In Transparency we Trust: Challenges for Applied Machine Learning in Neuroscience

Georg Starke (University of Basel), Eva de Clercq (University of Basel), Bernice Elger (University of Basel; University Center of Legal Medicine of Geneva and Lausanne, Geneva)

Machine learning (ML) constitutes the backbone of many applications of Artificial Intelligence in neuroscience, promising wide-ranging improvements of clinical care [1, 2]. To achieve ethically sound and trustworthy applications, transparency is supposed to be key [3]. Unfortunately, the very design of many ML applications can preclude transparent explanations of its inner workings [4], creating gaps in responsibility for potential short fallings [5], which in turn are thought to be inimical to trust. Does trust in neurotechnological ML hence require complete transparency? By drawing on the framework of Onora O'Neill, we will argue against over-emphasizing transparency. In her Reith Lectures, O'Neill challenges transparency as prime ideal of the information age and demonstrates that the weight given to this particular ideal has marginalized other, more basic obligations [6]. As O'Neill highlights, full and transparent disclosure of information may improve the trustworthiness of agents and institutions, but an obsession with transparency can also undermine trust. This may happen by information dumping, i.e. by disclosing heaps of unsorted or misleading data, confusing the addressee and obscuring crucial information. According to O'Neill, and recently re-iterated by Stephen John [3], transparency is thus not always a value in itself, and fostering trust requires addressing the deeper roots of widespread societal distrust. O'Neill's solution to the dilemma that transparency ought to increase trust but ends up undermining it is straightforward: put less emphasis on transparency and focus more on other conditions of trustworthiness such as the absence of willful deception, the capability to perform the entrusted task and an orientation on the trustee's interest [7-9]. With regard to ML in neuroscience, measures to foster trust could e.g. take the form of clinical tests for ML algorithms comparable to drug testing, focusing on efficacy and tackling conflicts of interests of the developers [10].

References

[1] Bzdok, D. and A. Meyer-Lindenberg, Machine Learning for Precision Psychiatry: Opportunities and Challenges. Biol Psychiatry Cogn Neurosci Neuroimaging, 2018. 3(3): p. 223-230.

[2] Janssen, R.J., J. Mourao-Miranda, and H.G. Schnack, Making Individual Prognoses in Psychiatry Using Neuroimaging and Machine Learning. Biol Psychiatry Cogn Neurosci Neuroimaging, 2018. 3(9): p. 798-808.

[3] John, S., Epistemic trust and the ethics of science communication: against transparency, openness, sincerity and honesty. Social Epistemology, 2018. 32(2): p. 75-87.

[4] Vayena, E., A. Blasimme, and I.G. Cohen, Machine learning in medicine: Addressing ethical challenges. PLoS Med, 2018. 15(11): p. e1002689.

[5] Bublitz, C., et al., Legal liabilities of BCI-users: Responsibility gaps at the intersection of mind and machine? Int J Law Psychiatry, 2018.

[6] O'Neill, O., A Question of Trust. The BBC Reith Lectures 2002. 2002, Cambridge: Cambridge University Press.

[7] O'Neil, C., Lying, Trust, and Gratitude. Philosophy & Public Affairs, 2012. 40(4): p. 301-333.

[8] Hardin, R., Gaming trust. Trust and Reciprocity, 2003. 6: p. 80-101.

[9] O'Neill, O., Autonomy and trust in bioethics. Gifford lectures. 2002, Cambridge ; New York: Cambridge University Press. xi, 213 p.

[10] Paulus, M.P., Q.J. Huys, and T.V. Maia, A Roadmap for the Development of Applied Computational Psychiatry. Biol Psychiatry Cogn Neurosci Neuroimaging, 2016. 1(5): p. 386-392.