). Neuroeconomics: why economics needs brains. *The Scandinavian Journal of Economics, 106*(3), 555–579. https://doi.org/10.1111/j.1467-9442.2004.00378.x.

Carter, C. S. (2014). Oxytocin pathways and the evolution of human behavior. *Annual Review of Psychology, 65*, 17–39. https://doi.org/10.1146/annurev-psych-010213-115110.

Clore, G., & Ketelaar, T. (1997). Minding our Emotions: On the Role of Automatic, Unconscious Affect. In R. S. Wyer (Ed.), *The Automaticity of Everyday Life* (pp. 105–120). Mahwah: Lawrence Erlbaum Associates.

Damasio, A. (1995). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Penguin Books.

Dastin, J. (2018, October 9). Amazon scraps secret AI recruiting tool that showed bias against women. Accessed November 21, 2019, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Dunbar, R. I. M., & Schultz, S. (2007). Evolution in the social brain. *Science, 317*, 1344–1347. https://doi.org/10.1126/science.1145463.

Eisenberger, N. I. (2013). Social ties and health: a social neuroscience perspective. *Current Opinions in Neurobiology, 23*(3), 407–413. https://doi.org/10.1016/j.conb.2013.01.006.

Evans, J. S. B. T. (2010). Intuition and reasoning: a dual-process perspective. *Psychological Inquiry, 21*(4), 313–326. https://doi.org/10.1080/1047840X.2010.521057.

Falk, E. B., & Bassett, D. B. (2017). Brain and social networks: fundamental building blocks of human experience. *Trends in Cognitive Sciences, 21*(9), 674–690. https://doi.org/10.1016/j.tics.2017.06.009.

Falk, E., & Scholz, C. (2018). Persuasion, influence, and value: perspectives from communication and social neuroscience. *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-122216.011821.

Fauconnier, G., & Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Franklin, S., Madl, T., D'Mello, S., & Snaider, J. (2014). LIDA: A systems-level architecture for cognition, emotion, and learning. *IEEE Transactions on Autonomous Mental Development, 6*(1), 19–41. https://doi.org/10.1109/TAMD.2013.2277589.

Frith, C. D. (2007). The social brain? *Philosophical Transactions of the Royal Society B, 362*, 671–678. https://doi.org/10.1098/rstb.2006.2003.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences, 8*, 8–11.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and Biases: the Psychology of Intuitive Judgment*. New York: Cambridge University Press.

Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Science, 11*(10), 435–441. https://doi.org/10.1016/j.tics.2007.09.003.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.

Hari, R., Henriksson, L., Malinen, S., & Parkkonen, L. (2015). Centrality of social interaction in human brain function. *Neuron, 88*(1), 181–193. https://doi.org/10.1016/j.neuron.2015.09.022.

Homan, R. W. (2003). Autonomy reconfigured: incorporating the role of the unconscious. *Perspectives in Biology and Medicine, 46*(1), 96–108.

Hurlemann, R., Patin, A., Onur, O. A., Cohen, M. X., Baumgartner, T., Metzler, S., et al. (2010). Oxytocin enhances amygdala-dependent, socially reinforced learning and emotional empathy in humans. *The Journal of Neuroscience, 30*(14), 4999–5007. https://doi.org/10.1523/JNEUROSCI.5538-09.2010.

Isen, A. M., & Diamond, G. A. (1989). Affect and automaticity. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended Thought* (pp. 124–52). New York: Guilford Press.

Janis, I. L. (1982). *Groupthink: A Psychological Study of Policy Decisions and Fiascos*. Boston: Houghton Mifflin Company.

Jiji. (2018, November 15). Over 80% of Japanese positive about robotic nursing care. Accessed November 23, 2019, https://www.japantimes.co.jp/news/2018/11/15/national/80-japanese-positive-robotic-nursing-care/#.XebK-OhKiM8.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Lewis, M., Denis Y., Yann N. D., Devi, P., & Dhruv, B. (2017, June 14). Deal or no deal? Training AI bots to negotiate. Accessed November 23, 2019, https://engineering.fb.com/ml-applications/deal-or-no-deal-training-ai-bots-to-negotiate/.

Lloyd, K. (2018, September 20). Bias amplification in artificial intelligence systems. Accessed November 23, 2019, https://arxiv.org/abs/1809.07842.

Logan, G. D. (1989). Automaticity and cognitive control. In S. J. UIeman & J. A. Bargh (Eds.), *Unintended Thought* (pp. 52–74). New York: Guilford Press.

Mitchell, G. (2014). Use of doll therapy for people with dementia: an overview. *Nursing Older People, 26*(4), 24–26. https://doi.org/10.7748/nop2014.04.26.4.24.e568.

Prinz, J. (2015). Is the moral brain ever dispassionate? In J. Decety & T. Wheatley (Eds.), *The Moral Brain: A Multidisciplinary Perspective* (pp. 51–67). Cambridge, MA: The MIT Press.

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology, 16*, 235–239. https://doi.org/10.1016/j.conb.2006.03.001.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schilcht, T., et al. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences, 36*, 393–462. https://doi.org/10.1017/S0140525X12000660.

Sharkey, A., & Sharkey, N. (2011). Children, the elderly, and interactive robots. *IEEE Robotics and Automation Magazine, 18*(1), 32–38. https://doi.org/10.1109/MRA.2010.940151.

Sharkey, A., & Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology, 14*(1), 27–40. https://doi.org/10.1007/s10676-010-9234-6.

Smith, A. & Monica, A. (2017). 4. Americans' attitudes toward robot caregivers. Accessed November 23, 2019, https://www.pewresearch.org/internet/2017/10/04/americans-attitudes-toward-robot-caregivers/.

Smith, E. R. (1997). Preconscious automaticity in a modular connectionist system. In R. S. Wyer (Ed.), *The Automaticity of Everyday Life* (pp. 187–202). Mahwah: Lawrence Erlbaum Associates.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435.

Sweeney, L. (2013). Discrimination in online Ad delivery. *Queue, 10*(3), 10 (20 pages). https://doi.org/10.1145/2460276.2460278.

Turner, M. (2000). Backstage cognition in reason and choice. In *Elements of Reason: Cognition, Choice, and the Bounds of Rationality* (pp.264–286). New York: Cambridge University Press.
Uleman, J. S., & Bargh, J. A. (1989). *Unintended Thought*. New York: Guilford Press.
Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.
Whitby, B. (2008). Computing machinery and morality. *AI & Society, 22*, 551–563. https://doi.org/10.1007/s00146-007-0100-y.
Wyer, R. S. (Ed.). (1997). *The Automaticity of Everyday Life*. Mahwah: Lawrence Erlbaum Associates.