Towards Trusted AI (TAI): A Phonetic Perspective

Pawel Polak[1], Roman Krzanowski[2]*

As we hand over more aspects of our agency to AI Systems and AI Agents, the question of the trust we have in these systems becomes more pressing. The current implementations for trusted AI (TAI), such as the IBM TAI framework and similar systems, are framed as a multi-dimensional space that encompasses concepts of fairness, robustness, explainability, transparency and accountability, and value alignment (Trusting AI 2021, Shahrdar et al. 2018, Cohen et al. 2019, EPRS 2020, Thiebes et al. 2020, McLeod 2020, Serafimova 2020). TAI is aimed at industrial, technical systems, so in these systems, the conceptual framework is interpreted as practical, operational, and technological requirements without any philosophical depth (i.e., they are "thin" concepts) (Williams 1985, Kirchin 2013, Pekka 2021). As thin concepts, these systems will be open to mistrust by the general public, because they do not carry any moral depth or admit any moral liability.

In social robotics, agents that need to be trusted, or trusted agents, should possess the quality of practical judgment (e.g., Coeckelbergh 2009, Leite et al. 2013, Campa 2016, Polak & Krzanowski 2020). The capacity for **practical judgment** in TAI systems is at most, if at all, implicit in requirements like value alignment and benevolence, yet it is not realized as a standalone property of such systems. Sound practical judgment for autonomous robots (a-robots) can be conceptualized as Aristotelian phronesis, however (Polak & Krzanowski 2020, Krzanowski & Polak 2021).

So, what is phronesis? It is not an exact science (like *episteme* in the Aristotelian sense) or art (i.e., the *techne*/craft or practical skill required to do something or produce something). Unlike the exact sciences, it is not based on ultimate (necessary) principles. The general principles of an exact science are "absolute," because they express abstract, nominal truths, such as in mathematics or pure logic. Ethics, however, deal with "enmattered things" and the changeability and variability that are inherent in concrete embodied facts. The objective of *phrones* is *eudaimonia*, which is the realization of a specific concept of good for a person (i.e., an actor). The focus of *phronesis* is the variability of a specific case from ultimate principles rather than the "generality" of the case. Phronesis cannot be taught like mathematics because there are no rules to teach in phronesis. Phronetic expertise can therefore only be gained through experience (Reeve 1992, Randall 1965). Phronetic TAI systems would need to self-improve their ethical capacities by learning from their own, and others', responses to situations. This means that rather than programming these systems with pre-defined responses to any possible situation, which is clearly an impossible task, these systems should have the ability to perfect their ethical skills independently.

Such a self-improvement capacity should also be regarded as a critical aspect of autonomy. In other words, the autonomy of an a-robot should not be restricted to the ability to act without human supervision but also extend to the ability to independently improve its capacity to act autonomously (see e.g., Polak & Krzanowski 2020). Now, this is what phronesis is not: It is not reliability as understood as something that can be predicted by an algorithm, nor does it provide the transparency required for explainable AI (XAI) (e.g., DARPA-BAA-16-532016, Miller 2017). Moreover, robustness in phronesis is not a strict adherence to requirements, as is the case in industrial TAI, but rather an adherence to a moral code or ethical principles; these are all "soft" concepts that are not easily formalized.

Phronesis, or Aristotelian phronesis to be precise, is a thick concept (Williams 1985, Kirchin 2013, Pekka 2021), and a computer-based implementation will probably not be easy. Despite these expected technical limitations, however, it seems that phronesis even if only as a guiding concept, may have a place in TAI systems, particularly for autonomous robots that perform social functions or exist in complex social situations where proper, moral judgment is expected from any agent (synthetic or human) and may even be vital (e.g., Coeckelbergh 2009, Leite et al. 2013, Campa 2016). The concept of phronesis as eudemonia means that the realization of a specific concept of good may facilitate the acceptance of AI systems in social settings. In fact, it is more critical for trusted AI systems to justify why they did something (i.e., justification) rather than be able to explain what that something was (i.e., formal reckoning), as postulated in XAI. Such a justification would require a phronetic "attitude" (i.e., a deeper, intuitive, contextual, understanding of the decision space) or "engagement with the world" (Smith 2019).

[1] The Pontifical University of John Paul II. Cracow. Poland.
[2] The Pontifical University of John Paul II. Cracow. Poland, rmkrzan@gmail.com, corresponding author.

Selected References

Campa, R. 2016. The Rise of Social Robots: A Review of the Recent Literature. *Journal of Evolution and Technology* - Vol. 26 Issue 1 – February 2016 - pgs 106-113.

Coeckelbergh, M. 2009. Personal Robots, Appearance, and Human Good: A Methodological Reflection on Roboethics. Int J Soc Robot (2009) 1: 217–221. DOI 10.1007/s12369-009-0026-2

DARPA-BAA-16-53. 2016. Explainable Artificial Intelligence. Broad Agency Announcement. Explainable Artificial Intelligence (XAI). August 10, 2016.Arlington, VA.

EPRS. 2020. The ethics of artificial intelligence: Issues and initiatives. Panel for the Future of Science and Technology. European Parliament.

Kirchin, S. 2013. Thick Concepts. Published to Oxford Scholarship Online: May 2013. Accessible at https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199672349.001.0001/acprof-9780199672349-chapter-1.

Krzanowski ,R., P. Polak. 2021. Aristotelian ethics in social robotics: Phronetic Robotics. ICAART 2021,4-6 February, 2021.

Leite, I. , C. Martinho, and A. Paiva. 2013. Social Robots for Long-Term Interaction: A Survey**.** Int J Soc Robot.

McLeod, C. 2020. "Trust", *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2020/entries/trust/>.

Miller, T. 2017. Explanation in Articial Intelligence: Insights from the Social SciencesarXiv:1706.07269v1 [cs.AI] 22 Jun 2017.

Polak, P. and R. Krzanowski. 2020. Phronetic Ethics ain Social Robotics. A new Approach to building ethical robots. Studies and Grammar and Rethotic. 63 (76) 2020

Randall, J. H. (1965). Aristotle (4th ed.). New York: Columbia University Press.

Reeve, C. D. C. (1992). Practices of Reason: Aristotle's Nicomachean Ethics. Oxford: Clarendon Press.

Serafimova, S. 2020. Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement. Humanities and Social Sciences Communications. 7:119 | https://doi.org/10.1057/s41599-020-00614-8

Trusting AI. 2021. Trusting AI. IBM WEB page. Available at https://www.research.ibm.com/artificial-intelligence/trusted-ai/#about-us. Accessed on 03..31. 2021.

Thiebes, S. S. Lins, and A. Sunyaev. 2020. Trustworthy artificial intelligence. Electronic Markets https://doi.org/10.1007/s12525-020-00441-4

Shahrdar, S. L. Menezes, and M. Nojoumian. 2019. A Survey on Trust in Autonomous Systems: Proceedings of the 2018 Computing Conference, Volume 2. 10.1007/978-3-030-01177-2_27.

Pekka, V. "Thick Ethical Concepts", *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2021/entries/thick-ethical-concepts/>.

Smith, B. C. 2019. The Promise of Artificial Intelligence. Reckoning and Judgment. The MIT Press, Cambridge.

Williams, B. 1985. Ethics and the Limits of Philosophy. Harvard University Press, Cambridge, Mass.