A Leap of Faith: Is There a Formula for "Trustworthy" AI?

by-the-organizing-committee-of-the-second-international-summit-on-human-genome-editing.
10. F. Baylis, "Questioning the Proposed Translational Pathway for Germline Gene Editing," *Nature Human Behavior* 3, no. 3 (2019): 200; J. B. Hurlburt, "Human Genome Editing: Ask Whether, Not How," *Nature* 565 (2019): 135; R. Andorno et al., "Geneva Statement on Heritable Human Genome Editing: The Need for Course Correction," *Trends in Biotechnology* 38, no. 4 (2020): 351-54.
11. Baylis, "Questioning the Proposed Translational Pathway," 200.
12. National Academy of Sciences, *Heritable Human Genome Editing*, 9-10.
13. B. Cwik, "Revising, Correcting, and Transferring Genes," *American Journal of Bioethics* 20, no. 8 (2020): 7-18.
14. Andorno et al., "Geneva Statement."
15. J. Halpern and D. Paolo, "Upstream Ethical Mapping of Germline Gene Editing," *American Journal of Bioethics* 20, no. 8 (2020): 1-4; R. Sparrow, "Yesterday's Child: How Gene Editing for Enhancement Will Produce Obsolescence—and Why It Matters," *American Journal of Bioethics* 19, no. 7 (2019): 6-15.
16. Halpern and Paolo, "Upstream Ethical Mapping," 2.
17. Quoted in B. M. Knoppers and E. Kleiderman, "Heritable Genome Editing: Who Speaks for 'Future' Children?," *CRISPR Journal* 2, no. 5 (2019): 285-92, at 288.
18. National Academy of Sciences, *Heritable Human Genome Editing*, 100-108.
19. Lander et al., "Adopt a Moratorium."
20. H. Ledford, "CRISPR Gene Editing in Human Embryos Wreaks Chromosomal Mayhem," *Nature* 583 (2020): 17-18.
21. J. A. Doudna, "The Promise and Challenge of Therapeutic Genome Editing," *Nature* 578 (2020): 229-36.
22. C. E. Dunbar et al., "Gene Therapy Comes of Age," *Science* (2019): doi:10.1126/science.aan4672.
23. Cwik, "Responsible Translational Pathways."
24. Angrist et al., "Reactions."
25. H. T. Greely, "CRISPR'd Babies: Human Germline Genome Editing and the 'He Jiankui Affair,'" *Journal of Law and the Biosciences* 6, no. 1 (2019): 111-83.
26. E. Y. Adashi and I. G. Cohen, "Disruptive Synergy: Melding of Human Genetics and Clinical Assisted Reproduction," *Cell Reports Medicine* 1, no. 6 (2020): 1-5.
27. E. Y. Adashi and I. G. Cohen, "Going Germline: Mitochondrial Replacement as a Guide to Genome Editing," *Cell* 164, no. 5 (2016): 832-35.

A Leap of Faith:
Is There a Formula for "Trustworthy" AI?

by MATTHIAS BRAUN, HANNAH BLEHER, and PATRIK HUMMEL

Artificial intelligence promises to shape and to transform clinical practice and public health.[1] At the same time, AI applications invite reflection on the extent to which they rely upon seemingly opaque black box technologies, reduce opportunities for human assessments and interventions, and automatize situations, decisions, and allocations that can have significant consequences. Doctors' potential reliance on opaque algorithms[2]—which might add to rather than decrease clinician workloads or even render a doctor "handmaiden to an AI" that is actually making the treatment decisions[3]—could affect patients' attitudes. In light of challenges like these, the ideal of trustworthy AI figures prominently in a number of reports, statements, and guidelines concerning the ethical use of AI.[4] For example, in the 2019 guidelines of the High-Level Expert Group on Artificial Intelligence appointed by the European Commission,[5] trustworthy AI is the target notion around which other principles and requirements are centered. The European Commission takes up these principles and requirements in a 2020 white paper, *On Artificial Intelligence: A European Approach to Excellence and Trust*, proposing policy options for the development of an "ecosystem of trust" in which AI is to be pursued and realized.[6]

The High-Level Expert Group argues that trust in AI can be achieved by building on (four) fundamental ethical principles, namely, respect for human autonomy, prevention of harm, fairness, and explicability. With these principles as a foundation, the group's guidelines define seven requirements that have to be met to "promote Trustworthy AI."[7] Key to the ecosystem of trust, as presented in *On Artificial Intelligence,* is a regulatory framework to ensure the security of AI applications and their concordance with fundamental rights and consumer rights and thereby to further promote the acceptance of AI technologies. In addition, the development of this ecosystem of trust as a "policy objective in itself"[8] is intended to provide legal certainty for citizens, companies, and public organizations.

So far, however, both proponents and critics of the notion of trustworthy AI have assumed flawed pictures of the nature of trust. The critics we are concerned with reject

Matthias Braun, Hannah Bleher, and Patrik Hummel, "A Leap of Faith: Is There a Formula for 'Trustworthy' AI?," *Hastings Center Report* 51, no. 3 (2021): 17-22. DOI: 10.1002/hast.1207

the notion of trustworthy AI too quickly, partly because their notion of trust is relatively narrow. The proponents, in turn, proceed as if the right formulaic procedures guarantee trust in AI, or at least render such trust a rational choice. We argue, though, that both trust and reliance retain an ambivalent and precarious status and resist promotion through formulaic approaches.

Trust as an Accelerator to Gain Societal Acceptance?

Empirical evidence indicates what could be described as a form of basic trust in AI.[9] Even though the concrete ends of the development of AI are unforeseeable, there is a sort of enthusiasm that AI will help to address severe societal challenges, for example, in medical care, public health, industry, or transportation. At the same time, people seem to have varying expectations about whether and which organizations develop and manage AI in the best interests of the public. Various surveys further suggest reservations or even fear about AI's potential to increase data-driven surveillance and invasion of privacy.[10] Respondents also expressed concern that "AI interjects greater possibilities for . . . lessening creativity and freedom of thought" and increases loneliness and isolation due to a reduction in direct human-to-human interactions.[11]

Precisely because societal perceptions of AI are ambiguous, it is not surprising that trust in AI-driven public health applications is highlighted as a key condition for efficacy and acceptability.[12] In stakeholder consultations and academic discourse, debates wage about the proper object of trust in this context and also about whether trustworthiness is a fruitful category to shape governance. Some observers argue that trust in AI is "conceptual nonsense. Machines are not trustworthy; only humans can be trustworthy (or untrustworthy)."[13] Authors like Thomas Metzinger, Joanna Bryson,[14] and Mark Ryan[15] maintain that trust can be placed only in entities that can possess emotive states or take responsibility. Conceiving of technical objects or animals as entities that could be trusted would amount to an illegitimate anthropomorphization.

We see two broad strategies of response to such an argument. First, which characteristics an entity must possess in order to be a potential addressee of trust is not, in fact, obvious. At least intuitively, a Bryson-, Metzinger- and Ryan-style claim that trust can be ascribed only by and between human agents seems quite narrow. For example, a blind person who lives with her guide dog and depends on seeing the world and possible obstacles through its eyes seems to grant what appears to be a considerable amount of trust to her dog. Another individual may need an AI-driven prosthesis to walk. Would it be wrong to speak of trust in the dog or the smart prosthesis? Moreover, at least some human institutions—a society's system of justice or its health care system, other subsystems, specific organizations, and broader institutions like science in general, for example—could, in principle, be appropriate objects of trust. The object of trust might seem less tangible than the guide dog or the prosthesis, and one might have general questions about whether trust in institutions reduces to trust in individuals associated with the institution. But at least prima facie, institutions seem to pose a further counterexample against an overly narrow view of which entities we can trust.

Scholars like Mark Coeckelbergh have argued[16] that there might be forms of trust that are *specific* to interactions between human agents but also that there is no reason to limit the notion of trust *in general* as obtaining only between humans. Instead, there can be different types of trust for various forms of interactions and actors. Even if some of them differ from the kind of trust typically obtaining between humans, they are still modes of trust. One way to bolster this suggestion is to consider that attitudes of the grantor rather than features of the recipient are what determines whether trust is present. According to this working hypothesis, one entity trusting another means, first of all, that the former dares to establish a particular kind of relation with the latter.

Second, Metzinger, Ryan, and others likely maintain that all this relation can involve in human-machine interaction is *reliance*. According to this view, a human relying on a machine makes certain assumptions about how the machine behaves but supposedly does not deploy the rich set of normative presuppositions distinctive of trust. Indeed, in everyday language, "trustworthy" is used interchangeably with a range of cognate notions, most notably reliability but also predictability, goodwill, and, to some extent, transparency. We can meaningfully ask what licenses application of these notions to an AI-driven system even if, strictly speaking, trustworthy AI is a misnomer. Skeptics about "trustworthy" AI need not be opposed to this project but may highlight that what we arrive at in the best case is a well-designed and -controlled technical system, nothing capable of humanlike agency and, hence, nothing that could be a potential addressee of trust.

Even if it were true, strictly speaking, that *trustworthy* AI is a misnomer, considering structural analogies between reliability (being an appropriate object of reliance) and trustworthiness would remain worthwhile. With regard to the blind person and her dog, one could argue that the decisive factor is not only whether the dog does its work correctly and effectively. At least as important is the blind person's leap of faith to proceed as if the dog will perform satisfyingly. Even if the dog does its work effectively and without error—is completely reliable—the blind person will not walk forward if she does not grant that the dog will guide her safely. Similar points apply to inanimate objects such as a prosthesis: the respective person must be ready to take a leap of faith with their smart prosthesis. Both trust and reliance are in play only if such leaps of faith are taken.

Thus, there is an analogous structure between reliability and trustworthiness. For someone or something to be considered trustworthy or reliable by a specific person, this per-

Taking the ambivalence of trust seriously invites abandoning the dichotomy of trust as a desirable goal and distrust as something to be avoided.

son has to take a leap of faith toward the respective entity. In the giving of trust or the decision to rely on an entity, a specific relation between the person and the respective entity is initiated. Putting trust in an entity involves an element of risk, the willingness to get involved and commit to something—in other words, taking a leap of faith.

Leaps of Faith as Gifts

What are these leaps of faith that we have just claimed can be directed at both humans and machines? One way to conceive the leap of faith is to construe it as a rational choice. Putting trust in a technology, an institution, or certain processes could be understood as a wager between science, economy, and society. The goal of the wager is to arrive at conditions in which, for those affected, trust or reliance is a good bet. Nevertheless, understanding trust or reliance as flowing from this kind of negotiation process is only one way of making sense of leaps of faith. It is necessary to see that leaps of faith, even as such rational choices, are dependent on conditions that are not clear in advance and are difficult to calculate. To better understand these two entangled dimensions, it is helpful to take a look at what it could mean to *give trust through a leap of faith*. Only by considering the fundamental gift aspects entailed in such a leap can one grasp the modes of giving trust in their entirety.

Gift theorists introduce a helpful distinction between exchanges and gifts.[17] Despite several differences in their arguments, these theorists highlight that gifts—unlike mere exchanges—do not rest on strategic or calculatory considerations. Instead, gifts are given in order to endow something, to lay the basis for a common ground, a bond among agents that is the background condition for acts of exchange. The donor gives something despite uncertainty about the effects of her act; the precarity of being rejected, exploited, or misunderstood; and potential disappointment and even injury. In particular, gifts do not rest solely on a presumption that others fulfill reciprocal obligations or that some strategic calculus such as a tit-for-tat reasoning will be vindicated. Gifts imply an asymmetric constellation in which the act of giving cannot be requested, incentivized, necessitated, or begged for. This is compatible with the idea that once the endowment of gift-giving has been given by the donor(s) and accepted by the recipient(s), acts of negotiation and exchange are necessary to maintain and to keep intact the social space opened up by this endowment. The point is that the initiation does not rest solely on such considerations.

This view on gifts has implications for how to understand, maintain, and reinforce the leaps of faith that are inherent to both trust and reliability. Of course, certain background conditions matter for whether such leaps occur. For example, for societal trust in a new technology, it matters that development processes are open, inclusive, and transparent. But, as the insights from gift theorists indicate, such conditions might not be sufficient. And even when they are, the endowment granted through a leap of faith remains fragile. For example, once violations of the trustor's goodwill are perceived, trust may be withdrawn. This is a live possibility for medical AI, where ambiguous antecedent perceptions could easily tip toward skepticism, disappointment, and resistance unless actual reservations, worries, and potentially destabilizing factors surrounding the technology are taken seriously. In dealing with these ambiguities, trustworthy AI may articulate necessary background conditions. But as shown, even the best concept of trustworthy AI has to reflect the insight that leaps of faith are taken, resist incentivization, and maintain a lingering sense of fragility and provisionality. This being said, once such leaps are taken, they open up or reinforce room for maneuver in social space. AI could, in principle, serve as a particularly good example for this generative potential of such leaps of faith, given the benefits it might provide across several domains of society.

Making AI Trustworthy?

Leaps of faith are inherent to both trust and reliance, and the logic of the gift appears to capture important features of how such leaps of faith are taken. In light of these points, how can AI satisfy conditions of trustworthiness or machine surrogates thereof, both in terms of its design and its implementation? The conditions of trust in digital health systems and what exactly it takes to advance "[t]owards trustable machine learning"[18] remain underexplored and contested.

There is, of course, nothing wrong with the European Commission's High-Level Expert Group's outlined principles on trustworthy AI themselves or their implementation through a regulatory framework. Quite the opposite is true. Strict respect for human dignity; a strong focus on individual acts of (informational) self-determination; avoid-

ance of undue discrimination based on sex, race, or gender; and opportunities to exercise control over data that have or will have an impact on the modes of people's freedom must guide AI development, use, and policy.

However, this is only one perspective on granting trust. A formulaic approach presumes that trusting or relying on AI is straightforwardly operationalizable rather than a fragile, precarious, and uncertain process. Indeed, the recent "Assessment List for Trustworthy AI" of the High-Level Expert Group applies such a formulaic approach.[19] This list, meant to support developers and deployers in implementing the seven key requirements of trustworthy AI, appears to rest on an understanding of trust as something that can be achieved by following procedures. Of course, criteria, parameters, and checklists are *relevant* to the genesis of social trust in a new technology. The High-Level Expert Group's proposed principles serve an important purpose by virtue of promoting certain background conditions of this process. However, it is not enough to leverage certain plausible ethical criteria as accelerators that pave the way for the acceptance of new technologies.

Three reasons for caution in particular are conceivable. To begin with, any checklist for addressing normative questions will be only as helpful as the normative reference points at hand for addressing them.[20] Requirements such as transparency, well-being, nondiscrimination, and fairness are in need of substantive accounts of how to make them explicit and how to adjudicate between mutually incompatible readings. Second, even if political and societal stakeholders have reasonably clear ideas in mind about what such requirements entail, there remains a need to deliberate on how to implement them in a given case. Finally, even if the key parties have agreed on and implemented a set of enforceable conditions, it would be illusory to consider these alone as a guarantee for social trust in AI or any other new technology. Indeed, even with all these principles in place and criteria met and an ensuing classification of an AI system as trustworthy in this sense, one hardly arrives at a guarantee that anyone will take a leap of faith. Whether trust will emerge and endure remains open.

Seeking to establish trust in this way also threatens to ignore the significance of *distrust*, which is relevant to trust in various and intricate ways. For example, Katherine Hawley maintains that trust can be fully understood only if we also attend to the conditions of distrust.[21] Besides conceptual considerations like these, distrust has important practical functions. Pierre Rosanvallon argues that, in democratic societies, distrust forces everyone to examine whose interests are represented, which goals are to be achieved in which way, and last but not least, who bears the burdens and who benefits from new technologies.[22] In this way, distrust plays an indispensable, constructive role in politics, society, and technology. To stabilize and ameliorate arrangements within and across these spheres, assumptions and prerequisites of trust must constantly be queried and scrutinized. Possibilities to voice distrust open up and advance societal negotiations and deliberative processes and maintain the leaps of faith that pave the way for trust, which, as we have argued, remains ambivalent and fragile.

Handle with Care: On Dealing with Remaining Fragilities

Leaps of faith inherent to attitudes of trust and reliance cannot be guaranteed or crafted easily. At least some of the rhetoric of the European Commission's white paper *On Artificial Intelligence* appears to gesture at this complication: referring to an entire, dynamic ecosystem widens the scope beyond the High-Level Expert Group's narrower focus in which the technology itself appears to be the locus and determinant of trustworthiness. In substance, however, the approach laid out in *On Artificial Intelligence* remains just as formulaic, seeking to supplement the expert group's "key requirements" with regulation that could "build trust."[23] Several further aspects of the nature of trust must be considered.

First, even if one subscribed to the goals of this formulaic approach to the desired ecosystem, the approach would remain incomplete. Trust presupposes the trustee's receptiveness to the attitudes of the trustor. If there is one common lesson from topics such as nuclear power, genetically modified crops, and geoengineering, it is that new technologies and strategies for navigating uncertainties about their safety and about conceptualizing their risk-benefit ratios and assessing their acceptability require careful attention to the narratives, perceptions, and attitudes of those affected.[24] It is thus somewhat surprising that, despite the high number of reports and guidelines on the ethics of AI that have recently been published,[25] there is a notable lack of calls for—and provision of—empirical evidence that explores stakeholder attitudes and expectations. Consider, as one example, black box issues in medical AI related to how and why an AI arrived at a given output. If neither clinicians, public health officials, nor system designers can answer such questions, why should patients and populations trust medical AI? One underexplored question is what the notion of explainability[26] could mean from the patient perspective, that is, what counts as a satisfying explanation of AI-driven recommendations and predictions.

Second, the focus of the formulaic approach—to generate trust in AI—requires a caveat. Given the diversity in attitudes and perspectives across societal stakeholders, some expectations will inevitably be disappointed. But preventing misplaced trust is not only difficult to achieve; it also threatens to neglect the fruitful role that *distrust* plays in social space. Taking the ambivalence of trust seriously invites abandoning the dichotomy of trust as a desirable goal and distrust as something to be avoided. Surely, AI developers should not build AI that is for some reason unworthy of trust—because of bias or a failure to track relevant facts, for instance. Then again, trying to convince every skeptic for good would be misguided: specifically, distrust motivates

and enables individuals to forgo technology, to constrain its power, and to exercise meaningful human control.[27] Enabling individuals to express attitudes of distrust will be a necessary (although insufficient) condition for trust.

Third, *controllability* will be key to giving room for both trust and distrust. At the individual level, controllability requires mechanisms for choosing whether to benefit from, but also to forego and to withhold data from, medical AI. At the societal level, controllability invites participative and ongoing debates in which the public shapes whether, how, and toward which ends AI is put to work.

Fourth, a formulaic approach presuming that trust can be ensured through a regulatory framework or straightforwardly crafted by reference to formulaic criteria also appears to ignore the role of participatory and deliberative processes for granting trust. The idea that a rigid set of principles and regulation will suffice to govern AI threatens to be an oversimplification. The implementation of requirements such as transparency or fairness and their respective configurations in the specific contexts of AI systems can succeed only on the basis of discursive, context-sensitive, and participatory processes. To move beyond the formulaic approach, societies therefore need to strive for open, inclusive, and receptive spaces for consultation and negotiation, participation, and contextualization at the interfaces of the political, economic, juridical, scientific, and societal spheres. The very point of such spaces is to allow narratives, background conditions, aims, and the specific use of AI to be debated in the light of stakeholder expectations in an open forum where distrust and dissent but also approval and endorsements can be articulated. Such spaces are indispensable to get clearer not only on which principles shall guide but also on how exactly their content is to be understood and how tensions and conflicts among the principles could be resolved in practice.[28]

While the High-Level Expert Group has started a feedback loop on its guidelines in a piloting process with more than 350 stakeholders (the result of which is the assessment list mentioned above), participation should not be limited to this: continuous, deliberative negotiation processes of society, science, politics, and economy are a prerequisite for a comprehensive perspective on whether and how to take leaps of faith. Even if a societal consensus on AI were reached, this would not represent a safe common ground but continue to imply a lingering sense of provisionality and incompleteness, as dynamics of renegotiation can always reemerge.

Finally, trust is not fully captured as a mere wager, and thus designing an ecosystem in which certain expectations and commitments are a good bet is worthwhile, but it is not guaranteed to generate trust. Even the most advanced principles and regulatory frameworks cannot necessitate the leaps of faith on which trust and reliance are based.

Although some believe that trustworthy AI is conceptual nonsense, we have explored what it would take to make sense of this notion. One must first recognize that trust cannot simply be ensured or guaranteed. Even if an AI system is presented as particularly trustworthy by certain institutions, companies, or even individuals, the question whether individuals or groups are willing to take the plunge and grant their trust remains open. It is an interplay between trust givers and receivers, a relational mode of interaction, that can, in principle, establish or enhance room for maneuver. Where trust is given, a dynamic of interaction is created that remains fragile. Throughout this process, the constitutive meaning of distrust is vital: distrust in this sense is not simply the expression of disappointed or broken trust. Instead, in social negotiation processes, distrust also marks the possibility to problematize which claims and rights are heard and recognized in the endowed social fabric. In this sense, distrust has a fruitful function in that it helps to indicate emerging shifts in power, justification, or representation.[29]

We thus offer the following suggestions on how to deal with the remaining fragilities of trust:

- be receptive and systematically explore stakeholder attitudes;
- be frank and emphasize that some expectations will be disappointed and that distrust is fruitful and need not be avoided;
- be provisional and stress deliberative and substantially open procedures, not rigid principles;
- be empowering and equip individuals with means and mechanisms for controllability;
- be humble and transparent about what exactly societies arrive at with these measures in place.

As should be clear, we do not understand these suggestions as sufficient conditions for arriving at trustworthy AI. But they are necessary for paving the way toward a slightly different, worthwhile target: not a mere acceleration of, or even a guarantee for, societal acceptance, but the provision of a pathway and a societal ground for responsible and inclusive development and deployment of AI.

Acknowledgments and Disclaimer

We are grateful to Sandra Fernau, Tabea Ott, Hannah Schickl, Stefanie Siewert, Max Tretter, and the anonymous reviewers for their helpful comments and criticism on earlier versions of this essay.

Open-access funding is enabled and organized by Projekt DEAL. This work is part of the research project SMART Start, which is funded by the German Ministry of Health (through grant ZMVI1-2519DAT400). The work was also supported by the Friedrich-Alexander-Universität Erlangen-Nürnberg Emerging Talents Initiative. The funders had no role in the study design, decision to publish, or preparation of the manuscript.

1. E. J. Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* (New York: Basic Books, 2019).

2. D. S. Watson et al., "Clinical Applications of Machine Learning Algorithms: Beyond the Black Box," *BMJ* 364 (2019): doi:10.1136/bmj.l886.

3. R. Sparrow and J. Hatherley, "High Hopes for 'Deep Medicine'? AI, Economics, and the Future of Care," *Hastings Center Report* 50, no. 1 (2020): 14-17.

4. L. Floridi et al., "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines* 28, no. 4 (2018): 689-707; The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*, 1st ed. (IEEE, 2019), at https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

5. High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* (Brussels, European Commission, 2019).

6. European Commission and Directorate-General for Communications Networks, Content and Technology, *On Artificial Intelligence—a European Approach to Excellence and Trust* (white paper) (Brussels: European Commission, 2020), 3, 9-25.

7. High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, 2.

8. European Commission and Directorate-General for Communications Networks, Content and Technology, *On Artificial Intelligence*, 3.

9. European Commission, Directorate-General for Communications Networks, Content and Technology, with coordination by the Directorate-General for Communication, *Attitudes towards the Impact of Digitisation and Automation on Daily Life* (Brussels: European Commission, 2017), doi:10.2759/835661; B. Zhang and A. Dafoe, *Artificial Intelligence: American Attitudes and Trends* (Oxford: Center for the Governance of AI, Future of Humanity Institute, University of Oxford, 2019), https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/us_public_opinion_report_jan_2019.pdf.

10. Pega, *What Consumers Really Think about AI: A Global Study. Insights into the Minds of Consumers to Help Businesses Reshape Their Customer Engagement Strategies* (Cambridge, MA: Pega, 2017), https://www.ciosummits.com/what-consumers-really-think-about-ai.pdf.

11. Edelman, "2019 Edelman AI Survey: Survey of Technology Executives and the General Population Shows Excitement and Curiosity Yet Uncertainty and Worries that Artificial Intelligence Could Be a Tool of Division," March 2019, https://www.edelman.com/sites/g/files/aatuss191/files/2019-03/2019_Edelman_AI_Survey_Whitepaper.pdf, p. 19.

12. World Health Organization, *Big Data and Artificial Intelligence for Achieving Universal Health Coverage: An International Consultation on Ethics* (Geneva: WHO, 2018).

13. T. Metzinger, *Ethics Washing Made in Europe* (Der Tagesspiegel, 2019), https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html.

14. J. Bryson, "AI & Global Governance: No One Should Trust AI," November 13, 2018, United Nations University, Centre for Policy Research, https://cpr.unu.edu/ai-global-governance-no-one-should-trust-ai.html.

15. M. Ryan, "In AI We Trust: Ethics, Artificial Intelligence, and Reliability," *Science and Engineering Ethics* 26 (2020): 2749-67.

16. M. Coeckelbergh, "Can We Trust Robots?," *Ethics and Information Technology* 14, no. 1 (2012): 53-60; M. Coeckelbergh, *AI Ethics* (Cambridge, MA: MIT Press, 2020).

17. J. Derrida, *Given Time: I. Counterfeit Money* (Chicago: University of Chicago Press, 1992), 204; P. Ricœur, *The Course of Recognition* (Cambridge, MA: Harvard University Press, 2005), 324; M. Hénaff, "Ceremonial Gift-Giving: The Lessons of Anthropology from Mauss and Beyond," in *The Gift in Antiquity*, ed. M. L. Satlow (Chichester, UK: Wiley-Blackwell, 2013), 12-24.

18. Nature Biomedical Engineering, "Towards Trustable Machine Learning," *Nature Biomedical Engineering* 2, no. 9 (2018): 709-10.

19. High-Level Expert Group on Artificial Intelligence, "The Assessment List for Trustworthy AI (ALTAI) for Self-Assessment," European Commission, 2020, https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

20. M. Braun and P. Hummel, "Contact-Tracing Apps: Contested Answers to Ethical Questions," *Nature* 583 (2020): 29-31.

21. K. Hawley, *How to Be Trustworthy* (Oxford: Oxford University Press, 2019).

22. P. Rosanvallon, *Die gute Regierung* (Berlin: Suhrkamp, 2018).

23. European Commission and Directorate-General for Communications Networks, Content and Technology, *On Artificial Intelligence*, 2, 3, 8.

24. S. Cave et al., *Portrayals and Perceptions of AI and Why They Matter* (London: Royal Society, 2018).

25. Floridi et al., "AI4People"; IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design*; High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*.

26. Floridi et al., "AI4People"; High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*.

27. F. Santoni de Sio and J. Van den Hoven, "Meaningful Human Control over Autonomous Systems: A Philosophical Account," *Frontiers in Robotics and AI* 5 (2018): doi:10.3389/frobt.2018.00015.

28. J. Whittlestone et al., "The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions," in *Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society* (New York: AAAI and ACM Digital Libraries, 2019).

29. P. Kalluri, "Don't Ask if Artificial Intelligence Is Good or Fair, Ask How It Shifts Power," *Nature* 583 (2020): doi:10.1038/d41586-020-02003-2.