Anthropomorphism in AI

Anthropomorphism in AI
Arleen Salles (1, 2)
Kathinka Evers (1)
Michele Farisco (1,3)

1: Centre for Research Ethics & Bioethics. Uppsala University, Uppsala, Sweden
Box 564, 751 22 Uppsala, Sweden.
2: 3: Science and Society Unit, Biogem, Biology and Molecular Genetics Institute, Via Camporeale, Ariano Irpino (AV), Italy

Abstract:

AI research is growing rapidly raising various ethical issues related to safety, risks, and other effects widely discussed in the literature. We believe that in order to adequately address those issues and engage in a productive normative discussion it is necessary to examine key concepts and categories. One such category is anthropomorphism.

It is a well-known fact that AI's functionalities and innovations are often anthropomorphized (i.e., described and conceived as characterized by human traits). The general public's anthropomorphic attitudes and some of their ethical consequences (particularly in the context of social robots and their interaction with humans) have been widely discussed in the literature. However, how anthropomorphism permeates AI research itself (i.e. in the very language of computers scientists, designers, and programmers), and what the epistemological and ethical consequences of this might be have received less attention.

In this paper we explore this issue. We first set the methodological/theoretical stage, making a distinction between a normative and a conceptual approach to the issues. Next, after a brief analysis of anthropomorphism and its manifestations in the public, we explore its presence within AI research with a particular focus on brain-inspired AI. Finally, on the basis of our analysis, we identify some potential epistemological and ethical consequences of the use of anthropomorphic language and discourse within the AI research community, thus reinforcing the need of complementing the practical with a conceptual analysis.

Acknowledgements

This research was supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2).

1- Introduction

AI research is growing rapidly raising various ethical issues related to safety, risks, and other effects widely discussed in the literature. We believe that in order to adequately address those issues and engage in a productive normative discussion it is necessary to examine key concepts and categories. One such category is anthropomorphism.

It is a well-known fact that AI's functionalities and innovations are often anthropomorphized (i.e., described and conceived as characterized by human traits). The general public's anthropomorphic attitudes and some of their ethical consequences (particularly in the context of social robots and their interaction with humans) have been widely discussed in the literature. However, how anthropomorphism permeates AI research itself (i.e. in the very language of computers scientists, designers, and programmers), and what the epistemological and ethical consequences of this might be have received less attention.

In this paper we explore this issue. We first set the methodological/theoretical stage, making a distinction between a normative and a conceptual approach to the issues. Next, after a brief analysis of anthropomorphism and its manifestations in the public, we explore its presence within AI research with a particular focus on brain-inspired AI. Finally, on the basis of our analysis, we identify some potential epistemological and ethical consequences of the use of anthropomorphic language and discourse within the AI research community, thus reinforcing the need of complementing the practical with a conceptual analysis.

2- A note on methodology and approach

Addressing the ethical issues raised by AI research and engaging in the normative discussion is a priority. However, a prerequisite for that discussion to be more informed and useful is arguably an analysis of concepts and categories underlying many of the ethical concerns. At present, much emphasis is given to the first, in particular, the mapping of the relevant ethical issues and the development of practical guidelines and recommendations (Tasioulas 2018, Boddington 2017,

Turner 2019). This practically oriented reflection usually refers to and makes use of a few fundamental principles common in the bioethical literature (e.g., human dignity, respect for autonomy, non-maleficence, beneficence, justice) and of additional considerations (such as the importance of human rights, promotion of human well-being, trust, transparency, accountability, and effectiveness) that are jointly taken to provide an adequate framework to achieve ethically sound AI.

However, in the relevant literature and documents there is a general tendency to assume that ethical reflection on AI is fundamentally applied. To illustrate, the Ethics Guidelines for Trustworthy AI developed by the High Level Experts Group on AI set up by the European Commission explicitly define AI ethics as "a sub-field of applied ethics, focusing on the ethical issues raised by the development, deployment and use of AI" (HLEG 2019)(p.11). Although not explicitly, the IEEE Ethically Aligned Design document suggests a similar emphasis on applied analysis of the issues. Thus conceived, ethical analysis of AI tends to revolve around a discussion of practical and normative issues with the goal of setting standards, policies, and guidelines.

A conceptual approach to the ethics of AI complements the practical discussion by focusing on an analysis of foundational notions. Ideally, conceptual analysis should be integral to the process of development of the technology itself and give priority to a reflection on foundational notions and interpretative categories (e.g., the values informing AI development) (Tegmark 2018, Havens 2016) thus effectively shaping AI from the very beginning. This is important because conceptual ambiguities exist and they have implications, both theoretical and practical. Therefore, conceptual clarification is necessary to achieve a balanced analysis of applied issues.

One foundational category in need of reflection is anthropomorphism. A number of recent articles have focused on anthropomorphism in human-robot interaction and have attempted to provide concrete steps for monitoring it (Darling 2012, 2017, Kaminski et al. 2017, Leong and Selinger 2018). However, this is not our main concern here. Instead, we attempt to give attention to anthropomorphism as found also in other AI related contexts and to identify its implications within AI research.

3- Anthropomorphism

Anthropomorphism is generally defined as the attribution of distinctively human-like feelings, mental states, and behavioural characteristics to inanimate objects, animals, and in general to

natural phenomena and supernatural entities (Airenti 2015, Epley, Waytz, and Cacioppo 2007). It is a widespread phenomenon that it not necessarily correlated to particular features of the anthropomorphized object itself (Airenti 2015), nor is it dependent on the anthropomorphized entity's ontological status. It is not uncommon for people to anthropomorphize religious figures, animals, the environment, and technological artefacts (from computational artefacts to robots) even when they lack any evolutionary connection with humans, and even if they are materially different from any living being. Thus, anthropomorphism does not describe existing physical features or behaviors but rather represents a particular human-like interpretation of existing physical features and behaviors that goes beyond what is directly observable (Epley et al. 2008). Epley and colleagues provide a general psychological theory of anthropomorphism intended to help in understanding what drives it and the variability in its manifestations. Rather than conceiving it as an invariant and somewhat extraordinary tendency of the human mind, the authors explain it as an ordinary cognitive process albeit accompanied by two key motivational determinants. Anthropomorphism, they argue, is a process of inductive inference -from readily available and "richly detailed" knowledge about humans in general and the self in particular- about characteristics of a non-human agent that is triggered by two motivational factors (Epley, Waytz, and Cacioppo 2007). One is the need to "to experience competence (Epley et al. 2008)," that is, to interact effectively with the surrounding world (understanding, predicting, controlling, and making sense of the uncertainty that permeates it). The second is the human need and desire to form social bonds with other humans which in the absence of humans can easily extend to forging human-like connection with non-human entities (be they animate or inanimate). In this sense, the authors consider that the motivational mechanisms of anthropomorphism are not unlike those that make humans develop a theory of mind, i.e. attributing mental life to other people (Epley 2018, Epley et al. 2008) even when anthropomorphism goes beyond that by also attributing emotional states, behavioural characteristics, and humanlike form to the anthropomorphized entity (Epley, Waytz, and Cacioppo 2007). These drivers would at least partly explain the variability found in anthropomorphic manifestations, for people are more likely to anthropomorphize when they want to compensate for the lack of social connections or when they want to understand their somewhat unpredictable environment (Waytz, Cacioppo, and Epley 2010).

Recent findings suggest that anthropomorphism may be triggered by attention to contingent social clues and that the extent of its occurrence may be influenced by psychiatric conditions or brain damage. A recent study with amygdala-damaged participants showed that while their ability to anthropomorphize was intact, their inability to perceive and process socially salient information reduced their spontaneous anthropomorphism of non-human stimuli (inanimate objects such as technology) in the absence of explicit social cues. This suggests that the amygdala may play an important role in a person's spontaneous tendency to anthropomorphize some entities where cues are not overtly social.(Waytz et al. 2019).

It is important to bear in mind the distinction between experiencing social emotions in a given situation, for example, "I am now bonding with my cat", and interpretations thereof, such as the belief that "my cat experiences bonding the same as I do". It is possible to experience a shared moment with non-humans without therefore believing that those who share that moment must have the same emotional experience. Sharing may presuppose a minimum of similarity in terms of mutuality (both animals, human and cat, bond), but it does not have to be self-projective. In other words, even if a human and a non-human have a shared bonding moment, their respective experiences of that moment may be quite different. Consequently, experiencing social e.g. emotional proximity with non-humans need not be anthropomorphic. This is important in part to safeguard ascriptions of sentience, intelligence, consciousness etc. to non-humans from the charge of being *ipso facto* anthropomorphic.[1]

Our focus here is on the tendency to anthropomorphize technology, particularly AIs whose inner workings, although created by humans, remain inherently opaque for lay people. Of interest is the fact that although in general there is a tendency to attribute human-like traits and motivations to technological artefacts, AI-anthropomorphism comes in many versions.

3.1 Anthropomorphism in the public

The tendency in popular culture to conceive of AIs as people (both emotionally, cognitively, and morally) is importantly influenced both by fictional narratives (literary science fiction, films and tv shows), and by media coverage of AI and robots (Bartneck 2013). It also reflects limited

[1] There is a debate on this issue that for lack of space we cannot address in this paper. Relevant reflections can be found in (Airenti, Cruciano, and Plebe 2019)

understanding of the state of AI and its capabilities. While it cannot be said that this anthropomorphic tendency is intended by the scientific community, it is to a great extent the product of flawed science communication that often results in false expectations about what the technology is and what it can do, and it can trigger overblown fears and unjustified expectations.

The above is different from the emotional type of anthropomorphic thinking evident in some specific group of users, for example, seniors who might have daily interaction with companion and therapeutic (social) robots. If we follow the account of Epley and colleagues, we can say that such anthropomorphism might be driven by the need for social connection. Indeed, these AIs are intended to address the emotional needs of specific users (Wortham, Theodorou, and Bryson 2017, Darling 2017), and therefore they are specifically designed to create the illusion of mutual caring (Darling 2017).

A more intellectualized type of anthropomorphism is found in users of computational programs or virtual assistants within software programs which are designed with anthropomorphic features in the hopes that this will facilitate understanding of the relevant technology, promote its acceptance, increase its effectiveness (Darling 2017, 2012, Zlotowski et al. 2015), and users' apparent competence when interacting with them (Epley, Waytz, and Cacioppo 2007). What these emotional and intellectualized manifestations of anthropomorphic thinking have in common is that they are specifically intended by AI designers: insofar as anthropomorphism is taken to be instrumentally valuable, its underlying mechanisms are explored so that they can be appropriately triggered by interactive AIs in different ways in the relevant contexts.

There are a number of widely discussed ethical consequences of "anthropomorphism by design" that we can only mention. For example, some are concerned about the possibility of mental manipulation and the extent to which purposely attributing anthropomorphic features to AI may make users more vulnerable to be steered in particular directions when making decisions (Hartzog 2015) (p. 802). Others argue that designing anthropomorphic features in AIs will promote a "misplaced" and time-costly emotional involvement in users (Bryson 2010). An additional concern is that socialization with entities that are not truly social, even if designed to offer new tools for enriching user's emotional life and even if beneficial in some specific cases (for example, when enhancing the well-being of those who would otherwise be social outcasts) is not truly meaningful and it can only limitedly replace the richness of human interactions (Bryson 2010) (Sparrow and Sparrow 2006). Of course, these concerns raise important questions

that deserve a separate discussion, such as, for example, whether human interactions are more valuable than others and how to understand human well-being.

Another common objection to anthropomorphism by design is that it is deceptive insofar as it would appear that in order to meet specific, e.g. emotional or social needs, AI must actually "fool" users by making them believe that they have the capacity to engage emotionally. But, is this deceptive? And if so, is this type of deception ethically objectionable or can it be justified? (Coeckelbergh 2012).

Finally, some raise privacy-related concerns. Indeed, human users may be disclosing information to AIs as if they were friends when in reality they are disclosing such information to corporations or remote robot operators (Kaminski et al. 2017) (Calo 2012). Aware of the risk of "machine masquerading" (Miller, Wolf, and Grodzinsky 2015) and its potential implications, some thinkers propose a distinction between "honest" and "dishonest" anthropomorphism in order to assess anthropomorphic manifestations (Kaminski et al. 2017). (Leong and Selinger 2018). [2]

3.2 **Anthropomorphism in AI developers and researchers**

Anthropomorphic language at times appears intrinsic to the field of AI research itself. Indeed, from Turing's descriptions of his machines (Proudfoot 1999, 2011) to recent accounts of AlphaZero's intellectual feats (Strogatz 2018) it is not uncommon to find researchers using terms typically used to describe human skills and capacities when referring to AIs and focusing on alleged similarities between humans and machines. Thus, it is not surprising that already in 1976, McDermott famously complained about AI's researchers' and programmers' use of 'wishful mnemonics' (e.g., terms like 'understand' or 'learn' referred to AI) which he considered to be misleading both for researchers and for the public (McDermott 1976). If we take the above-mentioned account by Epley and colleagues we could say that such anthropomorphism is due to the need to experience competence, i.e. to understand and control AI. However, other possible explanations. One is that anthropomorphism in AI is another illustration of how science has shifted from the eliminativism

[2] A different type of anthropomorphic tendency is suggested in some documents such as the recently published Ethics Guidelines for Trustworthy AI by the High-Level Expert Group on Artificial Intelligence (HLEG). The Guidelines call for trustworthy AI and, in the process, apply terms typically used to describe human features and capacities (trust, trustworthiness) to artificial entities and systems without addressing whether this is appropriate and if so why. We can\t address this here but the possibility of anthropomorphism in soft law is another interesting area to explore.

and "psychophobia" of the late 19th century and the beginning of the 20th century (Evers 2009) to a veritable inflation of anthropocentric mental terms that are applied even to non-living, artificial entities. Another possible interpretation of this inflation is that it is a sign of an intrinsic epistemic limitation/bias of AI researchers. Either way: how does this take place within AI research, and how do we account for the tendency to humanize AI in those who one would expect to be knowledgeable about AI and the extent to which it is not human-like?

We can take Stuart Russell's and Peter Norvig's seminal book on AI as a starting point. They suggest that, historically, two categories of AI definitions have been prevalent: human-centric and rationalist. Human-centric understandings rely on AI's comparison with humans, i.e. the extent to which it can think or act like humans do. Rationalist understandings conceive of AI in terms of an essential rationality or goal-directed behavior, i.e. thinking or acting rationally (Russell and Norvig 2010). As expected, human-centric approaches are particularly vulnerable to anthropomorphic interpretations: within this approach, AI is described as the effort to make computers think and to give them a mind (Haugeland 1985), or as the automation of activities typically associated with humans, such as decision-making, problem-solving, and learning (Bellman 1978). AI is thus often described as systems that act like humans, that is machines performing functions requiring intelligence if performed by humans (Kurzweil 1990), and AI development is understood as the attempt to make computers do things that no humans do better (Rich and Knight 1991).

Anthropomorphic interpretation within the AI research community can take different shapes. Very broadly, it can be an intentional attribution of typical human traits (e.g., cognitive attributes, intentionality, free will, emotions, etc.) to AI devices, it can be a more subtle attribution of expressive behaviors (e.g., smiling, looking for information, etc.) unwittingly anthropomorphizing AIs (Proudfoot 2011) or it can be thinking that AI follows a human-like way of working, and that its operations are consequently completely predictable and understandable. In the case of the latter, the underlying (questionable) assumption seems to be that there is mental similarity between humans and AI. This is the type of manifestation often found within brain inspired AI research.

3.2.1 Anthropomorphism in brain inspired AI

Anthropomorphic assumptions seem to underlie the descriptions of the mutual relationship between neuroscience and AI, claimed to be critical for advancing both fields. One initial goal of

creating correspondences between the functionalities of AI and the human brain was to promote better understanding of the brain, the self, and the behavior of biological organisms (Prescott 2015, Prescott and Camilleri 2018). If we add to this the fact that at the beginning of AI, in the early 50s, the only known systems carrying out complex computations were biological nervous systems, AI researchers' focus on the brain as a source for inspiration in the nascent field of AI is understandable (Ullman 2019b). It is true that neuroscience has benefitted from AI -both as a model for developing ideas about the working brain and as a tool for processing data sets- and the field of AI has also benefitted from successfully emulating brain activities. However, this mutual reinforcement stands on a background of anthropomorphic interpretation of AI (i.e. that AI uses a functional structure that is sufficiently similar to the human brain to give us new insights in both domains).

An anthropomorphic framework is not necessary but often appears to underlie the claim that AI, particularly Deep Neuronal Network (DNN), is key to gaining a better understanding of how the human brain works; and in how enthusiastically the achievements of AI, especially of DNN, have been received.

DNN architecture represents one of the most advanced and promising fields within AI research. It is implemented in the majority of AI-related existing applications, including translation services for Google, facial recognition software for Facebook, and virtual assistants like Apple's Siri (Watson 2019b). The widely hailed AlphaZero victory against the human Go world champion was the result of the application of DNN and reinforced learning. This success had a huge impact on people's imagination, contributing to increase the enthusiasm around AI uses to emulate and/or enhance human abilities (Silver et al. 2017, Silver et al. 2018).

And yet, caution is key. While actual artificial networks include many characteristics of neural computation, such as nonlinear transduction, divisive normalization, and maximum-base pooling of inputs, and they replicate the hierarchical organization of mammalian cortical systems (Hassabis et al. 2017), there are significant differences in structure. In a recent article, Ullman notes that almost everything we know about neurons (structure, types, interconnectivity) has not been incorporated in deep network (Ullman 2019a). In particular, while the biological neuronal architecture is characterized by a heterogeneity of morphologies and functional connections, the actual DNN uses a limited set of highly simplified homogeneous artificial neurons. However, while Ullman provides a balanced analysis of the technology and calls for avoiding

anthropomorphic interpretations of AI, his analysis at times suggests a subtle form of anthropomorphism, if not in the conceptual framework, at least in the language used. For instance, he wonders whether some aspects of the brain overlooked in actual AI might be key to reach Artificial General Intelligence (AGI), seemingly taking for granted (like the founders of AI) that the human brain is the (privileged) source of inspiration for AI, both as a model to emulate and a goal to achieve. Moreover, he refers to learning as the key problem of DNN and technically defines it as the adjustment of synapses to produce the desired outputs to their inputs. While he has a technical, non-biological definition of synapses (i.e., numbers in a matrix vs electrochemical connections among brain cells), the mere use of the term "synapse" might suggest an interpretation of AI as an emulation of biological nervous systems.

The problem with anthropomorphic language when discussing DNNs is that it risks masking important limitations intrinsic to DNN which make it fundamentally different from human intelligence (Marcus 2018). In addition to the issue of the lack of consciousness of DNN, which arguably is not just a matter of further development, but possibly of lacking the relevant architecture (Pennartz, Farisco, and Evers 2019), there are significant differences between DNN and human intelligence. It is on the basis of such differences that it has been argued that DNNs can be described as brittle, inefficient, and myopic compared to human brain (Watson 2019b). Brittle because what are known as "general adversarial networks" (GANs) - special DNNs developed to fool other DNNs- show that DNNs can be easily fooled through perturbation (Goodfellow, Shlens, and Szegedy 2015). This entails minimally altering the inputs (e.g., the pixels in an image), which results in outputs by the DNN that are completely wrong (e.g., misclassification of the image), showing that DNN lacks some crucial components of the human brain for perceiving the real world. Inefficient because in contrast to the human brain, current DNNs need a huge amount of training data to work (Watson 2019b, Marcus 2018). One of the main problems faced by AI researchers is how to make AI learning unsupervised, i.e. relatively independent from training data (Savage 2019). This current limitation of AI might be related also to the fact that, while the human brain relies on genetic, 'intrinsic' knowledge, DNNs lack it (Ullman 2019b). Finally, DNNs are myopic because while refining their ability to discriminate between single objects they often fails to grasp their mutual relationship (Watson 2019a).[3]

[3] Additional differences between DNN and human intelligence might be considered. For instance, the ability of the human brain to intuitively understand the world, or to efficiently learn about new concepts from a few examples, or

In short, even though DNNs achieved impressive results in complex scene classification, as well as semantic labeling and verbal description of scenes (LeCun, Bengio, and Hinton 2015), it seems reasonable to conclude that because they lack the crucial cognitive components of the human brain that enable it to make counterintuitive inferences and commonsense decisions the anthropomorphic hype around neural network algorithms and deep learning is overblown. DNNs do not recreate human intelligence: they introduce a new mode of inference that is better than ours in some respects and worse in others (Watson 2019b).

4- Can Anthropomorphism in AI research be justified? What are its implications?

At the conceptual level, the tendency to project typical human traits onto AI has been defended as a legitimate attempt to overcome the arrogance of humanistic thinking which considers humans unique and separate from other beings and entities when in fact almost every human trait can now be replicated and emulated artificially, i.e. programmed in artifacts. Some interpreters argue that we tend to deny such possibility because otherwise we would have to admit that we might also be the result of a kind of programming (e.g., genetic instructions)(Hammond 2018). On this view, the denial that AI resembles human traits implies a humanistic understanding of the human as unique and special, that is, the type of anthropocentric view widely criticized for philosophical and scientific reasons (Braidotti 2013, Dennett 1995).

However, this argument appears to mistakenly conflate ontological and moral considerations. Questioning the existence of human-like attributes in AI does not entail defending moral human exceptionalism but rather recognizing that from an ontological perspective humans and AI are different. Such questioning does not entail embracing a humanistic thinking, denying that other 'kinds' of intelligence are possible, or thinking that human-level intelligence is the best standard for intelligence, and it is compatible with the view of some researchers who argue that we must start looking for what is called "mindless intelligence", i.e. an intelligence without symbols (Pollack 2006). At the conceptual level, projecting human traits onto AI could result in positioning

to generalize or transfer generalized knowledge from one context to a new one, or the model-based learning characterizing the human brain are currently not implementable artificially (Hassabis et al. 2017).

humans as the model or paradigm and this entails going back to a form of human-centric ontology closed to potentially different forms of intelligence.

Furthermore, the anthropomorphism present in the way many AI researchers conceptualize and talk about AI has at least two significant epistemological implications one for the public and a different one for researchers themselves.

In the general public it inadvertently promotes misleading interpretations of and beliefs about what AI is and what its capacities are. As noted before, this represents a significant failure in scientific communication and engagement, and one that is not ethically minor. Rather than checking and managing lay persons anthropomorphic tendency it tends to support it. But to the extent that the tendency to anthropomorphize shapes how people behave towards the anthropomorphized entity, such anthropomorphism has ethical consequences. First, perceiving AIs as humanlike entails considering them as moral agents, and their actions as the result of an autonomous decision-making process (Waytz, Cacioppo, and Epley 2010), thereby having a normative impact that they should not have. Second, anthropomorphizing AIs may be the source both of overblown fears of AI (that they will make humans obsolete, for example) and of uncritical optimism (regarding the extent to which AIs could actually behave like humans and perform difficult tasks better than humans). Finally, such anthropomorphism might create ethical confusion by blurring moral and ontological boundaries.

Furthermore, anthropomorphic (implicit or explicit) interpretations of AI might also have epistemological impact on the AI research community itself, insofar as the search for biological and psychological realism (i.e., similarity with biological intelligence) might lead to underestimating the possibility of new theoretical and operational paradigms and frameworks thus ultimately limiting the development of AI.

Conclusion

Given the apparent deep-rootedness of anthropomorphism and its potential implications, it would be both misleading and risky not to make efforts to raise awareness of and limit its manifestations, especially within the scientific community. The risk to anthropomorphize appears particularly evident in brain-inspired AI research considered relevant for getting insights about the underpinnings of intelligence in human and other animals. While important in formalizing

concepts used by psychology and neuroscience in quantitative language and in explaining their role in intelligent behavior (Hassabis et al. 2017) the relevance of AI to understanding the human brain is limited by the fact that AI and the brain are not isomorphic in structure.

The above does not deny the importance or the potentiality of mutually beneficial collaboration between AI and neuroscience. AI has benefitted from and it is expected to continue to benefit from neuroscience. Yet, as Hassabis and colleagues point out, the adherence to biological plausibility of AI should not be slavishly enforced: for AI developers, biological plausibility is a guide rather than a strict requirement (Hassabis et al. 2017). The same is true, we think, when psychological and mental similarity between natural (e.g., human) and artificial intelligences are overemphasized. Understanding AI through the lens of human mental features risks reducing it to a sort of replica of the human mind and leads to a flawed and ultimately limited ethical analysis of the issues AI raises.

Acknowledgements

We would like to thank our colleagues in the Centre for Research Ethics and Bioethics at Uppsala University for their comments to a previous draft of the paper. This research was supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2).

References

Airenti, G. 2015. "The Cognitive Basis of Anthropomorphism: From Relatedness to Empathy." *Int J of Soc Robotics* 7:117-127.

Airenti, G., M. Cruciano, and A. Plebe. 2019. *The Cognitive Underpinnings of Anthropomorphism*. Lausanne: Frontiers Media.

Bartneck, C. 2013. "Robots in the theatre and the media." *Design & semantics of form & movement*:64-70.

Boddington, Paula. 2017. *Towards a code of ethics for artificial intelligence*, *Artificial Intelligence: foundations, theory, and algorithms,*. Cham, Switzerland: Springer.

Braidotti, Rosi. 2013. *The posthuman*. Cambridge: Polity.

Bryson, J. 2010. " Robots should be slaves." In *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issue*, edited by Yorick Wilks, 63-74. John Benjamin Publishing Company.

Calo, M.R. 2012. "Robots and Privacy." In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by P. Lin and G. Bekey. Cambridge, MA: MIT Press.

Coeckelbergh, M. 2012. "Are Emotional Robots Deceptive?" *IEEE Transactions on Affective Computing* 3 (4):388-393.

Darling, K. 2012. "Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects." We Robot Conference 2012, Miami; FL.

Darling, K. 2017. "“Who is Johnny?” Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy." In *Robot Ethics 2.0: From Autonomous Cars to Artificial intelligence*, edited by P. Lin, K. Abney and R. Jenkins. Oxford Scholarship Online.

Dennett, D. C. 1995. *Darwin's dangerous idea : evolution and the meanings of life*, *Penguin science series*. London: Allen Lane - Penguin Press.

Epley, N. 2018. "A Mind like Mine: The Exceptionally Ordinary underpinnings of Anthrpomorphism." *Journal of the Association for Consumer Research,* 3 (4):591 - 598.

Epley, N., A. Waytz, S. Akalis, and J. T. Cacioppo. 2008. "When we need a human: Motivational determinants of anthropomorphism." *Social Cognition* 26 (2):143-155.

Epley, N., A. Waytz, and J.T. Cacioppo. 2007. "On Seeing Human: A Three-Factor Theory of Anthropomorphism." *Psychological Review* 114 (4):864-886.

Goodfellow, I. J., J. Shlens, and C. Szegedy. 2015.

Hammond, K. 2018. *A New Philosophy on Artificial Intelligence*.

Hartzog, W. 2015. "Unfair and Deceptive Robots." *Maryland Law Review* 785:74.

Hassabis, D., D. Kumaran, C. Summerfield, and M. Botvinick. 2017. "Neuroscience-Inspired Artificial Intelligence." *Neuron* 95 (2):245-258. doi: 10.1016/j.neuron.2017.06.011.

Havens, John C. 2016. *Heartificial intelligence : embracing our humanity to maximize machines*. New York: Jeremy P. Tarcher/Penguin, an imprint of Penguin.

HLEG. 2019. Ethics Guidelines for Trustworthy AI. Brussels: European Commission.

Kaminski, M., M. Rueben, C. Grimm, and W.D. Smart. 2017. "Averting Robot Eyes." *Maryland Law Review* 76.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep learning." *Nature* 521 (7553):436-44. doi: 10.1038/nature14539.
Leong, B., and E. Selinger. 2018. "Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism." Conference on Fairness, Accountability, and Transparency, Atlanta, GA.
Marcus, G. 2018. Deep Learning: A Critical Appraisal.
McDermott, D. 1976. "Artificial Intelligence Meets Natural Stupidity." *SIGART Newsletter* 57:4.
Miller, K., M. Wolf, and F. Grodzinsky. 2015. "Behind the Mask: Machine Morality." *Journal of Experimental and Theoretical Artificial Intelligence* 27 (1).
Pennartz, C., M. Farisco, and K. Evers. 2019. "Indicators and Criteria of Consciousness in Animals and Intelligent Machines: An Inside-Out Approach." *Front Syst Neurosci* 13:25. doi: 10.3389/fnsys.2019.00025.
Pollack, J.B. 2006. "Mindless Intelligence." *IEEE Intelligent Systems* 21 (3):50-56.
Prescott, T. 2015. "Me in the machine." *New Scientist* 225 (3013):36-39.
Prescott, T., and D. Camilleri. 2018. "The Synthetic Psychology of the Self." In *Cognitive Architectures*, edited by MI. Aldinhas Ferreira, J. Silva Sequeira and R. Ventura. Cham, Switzerland: Springer.
Proudfoot, D. 1999. "How Human Can They Get?" *Science* 284 (5415).
Proudfoot, D. 2011. "Anthropomorphism and AI: Turing's much misunderstood imitation game." *Artificial Intelligence* 175:950-957.
Russell, Stuart, and Peter Norvig. 2010. *Artificial Intelligence: International Version: A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall.
Savage, N. 2019. "Marriage of mind and machine." *Nature* 571:S15-S17.
Silver, D., T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. 2018. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." *Science* 362 (6419):1140-1144. doi: 10.1126/science.aar6404.
Silver, D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis. 2017. "Mastering the game of Go without human knowledge." *Nature* 550 (7676):354-359. doi: 10.1038/nature24270.
Sparrow, R., and L. Sparrow. 2006. "In the hands of machines? The future of aged care." *Mind Mach* 16:141-161.
Strogatz, S. 2018. "One Giant Step for a Chess-Playing Machine." *New York Times*, December 26th, Science.
Tasioulas, J. 2018. "First Steps Towards an Ethics of Robots and Artificial Intelligence." *SSRN*.
Tegmark, M. 2018. *Life 3.0 Being Human in the Age of Artificial Intelligence*. New York, NY: Alfred A. Knopf.
Turner, J. 2019. *Robot Rules. Regulating Artificial Intelligence*. London: Palgrave Macmillan.
Ullman, S. 2019a. "Using neuroescience to develop artificial intelligence." *Science* 363 (6428):692-693.
Ullman, S. 2019b. "Using neuroscience to develop artificial intelligence." *Science* 363 (6428):692-693. doi: 10.1126/science.aau6595.

Watson, D. 2019a. "The Rhetoric and Reality fo Anthropomorphism in Artificial Intelligence." *Minds and Machines* 29:417-440.

Watson, D. 2019b. "The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence." *Minds & Machines* 29:417-440.

Waytz, A., J. Cacioppo, and N. Epley. 2010. "Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism." *Perspect Psychol Sci* 5 (3):219-32. doi: 10.1177/1745691610369336.

Waytz, A., J. T. Cacioppo, R. Hurlemann, F. Castelli, R. Adolphs, and L. K. Paul. 2019. "Anthropomorphizing without Social Cues Requires the Basolateral Amygdala." *J Cogn Neurosci* 31 (4):482-496. doi: 10.1162/jocn_a_01365.

Wortham, R. H., A. Theodorou, and J.J. Bryson. 2017. "Robot Transparency, Trust and Utility." *Connection Science*.

Zlotowski, J., D. Proudfoot, K. Yogeeswaran, and C. Bartneck. 2015. "Anthropomorphism: Opportunities and Challenges in Human-Robot Interaction." *iInt J of Soc Robotics* 7:347-360.