The Future of AI Ethics and the Role of Neurotechnology

Sara Berger, Francesca Rossi

IBM T.J. Watson Research Lab, IBM Research, USA

Abstract

As scientists and researchers, we are not only called to create and advance technology but also are obliged to ensure that new technology is designed and utilized in ways that are inclusive and trustworthy, that align with societal values, and that hold companies and people accountable to stringent and transparent standards. This responsibility is not new and is been widely discussed in the context of AI ethics. In this paper we point out that AI is gradually merging with neurotech, bringing new potential for positive impacts on our lives but also additional concerns. Therefore, we need to extend our capabilities to identify and address emerging ethics issues. However, we can and should exploit all the knowledge gathered and the capabilities developed in the area of AI ethics to accelerate the path towards addressing the expanded or even new issues raised by the combination of AI and neurotechnologies. This paper represents a first step in identifying the ethical issues and pointing out what neurotech brings to the discussion. It is also a call to action to AI ethics thought leaders and practitioners to support a larger multi-disciplinary and multi-stakeholder approach that includes expertise in AI+neurotech research, technology, and deployed solutions. Given that neurotechnologies are still emerging, there is a unique opportunity to learn from the past, think proactively about potential issues and negative impacts, and develop preventative technical, societal, and educational solutions before problems arise.

Keywords

Artificial Intelligence, AI Ethics, Neuroscience, Neurotechnology, Neuroethics

1. Current Issues in AI Ethics

AI is a science and a technology that has applications in almost every aspect of our everyday life. We use it when we swipe a credit card, when we search something on the web, when we take a picture with our cameras, when we give vocal commands to our phone or another device, and when we interact with many apps and social media platforms. Companies of every size and business model, all over the world, are adopting AI solutions to optimise their operations, create new services and work modalities, and help their professionals in making more informed and better decisions.

There is no doubt that AI is a powerful technology that has already imprinted itself positively on our ways of living and will continue to do so for years to come. At the same time, the transformations it brings to our personal and professional lives are very significant and fast, and this raises questions and concerns about the impact of AI on our society. AI systems need to be designed to be aware of, and to follow, important human values so that the technology

AIofAI'21: 1st Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies, Montreal, CA

asara.e.berger@ibm.com (S. Berger); francesca.rossi2@ibm.com (F. Rossi)

can help us make better, wiser decisions that are simultaneously human-value-aligned.

Here are some of main current AI ethics issues:

- **Data issues**: AI needs a lot of data, so questions about data privacy, storage, sharing, and governance are central for this technology. In some regions of the world, such as Europe, there are specific regulations to state fundamental rights for the *data subject*, the human being releasing personal data to an AI system that can then use it to make decisions affecting his/her life [1].
- **Explainability and trust**: Often the most successful AI techniques, such as those based on machine learning, are opaque in terms of allowing humans to understand how they reach their conclusions from the input data. This does not help when trying to build a system of trust between humans and machines, so it is important to adequately address concerns related to transparency and explainability (see [2] and [3] for example of tools that provide solutions to these issues). Without trust, a doctor will not follow the recommendation of a decision support system that can help in making better decisions for his/her patients.
- **Accountability**: Machine learning is based on statistics, so it always has a percentage of error, even if small. This happens even if no programmer actually made any mistake in developing the AI system. So, when an error occurs, who is responsible? To whom should we ask for a redress or a compensation? This raises questions related to responsibility and accountability.
- **Fairness**: Based on huge amounts of data that surround every human activity, AI is able to derive insights and knowledge on which to make decisions about or recommends decisions to a human being. However, we need to make sure that the AI system understands and follows the relevant human values in the context in which such decisions are made. A very important human value is fairness: we don't want AI systems to make (or recommend) decisions that could discriminate against or perpetuate harm across groups of people. How do we make sure that AI can act according to the most appropriate notion of fairness (or any other human value) in each scenario in which it is applied? See [4] for an open-source library and [5] for a description of the multiple dimensions of AI fairness.
- **Profiling and manipulation**: AI can interpret our actions and the data we share online to make a *profile* of us, a sort of abstract characterization of some of our traits, preferences, and values, to be used to personalize services (for example, to show us posts or ads that we most likely will appreciate). Without appropriate guardrails, this approach can twist the relationship between humans and online service providers by designing the service in order to make our preferences more clearly characterised, and thus the personalization easier to compute. This raises issues of human agency: are we really in control of our actions, or is AI being used to nudge us to the point of manipulating us?
- **Impact on jobs and larger society**: Since AI permeates our workplace functioning, it obviously has an impact on jobs (since it can perform some cognitive tasks that usually were done by humans), and these impacts need to be better understood and addressed [6] to make sure humans aren't disadvantaged. As mentioned earlier, AI is very pervasive and its applicability expands very rapidly, so any negative impacts of this technology could be extremely detrimental for individuals and society. The pace at which AI is being

applied within the workplace (and outside of it) also brings concerns about people and institutions having enough time to understand the real consequences of its use and avoid a possible negative impact.

- **Control and value alignment**: Although AI has a lot of applications, it is still very far from achieving forms of intelligence close to humans (or even animals). However, the fact that this technology is mostly unknown to the general public raises concerns about being able to control it and to make it aligned to our larger and sometimes disparate societal values, should it achieve a higher form of intelligence [7, 8].

No single company can address and solve all these issues alone. This is why AI ethics includes experts from many scientific and technological disciplines. Indeed, the AI ethics community includes AI experts, philosophers, sociologists, psychologists, lawyers, policy makers, civil society organizations, and many others. Only by including all the voices (those that produce AI, those that use AI, those that regulate AI, those that are impacted by AI's decisions, and those that understand how to evaluate the impact of a technology on people and society) can we understand how to identify and address the AI ethics concerns.

Technical solutions to AI ethics concerns, such as software tools and algorithms to detect and mitigate bias (see for example [4]), or to embed explainability into an AI system (see for example [2]) are certainly necessary. But they are not enough: also non-technical solutions, such as guidelines, principles, best practices, educational and re-skilling activities [9], standards [10], audits, and laws [11] are being considered. On top of these, there is the need to specific methodologies to operationalize AI ethics principles and create approriate governance around them.

2. AI, Neuroscience, and Neurotechnology

As we learn more about the nervous system and untangle the embodied and bidirectional interactions between our external environments and internal milieus, the need for new tools and capabilities increases. This is not only to meet the demands of basic bench, translational, and clinical neuroscience research by creating more advanced methods and materials for capturing neural signals and computing neural features, but also to provide novel therapies, generate new ways to restore or equalize human functions, and create resources to augment and enrich our existing skills and experiences. At the same time, the capabilities of AI are continuously expanding, becoming more complex, efficient, and faster due to numerous advances in computing power. AI, and especially machine learning, is increasingly being used in neuroscience applications. But the links between AI and neuroscience have been strong since since the first days of AI.

Indeed, conceptual linkages between AI and neuroscience have been around since the emergence of AI as a research field. Goals of emulating and augmenting human intelligence via machines that can "learn", common and often contentious (see [12]) brain-is-a-computer and computer-is-a-brain metaphors (and associated colloquialisms like "I don't have the bandwidth" or "my computer is out of memory") (see [13]), and more recent neuro-inspired computing techniques like neural nets, neuromorphic algorithms, and deep learning are all examples of the intertwined trajectories of the two fields. The associations between our minds and machine capabilities are only strengthening as AI becomes embedded into nearly every aspect of our lives – from our "smart" phones to our "smart" fridges and from our shopping habits to our social media, it permeates our jobs, our homes, our transportation, our healthcare (see [14]) and our interactions with others.

The inevitable movement beyond conjectural linkages into real world interactions between computation and neuroscience has begun. Albeit indirectly, AI already pervasively interacts with our nervous systems by influencing, reinforcing, and changing our behaviors and cognitive processes. While the concept of an "extended mind" is not new (see [15]), until relatively recently, humans were largely limited to extending their thoughts into the physical realm via representations such as symbols, writings, art, or spatial markers, stored on the walls of caves or on canvases, within books and diaries, or as signs in the environment; these tools and relics functioned as repositories of ideas, memories, directional aids, and external expressions of our internal selves. Now, however, we are extending the neural into the digital (see [16]), and the more pervasive AI and digital technologies become, the more intertwined and almost inseparable they are with our nervous systems and associated abilities and psychology.

Simultaneously, these indirect and often theoretical links between AI and neuroscience are transforming into direct and tangible ones, from one-way extensions of our minds into digital spaces to bidirectional connections between nervous systems and computers. Over the last few decades, we have seen a rise in the development and deployment of devices that exploit the advances in computing and the pervasiveness of AI to collect, interpret, infer, learn from, and even modify various signals generated throughout the entire nervous system (called *neurodata* or *neuroinformation*) – these devices are called *neurotechnologies* (i.e., ***neurotech***). Neurotech can interact with neurodata either invasively and directly through different kinds of surgical implants, like electrodes or devices implanted into or near neuronal tissues, or they can interact non-invasively and indirectly through wearable devices sitting on the surface of the skin, picking up signals or proxies of those signals from the head, body, or limbs. Generally, neurotech is divided into three categories - (1) *neurosensing*, which essentially "reads" neurodata by collecting, monitoring, or interpreting it; (2) *neuromodulating*, which "writes" data by changing the electrical activity, chemical makeup, and/or structure of the nervous system; and (3) *combinatorial* or bidirectional, which can both read and write neurodata, so to speak (see [17]).

As neurotechnology is still emerging, the state of the art is constantly evolving. At present [18], scientists can invasively record from hundreds of neurons simultaneously (a number which will soon become 1000s with the advent and increasing adoption of interfaces like neuralace and neural threads). Neurotech can decode and project specific forms of thought like imagined handwriting, typing, and other kinds of intended movements, very crudely reconstruct conscious and unconscious mental imagery [19], treat a gamut of chronic illnesses or injuries spanning neurological and psychological ailments, and is beginning to restore movement and sensation in people with missing or damaged limbs or those with spinal cord severation. Neuroscientists have also demonstrated the technological capability to transfer sensory and perceptual experiences and memories directly between animals, a capability that is slowly and rudimentarily being developed and tested between people [17]. While once relegated to the realm of science fiction, the merging of machines, bodies, and psyches is on the horizon due to the technological advancements enabled by neuroscience and AI.

3. Neuroethics

Given the important implications of neurotech on society, the relative immaturity of their techniques and inferences, and the growing direct-to-consumer push of their capabilities, there are concerns that the commercialization of neurotech and the commodification of neurodata is moving at a speed and scale that could proceed without proper policies and regulations in place to protect future consumers. Likewise, if history is any indication of the future, the increase in cautionary tales from AI applications resulting in community harms and reactionary mitigation strategies only further strengthens the need for proactive guardrails to be developed within the AI-enabled neurotech space.

However, in order for agreed-upon standards and best practices to be put in place, we must first understand the ethical concerns associated with neurotechnologies and how these compare to those seen in AI. The study of ethical principles and implications related to the development, deployment, and use of neurotechnologies (and associated neuroscience research and neurodata) is commonly referred to as a *neuroethics*, a nascent but growing field of inquiry emerging in the late 1990s and early 2000s out of medical and bio ethics [20]. Neuroethics is critical about the assumptions and intentions underlying neurotech and neuroscientific findings, concerned with questions about neurotech's impact on human self-understanding and the downstream effects of changes in this fundamental understanding on our biology, our psychology, and our society [16].

The ethical considerations surrounding neurotech are still being researched as there is yet so much to learn about the nervous system and about how, and the extent to which, neurotechnology will influence humanity. However, there are at least eight core neuroethics issues that consistently emerge which could pose significant societal, technical, and legal challenges. These are briefly defined below in a list of concept, rights, and values that could be impacted by the use of neurotechnology and therefore need to be protected:

- **Mental Privacy**: A condition met when one's neurodata is free from unconsented observation, intrusion, interpretation, collection, or disturbance by third parties or unauthorized neurotech devices.
- **Human Agency and Autonomy**: The ability to act or think with intention, in the absence of coercion or manipulation, with sufficient information to make rational choices in decisions regarding one's mind and body.
- **Human Identity**: The subjective, complex, and dynamic embodiment of various aspects of human reality - including but not limited to biology, culture, ecology, lived experiences, and historic socio-political situations - which together give rise to each person's unique ideas of meaning, relatedness to others and the world, and conceptions of self and ownership of life; a phenomena that is simultaneously unique and literally inscribed within the nervous system while also being influenced and constructed by external forces and societal constructs [21].
- **Fairness**: Equitable and just treatment of individuals irrespective of their choice to use neurotech or not or to participate in neurodata collection or not, and in manners regarding access to neurotech abilities and participation in neurotech design influence, neurotech solutions, and neurodata interpretations.

- **Accuracy**: The correctness of neurodata measurements, interpretations provided by neurotech, or code generated by neurotech for the purpose of modifying the nervous system.
- **Transparency**: The quality of being clear and open about the capabilities of neurotech, the usage of neurodata, and any inferences drawn from either.
- **Security**: A set of technologies, standards, and procedures which protect neurodata and data inferred from neurotech from access, disclosure, modification, or destruction by unauthorized users.
- **Well-being**: A prioritized state of physical and mental satisfaction (including health, safety, happiness, and comfort) achieved through both the avoidance of negligence and the prevention of harm, injury, or unreasonable risk of either in the design or implementation of neurotech (or the usage of associated neurodata).

Importantly, these concerns are not mutually exclusive and are considerably inter-related. For example, obtaining informed consent to collect neurodata would involve mental privacy, human autonomy and agency, transparency, and data security assurance; likewise, well-being is met when all other concerns are sufficiently addressed. Additionally, many of these concepts invoke previously established bioethical and medical principles related to beneficence, nonmaleficence, dignity, and justice, indicating that they are relevant to a broader range of applications *outside* of neurotech.

In practice, responding to these concerns will mean answering challenging questions on a contextual, case-by-cases basis, as the extent of risk changes depending on: the neurodata of interest, how the neurodata is being treated (e.g., is it being read, written, or both?), whose neurodata is being collected (and by whom for what purposes), what the participant's or end-user's overall literacy is in the space (e.g., do they understand what the neurotech's capabilities are or how sensitive their neurodata is?), and the location that the discussion or application is taking place (e.g., the specific impacted community, the cultural norms, the societal expectations, the associated politics and regulations, or any environmental considerations if applicable).

4. Ethics in the age of AI and Neurotech

When considering the list above in the context of AI ethics, it's clear that neurotech poses both familiar and new ethical challenges. AI ethics and Neuroethics largely converge along issues of data security, transparency, and general well-being. Areas where the fields' ethics issues overlap are important to highlight because it suggests that some of the existing solutions (both technological and not) that AI uses might be able to be applied to neurotech applications to mitigate these particular concerns. However, neurotech also poses risks which may not be sufficiently covered by existing AI regulations, governance frameworks, best practices, or company policies, or may be unique or *novel* based on the kinds of neurotech abilities, and this divergence of issues indicates the need to develop new preventative strategies, policies, and technological solutions.

In particular, two of the eight neuroethics considerations identified– mental privacy and human identity – entail wholly new kinds of challenges not presented by AI. These are of particular concern, given that it is feasible that neurotech could one day both collect neurodata and write new information into our nervous systems, all without being detected. Consent-based solutions may work with AI, but are not sufficient in the neurotech context. This is because the majority of our nervous system's signals are unconscious and outside of our awareness or control, making it difficult to establish what kind of neurodata we consent to share and also technically challenging to precisely pinpoint the kinds of data neurotech collects, interprets, or modulates in the first place. One day, it might be plausible to unknowingly or unintentionally provide neurotech with information that one wouldn't have otherwise, and thus in some applications of neurotech the presumption of privacy within one's own mind may no longer be a certainty. Furthermore, the fact that some neurotech can modify on-going neural activity and feedback (or write) data into the nervous system in real-time, a capability that no other technology has, produces new questions about how we can protect and ensure bodily autonomy and decisional capacity; this includes the potential for changing (purposefully or not, measurable or not) the integrity of our mental processes including our conceptions of identity. Because neurotech may one day be able to directly influence a person's behaviors, thoughts, emotions, memories, perceptions, or relationships between these phenomena, it poses challenging questions about free will, cognitive liberty, and notions of self-hood that we haven't had to truly address at this level before.

The remaining neuroethics considerations highlight challenges that are similar to those posed by AI, but that also may be appreciably different or *expanded* given the potential capabilities of neurotech and the sensitivity of neurodata. For example, fairness is a substantial component of AI ethics and some of the same issues surrounding equitable access and inclusion in design and interpretation are also in neuroethics. But, neurotech may one day allow us to infer and act on neurodata that we are unaware of (e.g., unconscious bias), as well as significantly augment or change our mental and physical abilities. Thus the risk to fairness is greater with neurotech, as these kinds of capabilities could perpetuate existing inequities and biases or create new avenues for discrimination or malicious targeting that are even harder to see because they are quite literally hidden *inside* of us. Additionally, underlying and unchallenged assumptions about what constitutes "normative" neurodata or what is considered desirable neurotech outcomes may also be biased against people with hidden disabilities or neurological differences. Moreover, neurotech interfaces may exacerbate or compound issues we are currently seeing with AI, if device sensors do not adequately account for different hair types or skin colors, or instructions and interfaces are designed in a way that widens instead of bridges the current digital or technological divides

An initial comparison of ethical concerns between neurotech and AI is summarized in the table below but may not be exhaustive as new considerations and differences may emerge with the unfolding of both AI and neurotech.

Once we identify the issues around the combined use of AI and neurotech, how can we address them? As the above table shows, some issues are really new or greatly expanded compared to AI, so we possibly need to deploy new solutions, both technical and not, to address them. However, the good news is that we don't need to start from scratch. Over the past 5 or so years, there has been a lot of foundational work completed to address AI ethics issues: we've constructed and utilized multi-stakeholder approaches to identify the concerning issues and their impacts, specified best practices, principles, and guidelines, built technical solutions and adopted educational/training methodologies, created governance frameworks and international standards, and even defined hard laws based on AI ethics considerations (such as the very recent one by the European Commission, see [11]). While doing all this, we've learned several lessons, identified challenges, listed the failures, and reported on successful approaches. By "we" we mean the whole society, not just AI experts: experts of many scientific disciplines, together with business leaders, policy makers, and civil society organizations.

Therefore, we can and should exploit all this knowledge and the developed capabilities to accelerate the path towards addressing the issues raised by the combination of AI and neurotech. The first step is of course to clearly map the relationship between the new and old issues, that we started doing in the previous sections and summarised in Table 1.

After this, we will be able to update and augment the current AI ethics frameworks and actions to also cover many if not all of the new issues. We will likely need to involve other experts from other disciplines that are now rarely found in existing AI ethics initiatives, such as neuroscience and neuroethics, to fully understand the current state of the art in neurotech and the real implications on humans and society. What is considered to be "multi-stakeholder" will now need to be greatly expanded if we want to correctly identify the issues in this broader technological/scientific context, define the relevant principles and values, and then build the necessary concrete actions.

5. Conclusions

AI is gradually merging with neurotech, bringing new potential for positively impacting our life but also new or expanded ethics concerns. Therefore, we need to expand our capabilities to address ethics issues.

With this paper, we wanted to point out this inevitable evolution and convergence to those who are actively thinking and working in AI ethics, with the hope to start a more general conversation and an accelerated path to addressing the identified neuroethics issues. AI is already a powerful and often positive technology in our life. Combined with neurotech, it will bring huge new benefits in healthcare, work, leisure, and many other aspects of our lives. But as we know, greater power comes with additional responsibilities. Knowledge should advance at the same pace as wisdom and awareness of human values and societal forces, so that technology progress can be beneficial for all.

Given that neurotechnologies are still emerging, there is a unique opportunity to learn from the past, think proactively about potential issues, and develop preventative technical, societal, and educational solutions *before* problems arise.

References

[1] European Parliament, General Data Protection Regulation, https://gdpr-info.eu/, 2018.
[2] IBM Research, AI explainability 360, https://aix360.mybluemix.net/, 2019.
[3] IBM Research, AI factsheet 360, https://aifs360.mybluemix.net/, 2019.
[4] IBM Research, AI fairness 360, https://aif360.mybluemix.net/, 2018.

[5] F. Rossi, How IBM is working toward a fairer AI, https://hbr.org/2020/11/how-ibm-is-working-toward-a-fairer-ai, 2020.

[6] E. Brynjolfsson, A. McAfee, The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies, W. W. Norton & Company, 2015.

[7] S. Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Viking, 2019.

[8] M. Tegmark, Life 3.0, Knopf, 2017.

[9] IBM, Everyday Ethics for AI, https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf, 2018.

[10] IEEE, P7000 standard series, https://ethicsinaction.ieee.org/p7000/, 2019.

[11] European Commission, Proposal for a regulation laying down harmonised rules on artificial intelligence (artificial intelligence act), https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence, 2021.

[12] M. Cobb, Why your brain is not a computer, The Guardian (2020). URL: https://www.theguardian.com/science/2020/feb/27/why-your-brain-is-not-a-computer-neuroscience-neural-networks-consciousness.

[13] B. Richards, Yes, the brain is a computer, Medium (2018). URL: https://medium.com/the-spike/yes-the-brain-is-a-computer-11f630cad736.

[14] Food, Drug Administration, AI/ML software as a medical device, https://www.fda.gov/media/145022/download, 2021.

[15] A. Clark, D. Chalmers, The extended mind, Analysis 58 (1998) 7–19.

[16] J. Illes, Neuroethics - Anticipating the Future, Oxford University Press, New York, New York, 2017.

[17] Royal Society, Ihuman: blurring lines between mind and machine, London: The Royal Society (2019).

[18] J. L. Contreras-Vidal, State-of-the-Art BCI Device Technology, https://www.fda.gov/media/90434/download, 2014.

[19] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, J. L. Gallant, Reconstructing visual experiences from brain activity evoked by natural movies, Current Biology 21 (2011) 1641–1646.

[20] J. Leefmann, C. Levallois, E. Hildt, Neuroethics 1995-2012. a bibliometric analysis of the guiding themes of an emerging research field, Frontiers in Human Neuroscience 10 (2016) 336.

[21] G. R. Gillett, The subjective brain, identity, and neuroethics, American Journal of Bioethics 9 (2009) 5–13.

Ethics Issue	**AI**	**Neurotech**	**Difference**	**Motivation**
Data issues	X	X	Expanded	Neurodata is also handled but may need more requirements due to data complexity, data sensitivity, and lack of agreed upon governance models.
Explainability and trust	X	X	Expanded	Less is known about the immediate and downstream impacts of these technologies on our nervous systems.
Accountability	X	X	Expanded	Immaturity of neurotech interfaces and sensors, lack of replication in findings, and limited expertise in neurodata interpretability.
Fairness	X	X	Expanded	Ability to act on neurodata we are unaware of and the potential to augment human abilities.
Profiling and manipulation	X	X	Expanded	More data, and of more intimate kind, can be used for these purposes.
Impact on society	X	X	Expanded	More applications are available when AI and neurotech are combined.
Control and value alignment	X	X	Same	
Mental privacy		X	New	AI only has access to external proxies of thoughts and behaviors, not neurodata.
Human autonomy and agency	X	X	Expanded	Neurotech can "write" into our nervous system and directly influence our minds and abilities .
Human identity		X	New	AI only has access to external proxies of identity (not neurodata) and can only indirectly influence identity, whereas neurotech may be able to directly influence it and do so in ways that may or may not be controllable or easily measurable.
Accuracy	X	X	Expanded	Lack of knowledge about the nervous system (and large levels of noise) may lead to lower accuracy and precision.
Transparency	X	X	Same	
Security	X	X	Expanded	Neurodata may deal with more private or sensitive information; certain neurodata will also be identifying.

Table 1
AI vs neurotech ethics issues.