Distrust of Artificial Intelligence: Sources & Responses from Computer Science & Law

Cynthia Dwork & Martha Minow

Social distrust of AI stems in part from incomplete and faulty data sources, inappropriate redeployment of data, and frequently exposed errors that reflect and amplify existing social cleavages and failures, such as racial and gender biases. Other sources of distrust include the lack of "ground truth" against which to measure the results of learned algorithms, divergence of interests between those affected and those designing the tools, invasion of individual privacy, and the inapplicability of measures such as transparency and participation that build trust in other institutions. Needed steps to increase trust in AI systems include involvement of broader and diverse stakeholders in decisions around selection of uses, data, and predictors; investment in methods of recourse for errors and bias commensurate with the risks of errors and bias; and regulation prompting competition for trust.

Works of imagination, from *Frankenstein* (1818) to the film *2001: A Space Odyssey* (1968) and the *Matrix* series (1999–2021), explore fears that human-created artificial intelligences threaten human beings due to amoral logic, malfunctioning, or the capacity to dominate.[1] As computer science expands from human-designed programs spelling out each step of reasoning to programs that automatically learn from historical data, infer outcomes for individuals not yet seen, and influence practices in core areas of society – including health care, education, transportation, finance, social media, retail consumer businesses, and legal and social welfare bureaucracies – journalistic and scholarly accounts have raised questions about reliability and fairness.[2] Incomplete and faulty data sources, inappropriate redeployment of data, and frequently exposed errors amplify existing social dominance and cleavages. Add mission creep – like the use of digital tools intended to identify detainees needing extra supports upon release to instead determine release decisions – and it is no wonder that big data and algorithmic tools trigger concerns over loss of control and spur decay in social trust essential for democratic governance and workable relationships in general.[3]

https://doi.org/10.1162/DAED_a_01918

Failures to name and comprehend the basic terms and processes of AI add to specific sources of distrust. Examining those sources, this essay ends with potential steps forward, anticipating both short-term and longer-term challenges.

Artificial intelligence signifies a variety of technologies and tools that can solve tasks requiring "perception, cognition, planning, learning, communication, and physical actions," often learning and acting without oversight by their human creators or other people.[4] These technologies are already much used to distribute goods and benefits by governments, private companies, and other private actors.

Trust means belief in the reliability or truth of a person or thing.[5] Associated with comfort, security, and confidence, its absence infers doubt about the reliability or truthfulness of a person or thing. That doubt generates anxieties, alters behaviors, and undermines cooperation needed for private and public action. *Distrust* is corrosive.

Distrust is manifested in growing calls for regulation, the emergence of watchdog and lobbying groups, and the explicit recognition of new risks requiring monitoring by corporate audit committees and accounting firms.[6] Critics and advocates alike acknowledge that increasing deployment of AI could have unintended but severe consequences for human lives, ranging from impairments of friendships to social disorder and war.[7] These concerns multiply in a context of declining trust in government and key private institutions.[8]

An obvious source of distrust is evidence of unreliability. Unreliability could arise around a specific task, such as finding that your child did not run the errand to buy milk as you requested. Or it could follow self-dealing: did your child keep the change from funds used to purchase the milk rather than returning the unused money to you? Trust is needed when we lack the time or ability to oversee each task to ensure truthful and accurate performance and devotion to the interests of those relying on the tasks being done.

Political theorist Russell Hardin explains trust as "encapsulated interest, in which the truster's expectations of the trusted's behavior depend on assessments of certain motivations of the trusted. I trust you because your interests encapsulate mine to some extent – in particular, because you want our relationship to continue."[9] Trust accordingly is grounded in the truster's assessment of the intentions of the trusted with respect to some action.[10] Trust is strengthened when I believe it is in your interest to adhere to my interests in the relevant matter.[11] Those who rely on institutions, such as the law, have reasons to believe that they comport with governing norms and practices rather than serving other interests.

Trust in hospitals and schools depends on assessments of the reliability of the institution and its practices in doing what it promises to do, as well as its responses to inevitable mistakes.[12] With repeated transactions, trust depends not only

on results, but also on discernable practices reducing risks of harm and deviation from expected tasks. Evidence that the institution or its staff serves the interests of intended beneficiaries must include guards against violation of those interests. Trust can grow when a hospital visibly uses good practices with good results and communicates the measures to minimize risks of bad results and departures from good practices.

External indicators, such as accreditation by expert review boards, can signal adherence to good practices and reason to expect good results. External indicators can come from regulators who set and enforce rules, such as prohibitions of self-dealing through bans on charging more than is justifiable for a procedure and prohibiting personal or institutional financial interests that are keyed to the volume of referrals or uses.[13] Private or governmental external monitors must be able to audit the behavior of institutions.[14] External review will not promote trust if external monitors are not themselves trusted. In fact, disclosure amid distrust can feed misunderstandings.[15]

Past betrayals undermine trust. Personal and collective experiences with discrimination or degradation – along lines of race, class, gender, or other personal characteristics – especially create reasons for suspicion if not outright distrust. Similarly, experiences with self-interested companies that make exploitative profits can create or sustain distrust. Distrust and the vigilance it inspires may itself protect against exploitation.[16]

These and further sources of distrust come with uses of AI, by which we mean: a variety of techniques to discern patterns in historical "training" data that are determinative of status (is the tumor benign?) or predictive of a future outcome (what is the likelihood the student will graduate within four years?). The hope is that the patterns discerned in the training data will extend to future unseen examples. Algorithms trained on data are "learned algorithms." These learned algorithms classify and score individuals as the system designer chose, equitably or not, to represent them to the algorithm. These representations of individuals and "outcomes" can be poor proxies for the events of interest, such as using re-arrest as a proxy for recidivism or a call to child protective services as a proxy for child endangerment.[17]

Distrust also results from the apparent indifference of AI systems. Learned algorithms lack indications of adherence to the interests of those affected by their use. They also lack apparent conformity with norms or practices legible to those outside of their creation and operations.

When designed solely at the directive of governments and companies, AI may only serve the interests of governments and companies – and risk impairing the interests of others.

Despite sophisticated techniques to teach algorithms from data sets, *there is no ground truth* available to check whether the results match reality. This is a basic challenge for ensuring reliable AI. We can prove that the

learned algorithm is indeed the result of applying a specific learning technique to the training data, but when the learned algorithm is applied to a previously unseen individual, one not in the training data, we do not have proof that the outcome is correct in terms of an underlying factual basis, rather than inferences from indirect or arbitrary factors. Consider an algorithm asked to predict whether a given student will graduate within four years. This is a question about the future: when the algorithm is applied to the data representing the student, the answer has not yet been determined. A similar quandary surrounds risk scoring: what is the "probability" that an individual will be re-arrested within two years? This question struggles to make sense even mathematically: what is the meaning of the "probability" of a nonrepeatable event?[18] Is what we perceive as randomness in fact certainty, if only we had sufficient contextual information and computing power? Inferences about the future when predicated on limited or faulty information may create an illusion of truth, but illusion it is.

Further problems arise because techniques for building trust are too often unavailable with algorithms used for scoring and categorizing people for public or private purposes. Familiar trust-building techniques include transparency so others can see inputs and outcomes, opportunities for those affected to participate in designing and evaluating a system and in questioning its individual applications, monitoring and evaluation by independent experts, and regulation and oversight by government bodies.

Trust in the fairness of legal systems increases when those affected participate with substantive, empowering choices within individual trials or panels reviewing the conduct of police and other officials. Could participation of those affected by AI help build trust in uses of AI?[19] Quite apart from influencing outcomes, participation gives people a sense that they are valued, heard, and respected.[20] Participatory procedures signal fairness, help to resolve uncertainties, and support deference to results.[21] Following prescribed patterns also contributes to the perceived legitimacy of a dispute resolution system.[22]

But there are few if any roles for consumers, criminal defendants, parents, or social media users to raise questions about the algorithms used to guide the allocation of benefits and burdens. Nor are there roles for them in the construction of the information-categorizing algorithms. Opportunities to participate are not built into the design of algorithms, data selection and collection protocols, or the testing, revision, and use of learning algorithms. Ensuring a role for human beings to check algorithmic processes can even be a new source of further inaccuracies. An experiment allowing people to give feedback to an algorithmically powered system actually showed that participation lowered trust – perhaps by exposing people to the scope of the system's inaccuracies.[23]

Suggestions for addressing distrust revolve around calls for "explainability" and ensuring independent entities access to the learned algorithms themselves.[24]

"Access" can mean seeing the code, examining the algorithm's outputs, and reviewing the choice of representation, sources of training data, and demographic benchmarking.[25] But disclosure of learning algorithms themselves has limited usefulness in the absence of data features with comprehensible meanings and explanations of weight determining the contribution of each feature to outcomes. Machine learning algorithms use mathematical operations to generate data features that almost always are not humanly understandable, even if disclosed, and whose learned combinations would do nothing to explain outcomes, even to expert auditors.

Regulation can demand access and judgments by qualified experts and, perhaps more important, require behavior attentive not only to narrow interests but also to broader public concerns. Social distrust of X-rays produced demands for regulation; with regulation, professional training, and standards alert to health effects, X-rays gained widespread trust.[26] Yet government regulators and independent bodies can stoke public fears if they contribute to misinformation and exaggerate risks.[27]

For many, reliance on AI arouses fears of occupational displacement. Now white collar as well as blue collar jobs seem at risk. One study from the United Kingdom reported that more than 60 percent of people surveyed worry that their jobs will be replaced by AI. Many believe that their jobs and opportunities for their children will be disrupted.[28] More than one-third of young Americans report fears about technology eliminating jobs.[29] Despite some predictions of expanded and less-repetitive employment, no one can yet resolve doubts about the future.[30] Foreboding may be exacerbated by awareness that, by our uses of technology, we contribute to the trends we fear. Many people feel forced to use systems such as LinkedIn or Facebook.[31] People report distrust of the Internet but continue to increase their use of it.[32]

Some distrust AI precisely because human beings are not visibly involved in decisions that matter to human beings. Yet even the chance to appeal to a human is insufficient when individuals are unaware that there is a decision or score affecting their lives.

As companies and governments increase their use of AI, distrust mounts considerably with misalignment of interests. Airbnb raised concerns when it acquired Trooly Inc., including its patented "trait analyzer" that operates by scouring blogs, social networks, and commercial and public databases to derive personality traits. The patent claims that "the system determines a trustworthiness score of the person based on the behavior and personality trait metrics using a machine learning system," with the weight of each personality trait either hard coded or inferred by a machine learning model.[33] It claims to identify

traits as "badness, anti-social tendencies, goodness, conscientiousness, openness, extraversion, agreeableness, neuroticism, narcissism, Machiavellianism, or psychopathy."[34] Although Airbnb asserts that the company is not currently deploying this software,[35] the very acquisition of a "trait analyzer" raises concerns that the company refuses to encapsulate the interests of those affected.[36]

Examples of practices harming and contrary to the interests of users abound in social media platforms, especially around demonstrated biases and invasions of privacy. Although social media companies offer many services that appeal to users, the companies have interests that diverge systematically from those of users. Platform companies largely profit off data generated by each person's activities on the site. Hence, the companies seek to maximize user "engagement." Each new data point comes when a user does – or does not – click on a link or hit a "like" button. The platform uses that information to tailor content for users and to sell their information to third parties for targeted advertising and other messages.[37] Chamath Palihapitiya, former vice president for "user growth" for Facebook, has claimed that Facebook is addictive by design.[38] Sean Parker, an original Facebook investor, has acknowledged that the site's "like" button and news feed keep users hooked by exploiting people's neurochemical vulnerabilities.[39]

Privacy loss is a particular harm resented by many. Privacy can mean seclusion, hiding one's self, identity, and information; it can convey control over one's personal information and who can see it; it can signal control over sensitive or personal decisions, without interference from others; or it can mean protection against discrimination by others based on information about oneself. All these meanings matter in the case of Tim Stobierski, who, shortly after starting a new job at a publishing house, was demonstrating a Facebook feature to his boss when an advertisement for a gay cruise appeared on his news feed.[40] He wondered, "how did Facebook know that I was interested in men, when I had never told another living soul, and when I certainly had never told Facebook?"[41] The Pew Research Center showed that about half of all Facebook users feel discomfort about the site's collection of their interests, while 74 percent of Facebook users did not know how to find out how Facebook categorized their interests or even how to locate a page listing "your ad preferences."[42] A platform's assumptions remain opaque even as users resent the loss of control over their information and the secret surveillance.[43]

Tech companies may respond that users can always quit. Here, too, a conflict of interests is present. Facebook exposes individuals to psychological manipulation and data breaches to degrees that they cannot imagine.[44] Most users do not even know how Facebook uses their data or what negative effects can ensue.[45] The loss of control compounds the unintended spread of personal information.

The interests of tech platforms and users diverge further over hateful speech. Facebook's financial incentive is to keep or even elicit outrageous posts because they attract engagement (even as disagreement or disgust) and hence produce

additional monetizable data points.[46] Facebook instructs users to hide posts they do not like, or to unfollow the page or person who posted it, and, only as a third option, to report the post to request its removal.[47] Under pressure, Facebook established an oversight review board and charged it with evaluating (only an infinitesimal fraction of) removal decisions. Facebook itself determines which matters proceed to review.[48] Directed to promote freedom of speech, not to guard against hatred or misinformation, the board has so far done little to guard against fomented hatred and violence.[49]

Large tech companies are gatekeepers; they can use their position and their knowledge of users to benefit their own company over others, including third parties that pay for their services.[50] As one observer put it, "social media is cloaked in this language of liberation while the corporate sponsors (Facebook, Google *et al.*) are progressing towards ever more refined and effective means of manipulating individual behavior (behavioral targeting of ads, recommendation systems, reputation management systems etc.)."[51]

The processes of AI baffle the open and rational debates supporting democracies, markets, and science that have existed since the Enlightenment. AI practices can nudge and change what people want, know, and value.[52] Differently organized, learned algorithms could offer people some control over site architecture and content moderation.[53]

Dangers from social media manipulation came to fruition with the 2020 U.S. presidential election. Some conventional media presented rumors and falsehood, but social media initiated and encouraged misinformation and disinformation, and amplified their spread, culminating in the sweeping erroneous belief that Donald Trump rather than Joe Biden had won the popular vote. False claims of rigged voting machines, despite the certification of state elections, reflected and inflamed social distrust.[54] The sustainability of our democratic governance systems is far from assured.

Building trust around AI can draw on commitments to participation, useable explanations, and iterative improvements. Hence, people making and deploying AI should involve broader and diverse stakeholders in decisions around what uses algorithms are put to; what data, with which features, are used to train the algorithms; what criteria are used in the training process to evaluate classifications or predictions; and what methods of recourse are available for raising concerns about and securing genuine responsive action to potentially unjust methods or outcomes. Creative and talented people have devised AI algorithms able to infer our personal shopping preferences; they could deploy their skills going forward to devise opportunities for those affected to participate in identifying gaps and distortions in data. Independent experts in academic and nonprofit settings – if given access to critical information – could provide much-needed audits of algo-

rithmic applications and assess the reliability and failures of the factors used to draw inferences.

Investment in participatory and information-sharing efforts should be commensurate with the risks of harms. Otherwise, the risks are entirely shifted to the consumers, citizens, and clients who are subjected to the commercial and governmental systems that deploy AI algorithms.

As AI escalates, so should accessible methods of recourse and correction. Concerns for people harmed by harassment on social media; biased considerations in employment, child protection, and other governmental decisions; and facial recognition technologies that jeopardize personal privacy and liberty will be echoed by known and unknown harms in finance, law, health care, policing, and war-making. Software systems to enable review and to redress mistakes should be built, and built to be meaningful. Designers responding that doing so would be too expensive or too difficult given the scale enabled by the use of AI algorithms are scaling irresponsibly. Responsible scaling demands investment in methods of recourse for errors and bias commensurate with the risks of errors and bias. AI can and must be part of the answer in addressing the problems created by AI, but so must strengthened roles for human participation. Government by the consent of the governed needs no less.[55]

Self-regulation and self-certification, monitoring by external industry and consumer groups, and regulation by government can tackle misalignment and even clashes in the interests of those designing the learning algorithms and those affected by them. Entities should compete in the marketplace for trust and reputation, face ratings by external monitors, and contribute to the development of industry standards. Trust must be earned.

AUTHORS' NOTE

We thank Christian Lansang, Maroussia Lévesque, and Serena Wong for thoughtful research and advice for this essay.

ABOUT THE AUTHORS

Cynthia Dwork, a Fellow of the American Academy since 2008, is the Gordon McKay Professor of Computer Science in the John Paulson School of Engineering and Applied Sciences and Radcliffe Alumnae Professor in the Radcliffe Institute for Advanced Study at Harvard University. Her work has established the pillars of fault-tolerant distributed systems, modernized cryptography to the ungoverned

interactions of the Internet and the era of quantum computing, revolutionized privacy-preserving statistical data analysis, and launched the field of algorithmic fairness.

Martha Minow, a Fellow of the American Academy since 1992, is the 300th Anniversary University Professor and Distinguished Service Professor at Harvard University. She also serves as Cochair of the American Academy's project on Making Justice Accessible: Designing Legal Services for the 21st Century. She is the author of, most recently, *Saving the News: Why the Constitution Calls for the Government to Act to Preserve the Freedom of Speech* (2021), *When Should Law Forgive?* (2019), and *In Brown's Wake: Legacies of America's Educational Landmark* (2010).

ENDNOTES

[1] See "AI in Pop Culture," ThinkAutomation, https://www.thinkautomation.com/bots-and-ai/ai-in-pop-culture/.

[2] Darrell M. West and John. R. Allen, *Turning Point: Policymaking in the Era of Artificial Intelligence* (Washington, D.C.: Brookings Institution Press, 2020).

[3] Although many are optimistic about new technologies, concerns over loss of control are growing. See Ethan Fast and Eric Horvitz, "Long-Term Trends in the Public Perception of Artificial Intelligence," in *AAAI '17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (Menlo Park: Calif.: Association for the Advancement of Artificial Intelligence, 1979), 963.

[4] National Security Commission on Artificial Intelligence, *Final Report: Artificial Intelligence in Context* (Washington, D.C.: National Security Commission on Artificial Intelligence, 2021), https://reports.nscai.gov/final-report/ai-in-context/.

[5] "Trust," Oxford English Dictionary Online, http://www.oed.com/view/Entry/207004 (accessed November 16, 2020).

[6] See Deloitte, *Managing Algorithmic Risks: Safeguarding the Use of Complex Algorithms and Machine Learning* (London: Deloitte Development, LLC, 2017), https://www2.deloitte.com/content/dam/Deloitte/us/Documents/risk/us-risk-algorithmic-machine-learning-risk-management.pdf; Simson L. Garfinkel, "A Peek at Proprietary Algorithms," *American Scientist* 105 (6) (2017), https://www.americanscientist.org/article/a-peek-at-proprietary-algorithms; and Stacy Meichtry and Noemie Bisserbe, "France's Macron Calls for Regulation of Social Media to Stem 'Threat to Democracy,'" *The Wall Street Journal*, January 29, 2021, https://www.wsj.com/articles/frances-macron-calls-for-regulation-of-social-media-to-stem-threat-to-democracy-11611955040.

[7] Such as National Security Commission on Artificial Intelligence, *Final Report*, "Chapter 7: Establishing Justified Confidence in AI Systems" and "Chapter 8: Upholding Democratic Values: Privacy, Civil Liberties, and Civil Rights in Uses of AI for National Security." See also Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values* (Washington, D.C.: Executive Office of the President, 2014), https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

[8] Lee Rainie, Scott Keeter, and Andrew Perrin, "Trust and Distrust in America," Pew Research Center, https://www.pewresearch.org/politics/2019/07/22/trust-and-distrust-in-america/. On the psychology of distrust, see Roy J. Lewicki and Edward C. Tomlinson, "Distrust," Beyond Intractability, December 2003, https://www.beyondintractability.org/essay/distrust. See also Embedded Ethics @ Harvard, https://embeddedethics.seas.harvard.edu/; "Ethics, Computer, and AI: Perspectives from MIT, Human Contexts and Ethics," MIT News, March 18, 2019, https://news.mit.edu/2019/ethics-computing-and-ai-perspectives-mit-0318; and Berkeley Computing, Data Science, and Society, https://data.berkeley.edu/hce.

[9] Russell Hardin, *Trust and Trustworthiness* (New York: Russell Sage Foundation, 2002), xix.

[10] Ibid., xx.

[11] Ibid., 4.

[12] See Pierre Lauret, "Why (and How to) Trust Institutions? Hospitals, Schools, and Liberal Trust," *Rivista di Estetica* 68 (2018): 41–68, https://doi.org/10.4000/estetica.3455.

[13] AMA Council on Ethical and Judicial Affairs, "AMA Code of Medical Ethics' Opinions on Physicians' Financial Interests," Opinion 8.0321–Physicians' Self-Referral, *AMA Journal of Ethics* (August 2015), https://journalofethics.ama-assn.org/article/ama-code-medical-ethics-opinions-physicians-financial-interests/2015-08. For analogous treatment of AI, see Matthew Hutson, "Who Should Stop Unethical A.I.?" *The New Yorker*, February 15, 2021, https://www.newyorker.com/tech/annals-of-technology/who-should-stop-unethical-ai.

[14] See Michael Kearns and Aaron Roth, "Ethical Algorithm Design Should Guide Technology Regulation," The Brookings Institution, January 13, 2020, https://www.brookings.edu/research/ethical-algorithm-design-should-guide-technology-regulation/.

[15] See Ethan Zuckerman, *Distrust: Why Losing Faith in Institutions Provides the Tools to Transform Them* (New York: W. W. Norton & Company, 2021), 60 (describing tendencies of people living with broken institutions to wrongly see patterns and conspiracies in random occurrences).

[16] Roderick M. Kramer, "Rethinking Trust," *Harvard Business Review*, June 2009, https://hbr.org/2009/06/rethinking-trust; and Christopher B. Yenkey, "The Outsider's Advantage: Distrust as a Deterrent to Exploitation," *American Journal of Sociology* 124 (3) (2018): 613.

[17] Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: St. Martin's Press, 2018).

[18] See, for example, A. Philip Dawid, "On Individual Risk," *Synthese* 194 (2017); and Cynthia Dwork, Michael P. Kim, Omer Reingold, et al., "Outcome Indistinguishability," in *STOC 2021: Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing* (New York: Association for Computing Machinery, 2021). For a treatment of individual probabilities in risk assessment tools, see Peter B. Imrey and A. Philip Dawid, "A Commentary on the Statistical Assessment of Violence and Recidivism Risks," *Statistics and Public Policy* 2 (1) (2015): 1–18; and Kristian Lum, David B. Dunson, and James E. Johndrow, "Closer than They Appear: A Bayesian Perspective on Individual-Level Heterogeneity in Risk Assessment," arXiv (2021), https://arxiv.org/abs/2102.01135.

[19] Tom R. Tyler, "Procedural Justice, Legitimacy, and the Effective Rule of Law," *Crime and Justice* 30 (2003): 283, https://www.jstor.org/stable/1147701?seq=1#metadata_info_tab_contents.

[20] See E. Allan Lind and Tom R. Tyler, eds., *The Social Psychology of Procedural Justice* (New York: Springer, 1988).

[21] Kees van den Bos, Lynn van der Velden, and E. Allan Lind, "On the Role of Perceived Procedural Justice in Citizens' Reactions to Government Decisions and the Handling of Conflicts," *Utrecht Law Review* 10 (4) (2014): 1–26, https://www.utrechtlawreview.org/articles/abstract/10.18352/ulr.287/.

[22] Rebecca Hollander-Blumoff and Tom R. Tyler, "Procedural Justice and the Rule of Law: Fostering Legitimacy in Alternative Dispute Resolution," *Journal of Dispute Resolution* 1 (2011).

[23] Donald R. Honeycutt, Mahsan Nourani, and Eric D. Ragan, "Soliciting Human-in-the-Loop User Feedback for Interactive Machine Learning Reduces User Trust and Impressions of Model Accuracy," in *Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing* (Menlo Park, Calif.: Association for the Advancement of Artificial Intelligence, 2020), 63–72, https://ojs.aaai.org/index.php/HCOMP/article/view/7464.

[24] See, for example, David Leslie, "Project ExplAIn," The Alan Turing Institute, December 2, 2019, https://www.turing.ac.uk/news/project-explain, which describes six kinds of "explanation types," including identifying who is involved in the development and management of an AI system and whom to contact for human review, as well as the effects that the AI system has on an individual and on wider society.

[25] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, et al., "Datasheets for Datasets," arXiv (2018), https://arxiv.org/abs/1803.09010; and Margaret Mitchell, Simone Wu, Andrew Zaldivar, et al., "Model Cards for Model Reporting," in *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2019), 220–229. See also Cynthia Dwork, Michael P. Kim, Omer Reingold, et al., "Outcome Indistinguishability," in STOC *2021: Proceedings of the 53rd Annual* ACM SIGACT *Symposium on Theory of Computing* (New York: Association for Computing Machinery, 2021).

[26] See Antony Denman, S. Parkinson, and Christopher John Groves-Kirby, "A Comparative Study of Public Perception of Risks from a Variety of Radiation and Societal Risks," presented at the 11th International Congress of the International Radiation Protection Association, Madrid, Spain, May 23–28, 2004.

[27] John M. Osepchuk, "A History of Microwave Heating Applications," *IEEE Transactions on Microwave Theory and Techniques* 32 (9) (1984): 1200, 1213.

[28] Jacob Douglas, "These American Workers Are the Most Afraid of A.I. Taking Their Jobs," CNBC, November 7, 2019 (37 percent of people surveyed aged eighteen to twenty-four expressed fear AI will take their jobs), https://www.cnbc.com/2019/11/07/these-american-workers-are-the-most-afraid-of-ai-taking-their-jobs.html; and "Technically Redundant: Six-in-10 Fear Losing Their Jobs to AI," Industry Europe, November 3, 2019, https://industryeurope.com/technically-redundant-six-in-10-fear-losing-their-jobs-to-ai/.

[29] Douglas, "These Americans Are the Most Afraid of AI Taking Their Jobs."

[30] James E. Bessen, Stephen Impink, Lydia Reichensperger, and Robert Seamans, "The Business of AI Startups" (Boston: Boston University School of Law, 2018), https://ssrn.com/abstract=3293275 or http://dx.doi.org/10.2139/ssrn.3293275.

[31] Since the writing of this essay, Facebook has been rebranded as Meta.

[32] Lee Raine and Janna Anderson, "Theme 3: Trust Will Not Grow, but Technology Usage Will Continue to Rise as a 'New Normal' Sets in," Pew Research Center, August 10, 2017, https://www.pewresearch.org/internet/2017/08/10/theme-3-trust-will-not-grow-but-technology-usage-will-continue-to-rise-as-a-new-normal-sets-in/.

[33] Sarabjit Singh Baveja, Anish Das Sarma, and Nilesh Dalvi, United States Patent No. 9070088 B1: Determining Trustworthiness and Compatibility of a Person, June 30, 2015, https://patentimages.storage.googleapis.com/36/36/7e/db298c5d3b280c/US9070088.pdf (accessed January 11, 2022).

[34] Ibid., 3–4.

[35] Aaron Holmes, "Airbnb Has Patented Software that Digs through Social Media to Root Out People Who Display 'Narcissism or Psychopathy,'" *Business Insider*, January 6, 2020, https://www.businessinsider.com/airbnb-software-predicts-if-guests-are-psychopaths-patent-2020-1.

[36] See text above from endnotes 14–17.

[37] Tero Karppi, *Disconnect: Facebook's Affective Bonds* (Minneapolis: University of Minnesota Press, 2018).

[38] Ibid.

[39] Tero Karppi and David B. Nieborg, "Facebook Confessions: Corporate Abdication and Silicon Valley Dystopianism," *New Media & Society* 23 (9) (2021); Sarah Friedman, "How Your Brain Responds Every Time Your Insta Post Gets a 'Like,'" *Bustle*, September 21, 2019, https://www.bustle.com/p/how-your-brain-responds-to-a-like-online-shows-the-power-of-social-media-18725823; and Trevor Haynes, "Dopamine, Smartphones, and You: A Battle for Your Time," Science in the News Blog, Harvard Medical School, May 1, 2018, https://sitn.hms.harvard.edu/flash/2018/dopamine-smartphones-battle-time/.

[40] Tim Stobierski, "Facebook Ads Outed Me," INTO, May 3, 2018, https://www.intomore.com/you/facebook-ads-outed-me/#.

[41] Ibid.

[42] Paul Hitlin and Lee Rainie, "Facebook Algorithms and Personal Data," Pew Research Center, January 16, 2019, https://www.pewresearch.org/internet/2019/01/16/facebook-algorithms-and-personal-data/. New "post-cookie" advertising schemes enlist the browser to perform user categorization previously carried out by advertising networks. Heralded as privacy-preserving because the browser is local to the user's machine, these systems are designed to carry out the same segmentation that so many find objectionable; see Bennet Cyphers, "Google's FLoC Is a Terrible Idea," March 3, 2021, Electronic Frontier Foundation. Standard notions of privacy do not ensure fair treatment. Cynthia Dwork and Deirdre K. Mulligan, "It's Not Privacy and It's Not Fair," *Stanford Law Review* 66 (2013), https://www.stanfordlawreview.org/online/privacy-and-big-data-its-not-privacy-and-its-not-fair/ ("The urge to classify is human. The lever of big data, however, brings ubiquitous classification, demanding greater attention to the values embedded and reflected in classifications, and the roles they play in shaping public and private life.")

[43] Moreover, "privacy solutions can hinder efforts to identify classifications that unintentionally produce objectionable outcomes–for example, differential treatment that tracks race or gender–by limiting the availability of data about such attributes." Dwork and Mulligan, "It's Not Privacy and It's Not Fair."

[44] See Karppi and Nieborg, "Facebook Confessions," 11.

[45] Ibid.

[46] Karppi, *Disconnect: Facebook's Affective Bonds*.

[47] Eugenia Siapera and Paloma Viejo-Otero, "Governing Hate: Facebook and Digital Racism," *Television & New Media* 22 (2) (2020): 112–113, 122.

[48] Evelyn Douek, "What Kind of an Oversight Board Have You Given Us?" *University of Chicago Law Review Online*, May 11, 2020, https://lawreviewblog.uchicago.edu/2020/05/11/fb-oversight-board-edouek/.

[49] See Andrew Marantz, "Why Facebook Can't Fix Itself," *The New Yorker*, October 12, 2020.

[50] Lina M. Khan, "Sources of Tech Platform Power," *Georgetown Law Technology Review* 2 (2) (2018): 325.

[51] Joshua-Michéle Ross, "The Question Concerning Social Technology," Radar, May 18, 2009, http://radar.oreilly.com/2009/05/the-question-concerning-social.html; and "Do Social Media Threaten Democracy?" *The Economist*, November 4, 2017, https://www.economist.com/leaders/2017/11/04/do-social-media-threaten-democracy.

[52] Ibid.; and *The Social Dilemma*, dir. Jeff Orlowski (Boulder, Colo.: Exposure Labs, Argent Pictures, The Space Program, 2020).

[53] See Daphne Keller, "The Future of Platform Power: Making Middleware Work," *Journal of Democracy* 32 (3) (2021), https://www.journalofdemocracy.org/articles/the-future-of-platform-power-making-middleware-work/.

[54] See, for example, Aaron Blake, "Trump's 'Big Lie' Was Bigger Than Just a Stolen Election," *The Washington Post*, February 12, 2021, https://www.washingtonpost.com/politics/2021/02/12/trumps-big-lie-was-bigger-than-just-stolen-election/; Melissa Block, "Can The Forces Unleashed By Trump's Big Election Lie Be Undone?" NPR, January 16, 2021, https://www.npr.org/2021/01/16/957291939/can-the-forces-unleashed-by-trumps-big-election-lie-be-undone; and Christopher Giles and Jake Horton, "U.S. Election 2020: Is Trump Right about Dominion Machines?" BBC, November 17, 2020, https://www.bbc.com/news/election-us-2020-54959962.

[55] See Richard Fontaine and Kara Frederick, "Democracy's Digital Defenses," *The Wall Street Journal*, May 7, 2021, https://www.wsj.com/amp/articles/democracys-digital-defenses-11620403161.