



Is trust between AI institutions and the public “morally rotten?”

Title	Is trust between AI institutions and the public “morally rotten?”
Author(s)	Carter, Sarah
Publication Date	2020
Publisher	Machine Ethics Research Group, School of Computer Science, University College Dublin
Repository DOI	10.5281/zenodo.3938851

Is Trust between AI Institutions and the Public “Morally Rotten?”

Developing artificial Intelligence (AI) technology has become a business of power. AI innovation is increasingly centralized in a few large companies – mainly, Google, Facebook, and Apple.¹ Specialized data scientists - the backbone of these institutions - understand how AI functions, further creating a power dynamic between the layperson and the corporatized specialist. In the age of COVID-19, we have all become more reliant on technology companies and innovations to fulfill the needs of our new digital lives. We have little choice but to trust them in developing the AI technologies of the future.

Whether these companies are deserving of our trust, however, is another discussion. “How we understand the nature of trust,” writes Magrit Sutrop in her paper *Should we trust artificial intelligence*, “will make a difference in what we say about the conditions in which trust is justified” (pg. 500).² Selecting Annette Baier’s understanding of trust as a lens, I explore the question – is trusting corporations to develop AI morally justifiable?

Before answering this, we must first describe our lens-of-choice in more detail. Annette Baier was a feminist theorist who challenged the traditional views of philosophers by incorporating the role of power into discussions of trust. In her primary work, *Trust and Anti-Trust*, she describes the previous work on trust by (male) scholars – such as Nietzsche and Kant - and critiques them.³ Her primary critiques were that they focused on trust relationships between those of equal standing that were contractarian in nature. This kind of trust establishes explicit repercussions for violating the relationship but is limiting – parties in a contract risk less and thereby set to gain less. In addition, this kind of trust ignores the most vulnerable of society and undermines their moral agency. “Contract,” she further writes, “is a device for traders, entrepreneurs, and capitalists, not for children, servants, indentured wives, and slaves” (pg. 247). At their core, arguments of trust put forth by previous philosophers failed to consider the role of power in trust.

To rectify this shortcoming, Baier captures the interplay between power and trust by conceptualizing trust in terms of reliance, vulnerability, and good will. Trust, she argues, involves relying on another party to care for something entrusted to them. This trust, however, is not in and of itself a virtue – a trusting community can be exploited by a powerful few. To tackle this, she further adds a test to determine the *moral decency* of the trust relationship, called the “expressibility test” (pg. 255). Summarized, the expressibility test states that a trust relationship is morally decent if both parties could disclose the reasons for being confident in the trust relationship, without undermining the relationship itself. If one party, for instance, is confident in the trust relationship because they can threaten the other party, then the relationship is morally indecent; in contrast, if both parties have the best interests of the other at heart, the relationship is morally decent. Through defining trust as reliability and capturing the role of good will in her expressibility test, Baier provides a concrete mechanism for assessing the moral decency of trust where one party has power over another.

As mentioned previously, corporations are developing a monopoly on AI technology. When assessing the strength of this trust relationship between the public and increasingly powerful companies developing AI, we require a moral test of trust that adds power into the equation – such as Baier’s expressibility test. With this in hand, we can now ask the question - is trusting corporations to develop AI *morally decent*?

In order to answer this, we must first define the nature of the trust relationship between the public and the institutions developing AI technology – that is, what is the public *entrusting*? Or, from Baier’s point of view, what is it that the public values that has been entrusted to the corporations’ care? Bioethicist David Resnick argues that there are three items that the public trusts to scientific or innovative enterprises: 1.) public resources (such as: materials, funding, education); 2.) to provide expert knowledge and commentary on policy decisions; and 3.) to innovate and benefit society (such as improving engineering, medicine, or agriculture).⁴ AI developers are similarly entrusted – young data scientists are educated in public institutions, with the expectation that they will provide a meaningful contribution to national policy decisions and further innovations that will benefit society. Thus, from a Baier point of view, the public relies on the companies to use the resources entrusted to them for the public benefit – especially during times of crisis, such as the COVID-19 pandemic.

Now, to the expressibility test – what are companies’ reasons for being confident in the trust relationship? While it may not be possible to peer into the boardroom of Google or Facebook, we may be able to glean what conditions they rely on through how they react to being caught *breaking* the trust relationship. Facebook’s Cambridge Analytica scandal, where user data was shared for algorithm-fueled targeted political advertising in the US 2016 election, was a violation of trust in AI-developing institutions to benefit society. While no technical data breach occurred, it was perceived by many users as a breach of trust.⁵ While Facebook has since instituted some changes,⁶ they have been critiqued for being quite minimal – *Surveillance Capitalism* author Shoshana Zuboff argues that Facebook ultimately waited out the user backlash, relying on the public’s “forgetfulness” in lieu of meaningful action.⁷ If this is true, it suggests that Facebook relies on the public’s *forgetfulness of transgressions* to sustain the trust relationship.

Can knowledge of this condition exist in harmony with the public’s reliance on these companies to develop AI technology to benefit society?³ This condition would likely not bode well with the average user, and the knowledge threatens the trust of the public in Facebook AI development and deployment. This therefore fails Baier’s expressibility test, and the trust relationship is morally indecent. On the contrary, not only would this trust relationship be morally indecent, but Baier would view it as “*morally rotten*” – a particularly heinous relationship that relies on secrecy (pg. 255). From the side of the corporations, then, the trust relationship between the public and AI-developing institutions is not morally justifiable.

Baier’s expressibility test, however, doesn’t just consider the perspective of the powerful party, but the moral agency of the other party as well; it stresses *mutual* reliance and *mutual* trust (pg. 259).³ What reasons then, do the public, or the governments that represent them, have to be confident in the trust relationship?

The 2019 *Ethics Guidelines for Trustworthy AI* suggests that the public relies on liability over accountability,⁸ limiting the trust relationship to “contractarian trust” (pg. 251).³ The concept of liability is, of course, related to accountability, but it emphasizes a different aspect – liability emphasizes *what harm* was done to a victim, while accountability emphasizes *who* caused an action to occur.⁹ *Ethics Guidelines for Trustworthy AI* seems to emphasize harm, defining accountability as “including auditability, minimization and reporting of negative impact, trade-offs, and redress” (pg. 14).⁸ In addition, the European Commission’s *Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things, and Robotics* aims to assess how AI technology could fit into existing EU product safety directives and furthers the “redress” described in 2019.¹⁰

This accountability-turned-liability would likely be viewed by Baier as the limited “contractarian” trust, the order of trust emphasized by many (male) theorists and critiqued in her work (pgs. 247-253).³ As mentioned previously, while this kind of trust establishes repercussions for violating the relationship, it comes at a cost – fewer risks are taken yielding fewer benefits. While the audit regimes described in the European Commissions’ *Ethics Guidelines for Trustworthy AI* may be critical from a legal point of view, a part of the trust relationship – to elevate it from “contractarian” to “morally decent” – remains missing.

When considering both parties, the trust relationship between the public and corporations to develop AI is morally unjustifiable because they do not meet the requirements of a morally decent relationship. Using the work of Annette Baier as a lens, I have explored the reasons that institutions developing AI are confident in their trust relationship with the public, arguing that they rely on collective *forgetfulness* and therefore would be morally rotten when subjected to Baier’s expressibility test. On the other side of the relationship, the public and the governments that represent them rely heavily on liability over accountability, thereby substituting a more genuine trust relationship for a contractarian one.

We may, however, be able to further improve the trust relationship by contextualizing these challenges in the expansive literature around AI trust and trustworthiness, outside of the scope of this brief abstract. For example, we could perhaps explore reframing discussions around trustworthy AI as an academic and scientific enterprise, not a corporate one. In addition, further analysis could be conducted to assess to what extent accountability has been surrendered for liability in guidelines and policies. Pending these future explorations, we may be able to promote a trust relationship between AI-developing institutions and the public that Baier herself would deem morally decent.

Literature Cited

1. Asaro P. What is an “Artificial Intelligence Arms Race” Anyway? *ISJLP*. 2019;15:45-64. <https://perma.cc/D37Q-33LB>].
2. Sutrop M. Should we trust artificial intelligence? *Trames*. 2019;23(4):499-522. doi:10.3176/tr.2019.4.07
3. Baier A. Trust and Antitrust. *Ethics*. 1986;96(2):231-260. <https://about.jstor.org/terms>.
4. Resnik DB. Scientific REasearch and Public Trust. *Sci Eng Ethics*. 2011;17(3). doi:10.1007/s11948-010-9210-x
5. Kozłowska I. Facebook and Data Privacy in the Age of Cambridge Analytica. The Henry M. Jackson School of International Studies. <https://jsis.washington.edu/news/facebook-data-privacy-age-cambridge-analytica/>. Published April 30, 2018. Accessed May 12, 2020.
6. Schroepfer M. An Update on Our Plans to Restrict Data Access on Facebook. *Faceb Newsroom*. 2018:1-5. <https://newsroom.fb.com/news/2018/04/restricting-data-access/>.
7. Zuboff S. *The Age of Surveillance Capitalism : The Fight for a Human Future at the New Frontier of Power*. 1st ed. New York: PublicAffairs; 2019.
8. EUROPEAN COMMISSION. High-Level Expert Group on Artificial Intelligence. 2019:2-36. <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top>.
9. Nissenbaum H. Accountability in a computerized society. *Sci Eng Ethics*. 1996;2:25-42.
10. EUROPEAN COMMISSION. *Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things, and Robotics.*; 2020. doi:10.1017/CBO9781107415324.004