Create no evil? A call to cross ethical boundaries

Laurens R. Krol[1,*], Thorsten O. Zander[1]

[1]Brandenburg University of Technology, Cottbus-Senftenberg, Germany

[*]Correspondence: krol@b-tu.de

Abstract:

In order to better inform the public of the potential risks involved in neurotechnology, we call upon the community to explicitly demonstrate such risks. By doing so in a carefully controlled environment, we can safely identify the ethical boundaries that neurotechnology could potentially, but should not, cross. This can also reveal the conditions under which society would, or would not, adopt consumer neurotechnology.

Body:

Neurotechnology appears to currently be at the cusp of societal adoption, with increasing commercial interest in direct-to-consumer hard- and software (Ienca et al., 2018) and general popular interest due to widely publicised results (Musk & Neuralink, 2019). Today, applications of neuroergonomics allow brain activity to be analysed in real time in the workplace (Parasuraman & Rizzo, 2007), and similar approaches are being pursued for everyday situations (Zander & Krol, 2017). At the same time, experts know that there are a number of potential ethical, legal, and societal risks associated with this technology. Consumer devices with access to their users' brain activity can engender serious issues relating to privacy (Mecacci & Haselager, 2019), data ownership (Fairclough, 2014), human agency (Haselager, 2013), cognitive liberty (Sententia, 2004), and even fundamental human rights (Ienca & Andorno, 2017).

We believe that, for a successful, enduring, and societally beneficial uptake of neurotechnology by the general population, this population must be able to have an informed discussion about both the possible benefits and the possible risks. There are a number of ways in which we, scientists, researchers, and developers, can help and support this process. In particular, whereas research generally focuses on demonstrating what is or will be possible, we would here like to argue for a change of perspective: we may also deliberately try to demonstrate applications of neurotechnology that *ought not* to be possible.

The field of human-computer interaction has long been aware of the societal and ethical implications of its work, and has called for active engagement in order to tackle perceived societal problems (Hochheiser & Lazar, 2007). One approach we may adopt is value-sensitive design (Friedman et al., 2013). This approach encourages designers to use a tripartite methodology involving conceptual, empirical, and technological investigations. For example, conceptually, we may identify and define specific relevant societal values. This can be done by cooperating with colleagues from social sciences and humanities, and reaching out to communities. Empirically, we may gather data of user behaviour with respect to those values, and/or of how technology has previously influenced or interacted with those values. Technologically, we may then consider how or to what extent neurotechnology can be used to support, or indeed violate, those values.

In fact, we here propose that the community take it one step further, and experimentally demonstrate how, exactly, neurotechnology may violate societal values. Whereas previous neuroethical discussions have largely been about *possible*, future risks, the existence of a demonstrator would ideally remove any

doubt that the risk is not merely theoretical. This will also make it easier to communicate specific findings, and make the issue more tangible to the public. For example, what better way to demonstrate the risks of cognitive probing (Krol et al., 2020) than by building a demonstrator that does the exact thing a user *wouldn't* want it to do? Similarly, our earlier demonstration of implicit cursor control (Zander et al., 2016) shows not only how neuroadaptive technology can lead to effortless, goal-oriented interaction, but also how this same technology can be abused by computers to obtain access to our preferences without us being aware of this happening. The public should be explicitly informed about both sides of this coin.

To be clear, we are suggesting to identify and then *explicitly violate* societal values—but in a careful, controlled manner that maximises the resulting information content for public discourse. By experimentally demonstrating relevant risks in isolated, safe environments, we may firstly prevent these risks from reaching the public at large—or at least place the public in a position of being sufficiently informed to accept the risks. Secondly, once any such risks have been established, we will be in a better position to develop safeguards and protections against them.

Such research must of course adhere to strict guidelines. A full ethical review will be required before any data is collected, and participants must, at the very least, be given a full debriefing explaining what they were subjected to, for what reason, and what happens next with their data.

This call to action can also be seen as echoing Haselager et al.'s (2021) plea for an experimental philosophy of neurotechnology. They suggest that neurotechnology can be used to deliberately 'confuse' certain human concepts such as agency, and can thus be used to experimentally investigate the relevant dimensions. This is one approach that could simultaneously be used to establish risks and ethical boundaries.

The recommendations made by the OECD concerning the responsible use of neurotechnology (OECD, 2019) call upon us to 'First and foremost, promote beneficial applications of neurotechnology', and to prevent adverse applications. Our current suggestion may appear contradictory to these recommendations. However, the OECD also calls upon us to anticipate potential misuse, enable societal deliberation, and foster communication that 'avoids hype, overstatement, and unfounded conclusions, both positive and negative'. We believe these latter points are best addressed by creating explicit examples of neurotechnology that can objectively demonstrate what is possible—especially when what is possible may not be desirable.

In short, in this conference contribution, we propose a future direction of research that explicitly tries to 'create evil'. In the laboratory, we can do this in a safe and controlled manner to demonstrate the risks of neurotechnology before they are demonstrated in the wild. This research seeks to inform the public, and as such, serves to prevent neurotechnology that unduly affects societal values. Ultimately, this will allow consumers and markets to identify and weed out applications that are unsafe or undesirable, giving greater chances to those that are not. In this way, we hope neurotechnology will be better guided to truly support and benefit society.

References:

Collingridge, D. (1980). *The social control of technology*. London, UK: Frances Pinter.

Fairclough, S. H. (2014). Physiological data must remain confidential. *Nature, 505*, 263. doi: 10.1038/505263a

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. In N. Doorn, D. Schuurbiers, I. van de Poel, & M. E. Gorman (Eds.), *Early engagement and new technologies: Opening up the laborator*y (pp. 55–95). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-007-7844-3_4

Haselager, P. (2013). Did I do that? Brain-computer interfacing and the sense of agency. *Minds and Machines, 23*(3), 405–418. doi: 10.1007/s11023-012-9298-7

Haselager, P., Mecacci, G., & Wolkenstein, A. (2021). Can BCIs enlighten the concept of agency? A plea for an experimental philosophy of neurotechnology. In O. Friedrich, A. Wolkenstein, C. Bublitz, R. J. Jox, & E. Racine (Eds.), *Clinical neurotechnology meets artificial intelligence: Philosophical, ethical, legal and social implications* (pp. 55–68). Cham, Switzerland: Springer. doi: 10.1007/978-3-030-64590-8_5

Hochheiser, H., & Lazar, J. (2007). HCI and societal issues: A framework for engagement. *International Journal of Human–Computer Interaction, 23*(3), 339–374. doi:10.1080/10447310701702717

Ienca, M., & Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. *Life Sciences, Society and Policy, 13*(1), 5. doi: 10.1186/s40504-017-0050-1

Ienca, M., Haselager, P., & Emanuel, E. J. (2018). Brain leaks and consumer neurotechnology. *Nature Biotechnology, 36*, 805–810. doi: 10.1038/nbt.4240

Krol, L. R., Haselager, P., & Zander, T. O. (2020). Cognitive and affective probing: a tutorial and review of active learning for neuroadaptive technology. *Journal of Neural Engineering, 17*(1), 012001. doi: 10.1088/1741-2552/ab5bb5

Mecacci, G., & Haselager, P. (2019). Identifying criteria for the evaluation of the implications of brain reading for mental privacy. *Science and Engineering Ethics, 25*(2), 443–461. doi: 10.1007/s11948-017-0003-3

Musk, E., & Neuralink. (2019). An integrated brain-machine interface platform with thousands of channels. *bioRxiv*. doi: 10.1101/703801

OECD. (2019). Recommendation of the Council on responsible innovation in neurotechnology. (OECD/LEGAL/0457)

Parasuraman, R., & Rizzo, M. (2007). *Neuroergonomics: The brain at work*. Oxford, UK: Oxford University Press.

Sententia, W. (2004). Neuroethical considerations: Cognitive liberty and converging technologies for improving human cognition. *Annals of the New York Academy of Sciences, 1013*(1), 221–228. doi: 10.1196/annals.1305.014

Zander, T. O., & Krol, L. R. (2017). Team PhyPA: Brain-computer interfacing for everyday human-computer interaction. *Periodica Polytechnica Electrical Engineering and Computer Science, 61*(2), 209–216. doi: 10.3311/PPee.10435

Zander, T. O., Krol, L. R., Birbaumer, N. P., & Gramann, K. (2016). Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity. *Proceedings of the National Academy of Sciences, 113*(52), 14898–14903. doi: 10.1073/pnas.1605155114