Misplaced Trust and Distrust.
How Not to Engage with AI in Clinical Neuroscience

Georg Starke[1], Marcello Ienca[1]

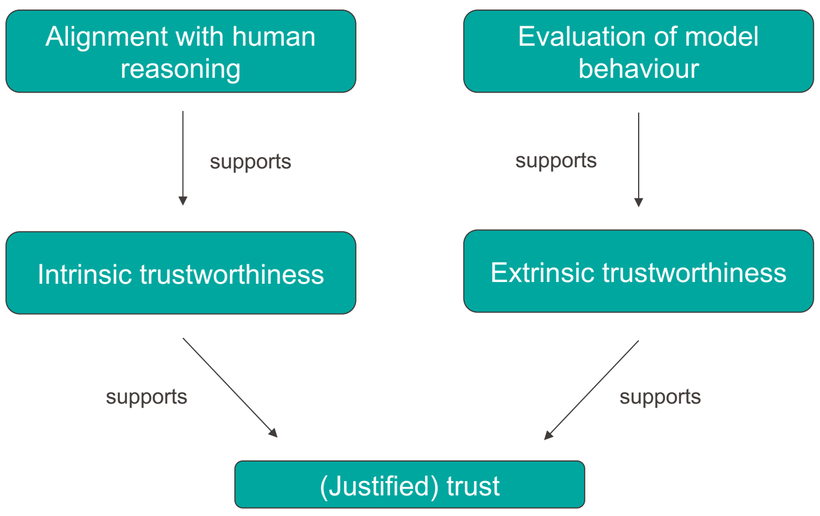
[1] École Polytechnique Fédérale de Lausanne, College of Humanities

Motivation

- There is a proliferation of different theories of trust in AI [e.g., 1-4].
- We propose to take the opposite approach and look at failed forms of trust and distrust.
- We suggest a general taxonomy when trust or distrust are ethically and epistemically justified.

Trust and Trustworthiness

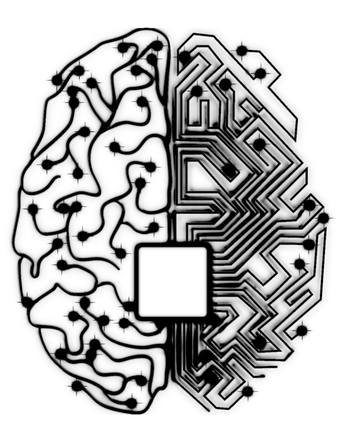
- We follow a minimal definition of trust as giving discretionary authority to an AI with view to a specific task [5].
- The focus of ethical analysis should be trustworthiness [6] as well as the respective reasons for trust and distrust.
- Trust can be intrinsic and extrinsic [4], and so can trustworthiness.



AI in Clinical Neuroscience

- Potential applications for AI in neuroscience are wide-ranging, from prediction to diagnosis, from brain modeling to assistive neurotechnologies [7].
- Frequently, these applications rely on epistemically opaque forms of AI such as convoluted neural nets (CNNs).
- For our purposes, we focus on one particular type of application: an implanted, AI-based and brain-responsive (i.e., closed-loop) neuromodulator for the treatment of idiopathic generalized epilepsy [8].

- While we believe that the increasing integration of human and artificial intelligence brings about what we call *hybrid minds* [9], we still distinguish between trustors (neuromodulator users), trustees (epilepsy patients) and the entrusted task (epilepsy prediction).



Trust and Distrust in AI in Clinical Neuroscience

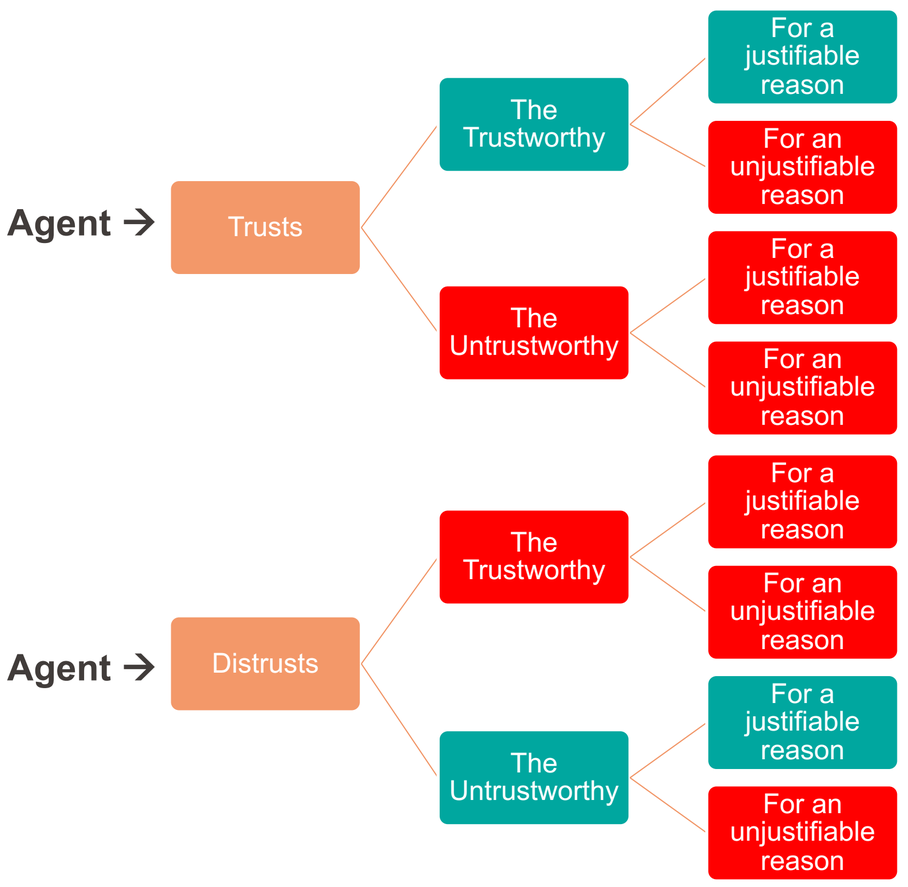
8 hypothetical cases

A: Trust

1. Trust based on assessment of trustworthiness e.g. past performance history, validation, robustness, and alignment with expert knowledge. Approval by FDA/EMA.
2. Trust based on aggressive online marketing. Problematic since reasoning may as well extend to untrustworthy technologies.
3. Trust in generally unreliable, untested device because it happens to work in one specific instance.
4. Trust in unreliable, untested device because it was promoted by an influencer.

B: Distrust

1. Distrust (i.e., not use the device) in a particular instance despite properties of case A1, due to noting a potential damage.
2. Distrust rooted in a user's antisemitism, because the manufacturer's CEO happents to be Jewish.
3. Distrust a device because its AI-component has only been trained on a non-comparable population.
4. Distrust B3 device based on a conspiracy theory, e.g., that it was developed to secretely manipulate users' thoughts.



Implications

By employing the "tripartite analysis of knowledge" to trust in AI, our work highlights the necessity to consider both normative as well as epistemological conditions of trust and distrust in clinical AI. Our research promotes a finer-grained analysis of the dangers involved in specific forms of trust and distrust, enables a better connection of debates in bioethics and empirical human-machine interaction studies, and counters dangers of ethics washing, by focusing on failed forms of trust and distrust.

Acknowledgments

This work has been supported by the ERA-NET NEURON project HYBRIDMIND (Swiss National Science Foundation 32NE30_199436).

Contact

Georg Starke, MD PhD, georg.starke@epfl.ch
Marcello Ienca, PhD, marcello.ienca@epfl.ch

References

Starke, G., & Ienca, M. (2022). Misplaced Trust and Distrust: How Not to Engage with Medical Artificial Intelligence. *Cambridge Quarterly of Healthcare Ethics,* 1-10. doi:10.1017/S0963180122000445

1. Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329-335.
2. Ferrario, A., Loi, M., & Viganò, E. (2020). In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy & Technology*, *33*(3), 523-539.
3. Starke, G., van den Brule, R., Elger, B. S., & Haselager, P. (2022). Intentional machines: A defence of trust in medical artificial intelligence. *Bioethics*, *36*(2), 154-161.
4. Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021, March). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 624-635).
5. Nickel, P. J. (2022). Trust in medical artificial intelligence: a discretionary account. *Ethics and Information Technology*, *24*(1), 1-10.
6. O'Neill O. (2013). Trust before trustworthiness? In: Archard D, Deveaux M, Manson NC, Weinstock D, eds. *Reading Onora O'Neill*. Oxford: Routledge; 2013:237-8.
7. Ienca, M., & Ignatiadis, K. (2020). Artificial intelligence in clinical neuroscience: methodological and ethical challenges. *AJOB neuroscience*, *11*(2), 77-87.
8. Nair, D. R., Laxer, K. D., Weber, P. B., Murro, A. M., Park, Y. D., Barkley, G. L., ... & Morrell, M. J. (2020). Nine-year prospective efficacy and safety of brain-responsive neurostimulation for focal epilepsy. *Neurology*, *95*(9), e1244-e1256.
9. Soekadar, S., Chandler, J., Ienca, M., & Bublitz, C. (2021). On the verge of the hybrid mind. *Morals & Machines*, *1*(1), 30-43.