Science and Engineering Ethics (2020) 26:2371–2379
https://doi.org/10.1007/s11948-020-00235-z

INTRODUCTION

Editorial: Shaping Ethical Futures in Brain-Based and Artificial Intelligence Research

Elisabeth Hildt[1] · **Kelly Laas**[1] · **Monika Sziron**[1]

Published online: 4 August 2020

In discussions of the ethical and social implications of artificial intelligence (AI), a good starting point is: How is AI and AI-related technology similar to and different from related contexts and technology? It seems helpful not to attempt to start the debate on ethical and social implications of AI from scratch, but to refer and relate to already existing and sometimes longstanding debates and research. Two fields seem to be particularly useful here: The discussion of philosophical, ethical and social aspects of technology on the one hand, and the context of brain-related sciences, philosophy of mind, bioethics and neuroethics on the other hand.

With this topical collection, we focus on the second approach, i.e. we think about ethical implications of AI and neurotechnology by thinking about and drawing upon similarities and differences between brains and computer technology, neuroscience, behavioral, and cognitive science and computer science, starting from concepts and conceptions used to describe characteristics of humans and non-human animals and reflecting on how these can be used in or transferred to AI.

Brain-Based and Artificial Intelligence

Technology plays an increasingly important role in all of our lives. Video conferencing allows us to work and socialize from home, databases and algorithms allow us to synthesize huge quantities of data, and artificial intelligence helps us to make informed decisions in settings from the doctor's office to the loan office. Recent estimates predict that investment in artificial intelligence will be around $15 trillion by 2030, and investment in information and communication technologies (ICT) might even top this (Holmes 2019). Since Alan Turing's paper, "Computing Machinery and Intelligence" discussed how to build intelligent machines and how to test their intelligence, artificial intelligence has been a growing field

✉ Elisabeth Hildt
ehildt@iit.edu

[1] Center for the Study of Ethics in the Professions, Illinois Institute of Technology, 10 W. 35th Street, Chicago, IL 60616, USA

(Turing 1950). The 1970s saw an expansion of AI into different research fields such as machine learning, robotics, intelligent control and pattern recognition, and today's discoveries and advancements continue to lead to new potential applications (Pan 2016). These include AI neural networks developed by Google, IBM, Microsoft, Facebook and Apple to name a few, the expanding use of facial recognition software in criminal justice, education, and advertising, and autonomous systems that are programmed to make ethical decisions in a variety of different settings.

While AI is affecting nearly every sector of industry, every discipline in academia, and every part of our daily lives, the shared histories and futures of AI and neuroscience is of primary interest when we consider the phrase 'brain-based and artificial intelligence.' Using neuroscience (the study of brain-based intelligence) to develop and inspire AI (artificial intelligence) is not a novel notion, but one that has been increasingly adopted over the years. We see this in current models and formulations of deep learning and deep networks where the structures within biological brains, layered neurons, serve as role models for AI. It has become apparent that while the trajectory of influence in 'brain-based and artificial intelligence' may be assumed to be that brain-based intelligence inspires artificial intelligence, the opposite may also be true. There are also debates as to whether biological brains are still good models for artificial intelligence (Brooks et al. 2012).

Neuroscience has played a considerable role in inspiring and guiding AI development (Hassabis et al. 2017; Ullman 2019): In particular deep networks and reinforcement learning approaches draw from and rely heavily on direct analogies to brains. As Hassabis et al. (2017) review, recent AI work on attention, episodic memory, working memory and continual memory has been inspired by neuroscience. Brain-inspired models are considerably less complex than brains, however. Brains consist of a broad variety of different types of neurons, connected through excitatory and inhibitory synapses, that build complex neural circuits and brain networks. There is clearly less heterogeneity and less complex circuits in AI models which are built of relatively simple and homogenous artificial neurons. It is unclear whether current models are adequate to achieve complex human-like cognitive abilities, or whether more refined brain-inspired or totally different approaches will be needed. Hassabis et al. (2017) consider neuroscience particularly relevant for future AI research and development in areas like intuitive understanding of the physical world, efficient learning, transfer learning, imagination and planning, and "virtual brain analytics."

While the focus is often on the role neuroscience has on shaping AI, there are also manifold ways in which AI research has positively influenced neuroscience research (Hassabis et al. 2017). Concepts developed through work with algorithms, neural networks, and reinforcement learning may provide new approaches that help better understand brain-based intelligence, while artificial neural networks may serve as simulations that provide insights which help to better understand brain processes. Simulation neuroscience, a new research field that aims to build a digital copy of the brain, relies heavily on these interdisciplinary connections and fosters collaborations between researchers from neuroscience and computer science (Markram 2006; Fan and Markram 2019).

In order for transfers between neuroscience and AI and vice versa to be effective, a clear understanding of key terms and concepts used by both fields is required, as well as a high awareness of similarities and differences in the concepts used.

Some Thoughts on Terminology

Within the phrase 'brain-based and artificial intelligence' and the conversations it inspires is a cyclone of debate in regard to terminology. This is not surprising as within both fields, artificial intelligence and neuroscience, terminology is not always conclusive. Name the topic within these fields and professionals, philosophers, and laypersons will disagree as to what the 'right' term to use may be in any given case, and what its 'correct' meaning is. This is bound to be a persistent condition of brain-based and artificial intelligence, but it should not be one that discourages innovations in development, intimidates innovations in theory, or suffocates new ideas down into a maze of definition stalking. It will always be that new terms are being conceived and old terms are challenged. With this being said, having moderately stable terms and definitions is crucial for communication within and beyond each field. Whether one is speaking of weak AI, strong AI, artificial general intelligence, a superintelligence, machine learning, deep learning, deep networks etc., one should explain their understanding and use of the term. In this topical collection several variations in terms are used and as new work emerges there will be more variations in the future.

As this collection compiles ideas from computer scientists, neuroscientists, philosophers, and social scientists to name a few, terms are bound to contest one another. Examples include discussions about terms such as "deep learning" and "knowledge" in David Gunkel's commentary (2020) of Qin Zhu and colleagues' article, "Blame-Laden Moral Rebukes and the Morally Competent Robot: A Confucian Ethical Perspective" (2020).

It is important to broaden the horizon of terminology in brain-based and artificial intelligence, to accept the flexibility of terms in general, to question "normal" uses, and acknowledge that it may be necessary to modify concepts as fields associate with one another. A philosopher's definition of 'intelligence' is likely different from a neuroscientist's definition of 'intelligence.' Accepting the flexibility of terms may be one of the most arduous tasks for anyone working in these fields and as many have noted, should be done cautiously. As Jared Moore reflects, "Vague terms are the wagons of a modern gold rush into the promised riches of a mythic AI frontier" (Moore 2019, p. 2). Language is powerful and unreflected descriptions can lead to misunderstandings about a machine's true capabilities. Discussing issues of vocabulary and questions of what terms are best suited is essential for furthering ethical developments in every field involved in brain-based and artificial intelligence.

At the beginning of his 1950 article "Computing Machinery and Intelligence," Alan Turing raised the question of whether machines can "think." Rather than attempting to answer this question by providing definitions of the central terms "machine" and "think" and acceding to normal uses of the terms, he introduced and further developed what would become known as the Turing test,

> I propose to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words. (Turing 1950, p. 433)

Turing pointed to the problem that, with regard to questions concerning machines and their abilities, the answers we come to depend strongly on the terms we use and the meanings given to these terms. Inferring that these questions will not be solved by referring to the normal uses or definitions of words, but by a more empirical-based approach that considers what the machine actually does.

Particularly with human-AI interaction, strategies have to be further developed to think and talk about humans interacting with AI and robots. What do terms like decision-making, intelligence, or autonomy mean when applied to human beings and AI? We must take care in which conceptions we use for AI, so that meaningful descriptions can be given of AI capabilities and the roles of AI. It could prove to be very misleading to use the same term—for example intelligence—for humans and AI, by either not being aware of the problem or by tacitly assuming that the terms have different meanings in the two different contexts. Instead, as Totschnig (2020) does with the term "autonomy" in his article, "Fully Autonomous AI," it will be necessary to question the uncritical application and transfer of terms that have traditionally been used to describe human (and animal) characteristics, capabilities, and traits to AI and seek adequate ways to describe AI capabilities instead.

In Farisco et al. (2020), the authors propose a methodological model for a comprehensive ethical analysis of instances when AI is used to replace human actors in different contexts such as healthcare, education, and job market. Without a clear understanding of critical concepts such as the similarities and differences between natural and artificial intelligence, they argue, we can not clearly identify the relevant ethical issues related to the development and expanded use of AI. Indeed, in attempting to avoid the many pitfalls and biases present in human reasoning might end with us producing artificial agents that have an intelligence very different from our own (Butkus 2020).

In "Dignity and Dissent in Humans and Non-humans," Matthias (2020) asks whether animals and AI systems can claim dignity if, as Kant defines it, dignity stems from the moral autonomy of the individual. He argues that the answer to this question depends on what constitutes 'human dignity' and 'autonomy,' and what requirements we ask systems to meet for them to be seen as morally autonomous.

Another promising pathway is to develop conceptions that are applicable to both humans and AI without running into major contradictions. This will be very

challenging, especially as current conceptions like free will, rationality, or consciousness, are shaped by the first-person human perspective which is not accessible when describing computational abilities in AI.

Thinking about relevant conceptions is particularly important in neurotechnology, with direct connections and interactions between technology and the brain. In their article, "Correcting the Brain? The Convergence of Neuroscience, Neurotechnology, Psychiatry and Artificial Intelligence" the authors Stephen Rainey and Yasemin J. Erden explore key narrative differences between brain-based and artificial intelligence, and what this means in the use of neurotechnologies and AI in psychiatric practice and treatments (Rainey and Erden 2020). In the subsequent commentary, Jotterand and Bosco (2020) expand the discussion by contemplating how AI technology might *de*-humanize or *re*-humanize medicine.

Overall, it is particularly important to avoid a clash of differently shaped conceptions and understandings. The future of brain-based and artificial intelligence research will inevitably encounter discrepancies in the use of terms and concepts, but this should not hinder the ethical development of research in this field.

Societal Ethics Landscape

Efforts to shape ethical futures in artificial intelligence are not in short supply. Over 80 AI ethics guidelines have been published since the end of 2019 (Schiff et al. 2020), with the majority emanating from professional societies (e.g. IEEE, ACM), governments (e.g. European Union, China), non-profit organizations (e.g. AI Now, Future of Life) and corporations (e.g. Google, IBM). These normative documents focus not only on the potential benefits AI may have for different populations, but also try to provide sets of principles or frameworks for minimizing risks. The principles and ethical issues brought to the forefront by these guidelines vary according to the position, goals, and audience of the authoring institutions, but there is some convergence around a few principles. A 2019 survey of 84 AI guidelines found that the top five ethical principles that appeared in these documents included those of (1) transparency/explainability, (2) justice and fairness, (3) non-maleficence, (4) responsibility and (5) privacy (Jobin et al. 2019).

On a closer look, with regard to the socioethical implications of AI, a broad spectrum of topics, concepts and principles play a role (Asaro 2019; Bostrom and Yudkowsky 2014; Gunkel 2012, 2018; Lin et al. 2017; Stahl and Wright 2018; Yuste et al. 2017). These include:

- *Data Concerns:* Data management, data security, protection of personal data, surveillance, privacy, and informed consent.
- *Algorithmic Bias and Discrimination:* How to avoid bias and bias related problems? This points to questions of justice, equitable access to resources, and digital divide.
- *Autonomy:* When and how is AI autonomous, what are the characteristics of autonomous AI? How to develop rules for autonomous vehicles?

- *Responsibility:* Who is in control? Who is responsible or accountable for decisions made by AI?
- *Questions relating to AI capabilities:* Can AI ever be conscious or sentient? What would conscious or sentient AI imply?
- *Values and morality:* How to build in values and moral decision-making to AI? Are moral machines possible? Should robots be granted moral status or rights?

The questions of if and how autonomous systems should make ethical decisions have been of key interest to computer scientists, engineers and philosophers in the past 25 years (Allen et al. 2006; Anderson and Anderson 2007). How should we think about artificial moral agents, how should they be designed, and how can we effectively test the efficacy of these moral agents? Should these systems adopt a Kantian, utilitarian, or some other moral theory on which to base these judgements? In this topical collection, a number of the authors focus on decision-making, values, moral judgment and autonomy in AI. Nallur (2020) provides a survey of how computer scientists are working to implement ethical behavior in robots, unmanned autonomous vehicles and software systems. In the responding commentary, Bauer (2020) expands upon this by looking at small-scale and large-scale interactions among intelligence machines, and suggesting some potential ethical approaches that could be used in building ethical autonomous machines. With regard to decision-making in autonomous vehicles, Dubljević (2020) rejects the adoption of a single moral theory like utilitarianism and instead proposes the Agent-Deed-Consequence model of moral judgment that helps explain the flexibility and stability of human moral judgement that is currently not replicatiable in AI decision-making.

When and how is AI autonomous, and what are the characteristics of autonomous AI? The discussions around this topic range from more technical notions as to what counts as "autonomous" in terms of the need for human intervention (such as the different levels of vehicle autonomy in autonomous cars) to wider questions of machines as *subjects* of autonomy (SAE 2018; Moor 2006). What will an autonomous AI (in the philosophical rather than the technical sense) look like and what does this mean for its interactions with us? In "Fully Autonomous AI," Totschnig discusses the possibility of AI obtaining autonomy if it gains the ability to change its final goal, or its "utility function" (2020). The author discusses the finality argument, and after examining a number of objections to this argument, reflects on how an agent's utility function depends on its understanding of the world and its values. In the following commentary, Dennis (2020) provides a computer science view of this argument, and after looking at how an AI that cannot change its final goals might still be considered autonomous, she discusses computational models of values and ethics and how they may relate to the concept of values that Totschnig proposes.

Drawing away from Western conversations about what ethical framework AI systems should follow when working in real-world settings and making decisions, as well as the ontological properties a robot must have to decide whether it is an entity deserving of rights and dignity, Zhu et al. (2020) discuss how Confucian ethical traditions can help expand our discussions around AI ethics. As human–robot interactions continue to grow, what role should these autonomous systems play in trying to help shape our own ethical behavior? In his commentary, David Gunkel discusses

how Confucian, role-based ethics enhances the ways in which we can comprehend the moral standing of robots (2020).

In instances where artificial intelligence and the human mind converge, a host of existing and new ethical issues arise. Aicardi et al. (2020) provide an overview of the current state of neurorobotics development in the Human Brain Project, explore important social and ethical issues and investigate potential gaps in this collaborative project that need attention to ensure the development of neurorobotics is ethically sound, and socially acceptable and desirable. In the responding commentary, Taraban (2020) discusses the limits of using human cognition as a model for integrating neuroscience with robotics, and points out that to the extent that intelligent robots become a reality, more attention needs to be paid to the ethics of robot rights.

Reflections on ethics, values and decision-making in AI point to broader questions like: How do we want to shape human–robot interaction? Do we want robots to be as similar to humans as possible? How does AI technology influence human self-understanding and human–human relationships? And finally: What will the future of AI and brain-based intelligence be? While we would like to believe that our advances in predictive modeling would give us the ultimate answers, we cannot be certain. Based on the recent development of hundreds of AI ethics documents worldwide and progressive initiatives in neuroscience, it is increasingly clear that neuroscience, computer science, and humanities fields will be working together to shape the ethical futures of AI and neurotechnology. As neuroscience continues to understand and uncover mysteries of the brain, computer science will be nearby developing artificial intelligence that may or may not be brain-based. Continued work in neurotechnologies will encourage these two fields to work together and humanities fields will investigate how these developments in technology influence the human condition in various ways.

Notes on the Topical Collection

In this topical collection of Science and Engineering Ethics we attempt to contribute to the debate of socioethical implications of AI and neurotechnology. The topical collection dates from the workshop "Brain-based and Artificial Intelligence: Socioethical Conversations in Computing and Neurotechnology" held in May 2018 in Chicago, Illinois. In addition to articles deriving from presentations given at the workshop, the topical collection comprises additional submissions and commentaries.

References

Aicardi, C., Akintoye, S., Fothergill, B. T., Guerrero, M., Klinker, G., Knight, W., Klüver, L., Morel, Y., Morin, F. O., Stahl, B. C., & Ulnicane, I. (2020). Ethical and social aspects of neurorobotics. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00248-8.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems, 21*(4), 12–17. https://doi.org/10.1109/MIS.2006.83.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine, 28*(4), 15. https://doi.org/10.1609/aimag.v28i4.2065.

Asaro, P. M. (2019). AI ethics in predictive policing: From models of threat to an ethics of care. *IEEE Technology and Society Magazine, 38*(2), 40–53. https://doi.org/10.1109/MTS.2019.2915154.
Bauer, W. (2020). Expanding Nallur's landscape of machine implemented ethics. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00237-x.
Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Franish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence*. Cambridge: Cambridge University Press.
Brooks, R., Hassabis, D., Bray, D., & Shashua, A. (2012). Is the brain a good model for machine intelligence? *Nature, 482*, 462–463. https://doi.org/10.1038/482462a.
Butkus, M. (2020). The human side of artificial intelligence. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00239-9.
Dennis, L.A. (2020). Computational goals, values and decision-making. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00244-y.
Dubljević, V. (2020). Toward implementing the ADC model of moral judgment in autonomous vehicles. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00242-0.
Fan, X., & Markram, H. (2019). A brief history of simulation neuroscience. *Frontiers in Neuroinformatics*. https://doi.org/10.3389/fninf.2019.00032.
Farisco, M., Evers, K., & Sales, A. (2020). Towards establishing criteria for the ethical analysis of artificial intelligence. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00238-w.
Gunkel, D. (2020). Shifting perspectives. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00247-9.
Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. Boston: MIT Press.
Gunkel, D. J. (2018). *Robot rights*. Boston: MIT Press.
Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron, 95*(2), 245–258.
Holmes, F. (2019). AI will add $15 trillion to the world economy by 2030. *Forbes.* https://www.forbes.com/sites/greatspeculations/2019/02/25/ai-will-add-15-trillion-to-the-world-economy-by-2030/#4daa5d0f1852. Retrieved 27 May 2020.
Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2.
Jotterand, F., & Bosco, C. (2020). Keeping 'humans loop' in the age of artificial intelligence. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00241-1.
Lin, P., Abney, K., & Jenkins, R. (2017). *Robot ethics 2.0: From autonomous cars to artificial intelligence*. New York: Oxford University Press.
Markram, H. (2006). The blue brain project. *Nature Reviews Neuroscience, 7*, 153–160.
Matthias, A. (2020). Dignity and dissent in humans and non-humans. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00245-x.
Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18–21. https://doi.org/10.1109/MIS.2006.80.
Moore, J. (2019). AI for not bad. *Frontiers in Big Data*. https://doi.org/10.3389/fdata.2019.00032.
Nallur, V. (2020). Landscape of machine implemented ethics. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00236-y.
Pan, Y. (2016). Heading toward artificial intelligence 2.0. *Engineering, 2*(4), 409–413.
Rainey, S., & Erden, Y. J. (2020). Correcting the brain? The convergence of neuroscience, neurotechnology, psychiatry and artificial intelligence. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00240-2.
Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2020). What's next for AI ethics, policy and governance? A global overview. In *AIES '20: Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 153–158). https://doi.org/10.1145/3375627.3375804.
Society of Automotive Engineers, International. (2018). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*. J3016-201806. https://www.sae.org/standards/content/j3016_201806/. Retrieved 22 May 2020.
Stahl, B. C., & Wright, D. (2018). Ethics and privacy in AI and big data: Implementing responsible research and innovation. *IEEE Security & Privacy, 16*(3), 26–33. https://doi.org/10.1109/MSP.2018.2701164.

Taraban, R. (2020). The nature of neural computation in humans and machines. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00249-7.
Totschnig, W. (2020). Fully autonomous AI. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00243-z.
Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 49*, 433–460.
Ullman, S. (2019). Using neuroscience to develop artificial intelligence. *Science, 363*(6428), 692–693.
Yuste, R., et al. (2017). Four ethical priorities for neurotechnologies and AI. *Nature, 551*, 159–163.
Zhu, Q., Williams, T., Jackson, B., & Wen, R. (2020). Blame-Laden moral rebukes and the morally competent robot: A confucian ethical perspective. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00246-w.