Published in final edited form as:

Nature. 2017 November 08; 551(7679): 159–163. doi:10.1038/551159a.

Four ethical priorities for neurotechnologies and AI

Rafael Yuste [professor of biological sciences],
Columbia University, New York City, New York, USA

Sara Goering [associate professor of philosophy]
University of Washington, Seattle, USA

Abstract

Artificial intelligence and brain-computer interfaces must respect and preserve people's privacy, identity, agency and equality, say Rafael Yuste, Sara Goering and colleagues.

Consider the following scenario. A paralysed man participates in a clinical trial of a brain–computer interface (BCI). A computer connected to a chip in his brain is trained to interpret the neural activity resulting from his mental rehearsals of an action. The computer generates commands that move a robotic arm. One day, the man feels frustrated with the experimental team. Later, his robotic hand crushes a cup after taking it from one of the research assistants, and hurts the assistant. Apologizing for what he says must have been a malfunction of the device, he wonders whether his frustration with the team played a part.

This scenario is hypothetical. But it illustrates some of the challenges that society might be heading towards.

Current BCI technology is mainly focused on therapeutic outcomes, such as helping people with spinal-cord injuries. It already enables users to perform relatively simple motor tasks — moving a computer cursor or controlling a motorized wheelchair, for example. Moreover, researchers can already interpret a person's neural activity from functional magnetic resonance imaging scans at a rudimentary level[1] — that the individual is thinking of a person, say, rather than a car.

It might take years or even decades until BCI and other neurotechnologies are part of our daily lives. But technological developments mean that we are on a path to a world in which it will be possible to decode people's mental processes and directly manipulate the brain mechanisms underlying their intentions, emotions and decisions; where individuals could communicate with others simply by thinking; and where powerful computational systems linked directly to people's brains aid their interactions with the world such that their mental and physical abilities are greatly enhanced.

rafaelyuste@columbia.edu; sgoering@uw.edu.
A list of 25 co-authors in The Morningside Group appears in the online version.

A full list of authors accompanies this Comment online; see go.nature.com/2ij9bqt.

Such advances could revolutionize the treatment of many conditions, from brain injury and paralysis to epilepsy and schizophrenia, and transform human experience for the better. But the technology could also exacerbate social inequalities and offer corporations, hackers, governments or anyone else new ways to exploit and manipulate people. And it could profoundly alter some core human characteristics: private mental life, individual agency and an understanding of individuals as entities bound by their bodies.

It is crucial to consider the possible ramifications now.

The Morningside Group comprises neuro-scientists, neurotechnologists, clinicians, ethicists and machine-intelligence engineers. It includes representatives from Google and Kernel (a neurotechnology start-up in Los Angeles, California); from international brain projects; and from academic and research institutions in the United States, Canada, Europe, Israel, China, Japan and Australia. We gathered at a workshop sponsored by the US National Science Foundation at Columbia University, New York, in May 2017 to discuss the ethics of neurotechnologies and machine intelligence.

We believe that existing ethics guidelines are insufficient for this realm[2]. These include the Declaration of Helsinki, a statement of ethical principles first established in 1964 for medical research involving human subjects (go.nature.com/2z262ag); the Belmont Report, a 1979 statement crafted by the US National Commission for the Protection of Human Subjects of Biomedical and Behavioural Research (go.nature.com/2hrezmb); and the Asilomar artificial intelligence (AI) statement of cautionary principles, published early this year and signed by business leaders and AI researchers, among others (go.nature.com/2ihnqac).

To begin to address this deficit, here we lay out recommendations relating to four areas of concern: privacy and consent; agency and identity; augmentation; and bias. Different nations and people of varying religions, ethnicities and socio-economic backgrounds will have differing needs and outlooks. As such, governments must create their own deliberative bodies to mediate open debate involving representatives from all sectors of society, and to determine how to translate these guidelines into policy, including specific laws and regulations.

INTELLIGENT INVESTMENTS

Some of the world's wealthiest investors are betting on the interplay between neuroscience and AI. More than a dozen companies worldwide, including Kernel and Elon Musk's start-up firm Neuralink, which launched this year, are investing in the creation of devices that can both 'read' human brain activity and 'write' neural information into the brain. We estimate that current spending on neurotechnology by for-profit industry is already US$100 million per year, and growing fast.

Investment from other sectors is also considerable. Since 2013, more than $500 million in federal funds has gone towards the development of neurotechnology under the US BRAIN initiative alone.

Current capabilities are already impressive. A neuroscientist paralysed by amyotrophic lateral sclerosis (ALS; also known as Lou Gehrig's or motor neuron disease) has used a BCI to run his laboratory, write grant applications and send e-mails[3]. Meanwhile, researchers at Duke University in Durham, North Carolina, have shown that three monkeys with electrode implants can operate as a 'brain net' to move an avatar arm collaboratively[4]. These devices can work across thousands of kilometres if the signal is transmitted wirelessly by the Internet.

PROTECTING PRIVACY

Federated learning

When technology companies use machine learning to improve their software, they typically gather user information on their servers to analyse how a particular service is being used and then train new algorithms on the aggregated data. Researchers at Google are experimenting with an alternative method of artificial-intelligence training called federated learning. Here, the teaching process happens locally on each user's device without the data being centralized: the lessons aggregated from the data (for instance, the knowledge that the word 'weekly' can be used as an adjective and an adverb) are sent back to Google's servers, but the actual e-mails, texts and so on remain on the user's own phone. Other groups are exploring similar ideas. Thus, information systems with improved designs could be used to enhance users' ownership and privacy over their personal data, while still enabling valuable computations to be performed on those data.

Soon such coarse devices, which can stimulate and read the activity of a few dozen neurons at most, will be surpassed. Earlier this year, the US Defense Advanced Research Projects Agency (DARPA) launched a project called Neural Engineering System Design. It aims to win approval from the US Food and Drug Administration within 4 years for a wireless human brain device that can monitor brain activity using 1 million electrodes simultaneously and selectively stimulate up to 100,000 neurons.

Meanwhile, Google, IBM, Microsoft, Facebook, Apple and numerous start-ups are building ever-more-sophisticated artificial neural networks that can already outperform humans on tasks with well-defined inputs and outputs.

Last year, for example, researchers at the University of Washington in Seattle demonstrated that Google's FaceNet system could recognize one face from a million others. Another Google system with similar neural-network architecture far outperforms well-travelled humans at guessing where in the world a street scene has been photographed, demonstrating the generality of the technique. In August, Microsoft announced that, in certain metrics, its neural network for recognizing conversational speech has matched the abilities of even trained professionals, who have the option of repeatedly rewinding and listening to words used in context. And using electroencephalogram (EEG) data, researchers at the University of Freiburg in Germany showed in July how neural networks can be used to decode planning-related brain activity and so control robots[5].

Future neural networks derived from a better understanding of how real ones work will almost certainly be much more powerful even than these examples. The artificial networks in current use have been inspired by models of brain circuits that are more than 50 years old, which are based on recording the activity of individual neurons in anaesthetized animals[6]. In today's neuroscience labs, researchers can monitor and manipulate the activity of thousands of neurons in awake, behaving animals, owing to advances in optical methods, computing, molecular engineering and microelectronics.

We are already intimately connected to our machines. Researchers at Google calculated this year that the average user touches their phone nearly one million times annually (unpublished data). The human brain controls auditory and visual systems to decipher sounds and images, and commands limbs to hold and manipulate our gadgets. Yet the convergence of developments in neurotechnologies and AI would offer something qualitatively different — the direct linking of people's brains to machine intelligence, and the bypassing of the normal sensorimotor functions of brains and bodies.

FOUR CONCERNS

For neurotechnologies to take off in general consumer markets, the devices would have to be non-invasive, of minimal risk, and require much less expense to deploy than current neurosurgical procedures. Nonetheless, even now, companies that are developing devices must be held accountable for their products, and be guided by certain standards, best practices and ethical norms.

We highlight four areas of concern that call for immediate action. Although we raise these issues in the context of neurotechnology, they also apply to AI.

Privacy and consent.

An extraordinary level of personal information can already be obtained from people's data trails. Researchers at the Massachusetts Institute of Technology in Cambridge, for example, discovered in 2015 that fine-grained analysis of people's motor behaviour, revealed through their keyboard typing patterns on personal devices, could enable earlier diagnosis of Parkinson's disease[7]. A 2017 study suggests that measures of mobility patterns, such as those obtained from people carrying smartphones during their normal daily activities, can be used to diagnose early signs of cognitive impairment resulting from Alzheimer's disease[8].

Algorithms that are used to target advertising, calculate insurance premiums or match potential partners will be considerably more powerful if they draw on neural information — for instance, activity patterns from neurons associated with certain states of attention. And neural devices connected to the Internet open up the possibility of individuals or organizations (hackers, corporations or government agencies) tracking or even manipulating an individual's mental experience.

> "Some of the world's wealthiest investors are betting on the interplay between neuroscience and AI."

We believe that citizens should have the ability — and right — to keep their neural data private (see also ‘Agency and identity’). We propose the following steps to ensure this.

For all neural data, the ability to opt out of sharing should be the default choice, and assiduously protected. People readily give up their privacy rights to commercial providers of services, such as Internet browsing, social media or entertainment, without fully understanding what they are surrendering. A default of opting out would mean that neural data are treated in the same way that organs or tissues are in most countries. Individuals would need to explicitly opt in to share neural data from any device. This would involve a safe and secure process, including a consent procedure that clearly specifies who will use the data, for what purposes and for how long.

Even with this approach, neural data from many willing sharers, combined with massive amounts of non-neural data — from Internet searches, fitness monitors and so on — could be used to draw ‘good enough’ conclusions about individuals who choose not to share. To limit this problem, we propose that the sale, commercial transfer and use of neural data be strictly regulated. Such regulations

— which would also limit the possibility of people giving up their neural data or having neural activity written directly into their brains for financial reward — may be analogous to legislation that prohibits the sale of human organs, such as the 1984 US National Organ Transplant Act.

Another safeguard is to restrict the centralized processing of neural data. We advocate that computational techniques, such as differential privacy or ‘federated learning’, be deployed to protect user privacy (see ‘Protecting privacy’). The use of other technologies specifically designed to protect people’s data would help, too. Blockchain-based techniques, for instance, allow data to be tracked and audited, and ‘smart contracts’ can give transparent control over how data are used, without the need for a centralized authority. Lastly, open-data formats and open-source code would allow for greater transparency about what stays private and what is transmitted.

Agency and identity.

Some people receiving deep-brain stimulation through electrodes implanted in their brains have reported feeling an altered sense of agency and identity. In a 2016 study, a man who had used a brain stimulator to treat his depression for seven years reported in a focus group[9] that he began to wonder whether the way he was interacting with others — for example, saying something that, in retrospect, he thought was inappropriate — was due to the device, his depression or whether it reflected something deeper about himself. He said: “It blurs to the point where I’m not sure...frankly, who I am.”

Neurotechnologies could clearly disrupt people’s sense of identity and agency, and shake core assumptions about the nature of the self and personal responsibility — legal or moral.

People could end up behaving in ways that they struggle to claim as their own, if machine learning and brain-interfacing devices enable faster translation between an intention and an action, perhaps by using an ‘auto-complete’ or ‘auto-correct’ function. If people can control

devices through their thoughts across great distances, or if several brains are wired to work collaboratively, our understanding of who we are and where we are acting will be disrupted.

As neurotechnologies develop and corporations, governments and others start striving to endow people with new capabilities, individual identity (our bodily and mental integrity) and agency (our ability to choose our actions) must be protected as basic human rights.

We recommend adding clauses protecting such rights (‘neurorights’) to international treaties, such as the 1948 Universal Declaration of Human Rights. However, this might not be enough — international declarations and laws are just agreements between states, and even the Universal Declaration is not legally binding. Thus, we advocate the creation of an international convention to define prohibited actions related to neurotechnology and machine intelligence, similar to the prohibitions listed in the 2010 International Convention for the Protection of All Persons from Enforced Disappearance. An associated United Nations working group could review the compliance of signatory states, and recommend sanctions when needed.

Such declarations must also protect people’s rights to be educated about the possible cognitive and emotional effects of neurotechnologies. Currently, consent forms typically focus only on the physical risks of surgery, rather than the possible effects of a device on mood, personality or sense of self.

Augmentation.

People frequently experience prejudice if their bodies or brains function differently from most[10]. The pressure to adopt enhancing neurotechnologies, such as those that allow people to radically expand their endurance or sensory or mental capacities, is likely to change societal norms, raise issues of equitable access and generate new forms of discrimination.

Moreover, it’s easy to imagine an augmentation arms race. In recent years, we have heard staff at DARPA and the US Intelligence Advanced Research Projects Activity discuss plans to provide soldiers and analysts with enhanced mental abilities (‘super-intelligent agents’). These would be used for combat settings and to better decipher data streams.

Any lines drawn will inevitably be blurry, given how hard it is to predict which technologies will have negative impacts on human life. But we urge that guidelines are established at both international and national levels to set limits on the augmenting neurotechnologies that can be implemented, and to define the contexts in which they can be used — as is happening for gene editing in humans.

> “Outright bans of certain technologies could simply push them underground.”

Privacy and individuality are valued more highly in some cultures than in others. Therefore, regulatory decisions must be made within a culture-specific context, while respecting universal rights and global guidelines. Moreover, outright bans of certain technologies could simply push them underground, so efforts to establish specific laws and regulations must include organized forums that enable in-depth and open debate.

Such efforts should draw on the many precedents for building international consensus and incorporating public opinion into scientific decision-making at the national level[11]. For instance, after the First World War, a 1925 conference led to the development and ratification of the Geneva Protocol, a treaty banning the use of chemical and biological weapons. Similarly, after the Second World War, the UN Atomic Energy Commission was established to deal with the use of atomic energy for peaceful purposes and to control the spread of nuclear weapons.

In particular, we recommend that the use of neural technology for military purposes be stringently regulated. For obvious reasons, any moratorium should be global and sponsored by a UN-led commission. Although such commissions and similar efforts might not resolve all enhancement issues, they offer the best-available model for publicly acknowledging the need for restraint, and for wide input into the development and implementation of a technology.

Bias.

When scientific or technological decisions are based on a narrow set of systemic, structural or social concepts and norms, the resulting technology can privilege certain groups and harm others. A 2015 study[12] found that postings for jobs displayed to female users by Google's advertising algorithm pay less well than those displayed to men. Similarly, a ProPublica investigation revealed last year that algorithms used by US law-enforcement agencies wrongly predict that black defendants are more likely to reoffend than white defendants with a similar criminal record (go.nature.com/29aznyw). Such biases could become embedded in neural devices. Indeed, researchers who have examined these kinds of cases have shown that defining fairness in a mathematically rigorous manner is very difficult (go.nature.com/2ztfjt9).

Practical steps to counter bias within technologies are already being discussed in industry and academia. Such ongoing public discussions and debate are necessary to shape definitions of problematic biases and, more generally, of normality.

We advocate that countermeasures to combat bias become the norm for machine learning. We also recommend that probable user groups (especially those who are already marginalized) have input into the design of algorithms and devices as another way to ensure that biases are addressed from the first stages of technology development.

RESPONSIBLE NEUROENGINEERING

Underlying many of these recommendations is a call for industry and academic researchers to take on the responsibilities that come with devising devices and systems capable of bringing such change. In doing so, they could draw on frameworks that have already been developed for responsible innovation.

In addition to the guidelines mentioned above, the UK Engineering and Physical Sciences Research Council, for instance, provides a framework to encourage innovators to "anticipate, reflect, engage and act" in ways that "promote... opportunities for science and innovation

that are socially desirable and undertaken in the public interest". Among the various efforts to address this in AI, the IEEE Standards Association created a global ethics initiative in April 2016, with the aim of embedding ethics into the design of processes for all AI and autonomous systems.

History indicates that profit hunting will often trump social responsibility in the corporate world. And even if, at an individual level, most technologists set out to benefit humanity, they can come up against complex ethical dilemmas for which they aren't prepared. We think that mindsets could be altered and the producers of devices better equipped by embedding an ethical code of conduct into industry and academia.

A first step towards this would be to expose engineers, other tech developers and academic-research trainees to ethics as part of their standard training on joining a company or laboratory. Employees could be taught to think more deeply about how to pursue advances and deploy strategies that are likely to contribute constructively to society, rather than to fracture it.

This type of approach would essentially follow that used in medicine. Medical students are taught about patient confidentiality, non-harm and their duties of beneficence and justice, and are required to take the Hippocratic Oath to adhere to the highest standards of the profession.

The possible clinical and societal benefits of neurotechnologies are vast. To reap them, we must guide their development in a way that respects, protects and enables what is best in humanity. ■

References

1. Kay KN, Naselaris T, Prenger RJ & Gallant JL Nature 452, 352–355 (2008). [PubMed: 18322462]
2. Goering S & Yuste R Cell 167, 882–885 (2016). [PubMed: 27814514]
3. Sellers EW, Vaughan TM & Wolpaw JR Amyotrophic Lateral Sclerosis 11, 449–455 (2010). [PubMed: 20583947]
4. Ramakrishnan A et al. Sci. Rep. 5, 10767 (2015). [PubMed: 26158523]
5. Burget F et al. Preprint at http://arxiv.org/abs/1707.06633 (2017).
6. Hubel DH & Wiesel TN J. Physiol. (Lond.) 160, 106–154 (1962). [PubMed: 14449617]
7. Giancardo L, Sánchez-Ferro A, Butterworth I, Mendoza CS & Hooker JM Sci. Rep. 5, 9678 (2015). [PubMed: 25882641]
8. Nieto-Reyes A, Duque R, Montana JL & Lage C Sensors 17, 1679 (2017).
9. Klein E et al. Brain-Computer Interfaces 3, 140–148 (2016).
10. Parens E Shaping Our Selves: On Technology, Flourishing, and a Habit of Thinking (Oxford Univ. Press, 2014).
11. Kitcher P Science in a Democratic Society (Prometheus, 2011).
12. Datta A, Tschantz MC & Datta A Proc. Priv. Enhancing Technol. 2015, 92–112 (2015).

A man with a spinal-cord injury (right) prepares for a virtual cycle race in which competitors steer avatars using brain signals.

After having electrodes implanted in the brain to stimulate neural activity, some people have reported feeling an altered sense of identity.