

NOTES

Introduction: Judging Machines

- 1** M. W. Shelley, *Frankenstein, or The Modern Prometheus* (Dent, 1869).
- 2** D. Victor, “Microsoft Created a Twitter Bot to Learn from Users. It Quickly Became a Racist Jerk,” *New York Times*, 25 March 2018.
- 3** E. L. Eisenstein, *The Printing Press as an Agent of Change: Communications and Cultural Trans* (Cambridge University Press, 1980).
- 4** C. Juma, *Innovation and Its Enemies: Why People Resist New Technologies* (Oxford University Press, 2016).
- 5** K. Sale, *Rebels against the Future—The Luddites and Their War on the Industrial Revolution: Lessons for the Computer Age* (Basic Books, 1996).
- 6** B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano, “Sacrifice One for the Good of Many? People Apply Different Moral Norms to Human and Robot Agents,” in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2015), 117–124, <https://doi.org/10.1145/2696454.2696458>.
- 7** J. J. Thomson, “Killing, Letting Die, and the Trolley Problem,” *Monist* 59(2) (1976), 204–217, <https://doi.org/10.5840/monist197659224>.
- 8** Malle et al., “Sacrifice One for the Good of Many?”;

I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J. F. Bonnefon, C. Breazeal, et al., “Machine Behaviour,” *Nature* 568 (2019): 477–486;

P. Lin, K. Abney, and G. A. Bekey, eds., *Robot Ethics: The Ethical and Social Implications of Robotics* (MIT Press, 2014);

D. J. Gunkel, “The Other Question: Can and Should Robots Have Rights?” *Ethics and Information Technology* 20 (2018): 87–99;

B. J. Dietvorst, J. P. Simmons, and C. Massey, “Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err,” *Journal of Experimental Psychology: General* 144 (2015): 114–126;

R. V. Yampolskiy, “Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach,” in *Philosophy and Theory of Artificial Intelligence*, ed. V. C. Müller (Springer, 2013), 389–396;

R. V. Yampolskiy, *Artificial Intelligence Safety and Security* (Chapman and Hall/CRC, 2018);

E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, et al., “The Moral Machine Experiment,” *Nature* 563 (2018): 59–64;

S. Hajian, F. Bonchi, and C. Castillo, “Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016), 2125–2126, <https://doi.org/10.1145/2939672.2945386>;

J-F. Bonnefon, A. Shariff, and I. Rahwan, *The Moral Psychology of AI and the Ethical Opt-Out Problem*. The Ethics of Artificial Intelligence (2019).

9 Awad et al., “The Moral Machine Experiment”;

J-F. Bonnefon, A. Shariff, and I. Rahwan, “The Social Dilemma of Autonomous Vehicles,” *Science* 352 (2016): 1573–1576.

10 R. Baeza-Yates, “Data and Algorithmic Bias in the Web,” in *Proceedings of the 8th ACM Conference on Web Science* (ACM, 2016), <https://doi.org/10.1145/2908131.2908135>;

B. Friedman and H. Nissenbaum, “Bias in Computer Systems,” *ACM Transactions on Information Systems* 14 (1996): 330–347;

S. Hajian, F. Bonchi, and C. Castillo, “Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining”;

J. Buolamwini and T. Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” in *Conference on Fairness, Accountability, and Transparency* (ACM, 2018), 77–91.

HOW HUMANS JUDGE MACHINES

- 11** Dietvorst et al., “Algorithm Aversion.”
- 12** Dietvorst et al., “Algorithm Aversion.”
- 13** E. Broadbent, “Interactions with Robots: The Truths We Reveal about Ourselves,” *Annual Review of Psychology* 68 (2017): 627–652;
- C. Bartneck, T. Belpaeme, E. Friederike, T. Kanda, M. Keijsers, and S. Šabanovic, *Human-Robot Interaction: An Introduction* (Cambridge University Press, 2019).
- 14** S. Cave, C. Craig, K. Dihal, S. Dillon, J. Montgomery, B. Singler, and L. Taylor, *Portrayals and Perceptions of AI and Why They Matter* (2018);
- <https://royalsociety.org/~media/policy/projects/ai-narratives/AI-narratives-workshop-findings.pdf>.
- 15** H. C. Barrett, A. Bolyanatz, A. N. Crittenden, D. M. T. Fessler, S. Fitzpatrick, M. Gurven, et al., “Small-Scale Societies Exhibit Fundamental Variation in the Role of Intentions in Moral Judgment,” *Proceedings of the National Academy of Sciences* 113 (2016): 4688–4693;
- R. A. McNamara, A. K. Willard, A. Norenzayan, and J. Henrich, “Weighing Outcome vs. Intent across Societies: How Cultural Models of Mind Shape Moral Reasoning,” *Cognition* 182 (2019): 95–108;
- E. Awad, S. Dsouza, A. Shariff, I. Rahwan, and J.-F. Bonnefon, “Universals and Variations in Moral Decisions Made in 42 Countries by 70,000 Participants,” *Proceedings of the National Academy of Sciences* 117 (2020): 2332–2337.

Chapter 1: The Ethics of Artificial Minds

- 1** Awad et al., “The Moral Machine Experiment”;
- Bonnefon et al., “The Social Dilemma of Autonomous Vehicles”;
- P. Lin, “Why Ethics Matters for Autonomous Cars,” in *Autonomous Driving: Technical, Legal and Social Aspects*, ed. M. Maurer, J. C. Gerdes, B. Lenz, and H. Winner (Springer Berlin Heidelberg, 2016), 69–85, https://doi.org/10.1007/978-3-662-48847-8_4;

HOW HUMANS JUDGE MACHINES

Lin et al., *Robot Ethics 2.0*.

- 2 Bonnefon et al., “The Social Dilemma of Autonomous Vehicles.”
- 3 Awad et al., “The Moral Machine Experiment.”
- 4 Buolamwini and Gebru, “Gender Shades.”
- 5 D. Autor and A. Salomons, “Is Automation Labor Share-Displacing? Productivity Growth, Employment, and the Labor Share,” *Brookings Papers on Economic Activity* (2018): 1–87;

D. Acemoglu and P. Restrepo, *Artificial Intelligence, Automation and Work*. <http://www.nber.org/papers/w24196> (2018), <https://doi.org/10.3386/w24196>;
- A. Alabdulkareem, M. R. Frank, L. Sun, B. AlShebli, C. Hidalgo, and I. Rahwan, “Unpacking the Polarization of Workplace Skills,” *Science Advances* 4 (2018): eaa06030;
- E. Brynjolfsson, T. Mitchell, and D. Rock, “What Can Machines Learn, and What Does It Mean for Occupations and the Economy?,” *AEA Papers and Proceedings* 108 (2018): 43–47.
- 6 A. Martínez-Ballesté, H. A. Rashwan, D. Puig, and A. P. Fullana, “Towards a Trustworthy Privacy in Pervasive Video Surveillance Systems,” in *2012 IEEE International Conference on Pervasive Computing and Communications Workshops* (IEEE, 2012), 914–919;

A. Datta, M. C. Tschantz, and A. Datta, “Automated Experiments on Ad Privacy Settings,” *Proceedings on Privacy-Enhancing Technologies* (2015): 92–112;
- Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, “Unique in the Crowd: The Privacy Bounds of Human Mobility,” *Scientific Reports* 3 (3) (2013): 1376.
- 7 E. L. Denton, S. Chintala, A. Szlam, and R. Fergus, “Deep Generative Image Models Using a Laplacian Pyramid of Adversarial Networks,” in *Advances in Neural Information Processing Systems*, eds. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Curran Associates, 2015), 1486–1494;

P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2017), 5967–5976, <https://doi.org/10.1109/CVPR.2017.632>;

HOW HUMANS JUDGE MACHINES

A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” *arXiv:1511.06434 [cs]* (2015);

A. Odena, C. Olah, and J. Shlens, “Conditional Image Synthesis with Auxiliary Classifier GANs,” *arXiv:1610.09585 [cs, stat]* (2016).

8 S. Russell, “Take a Stand on AI Weapons,” *Nature* 521 (2015): 415.

9 A. Elder, “False Friends and False Coinage: A Tool for Navigating the Ethics of Sociable Robots,” *SIGCAS Computers and Society* 45 (2016) 248–254;

A. M. Elder, *Friendship, Robots, and Social Media: False Friends and Second Selves* (Routledge, 2017), <https://doi.org/10.4324/9781315159577>.

10 Denton et al., “Deep Generative Image Models”;

Radford et al., “Unsupervised Representation Learning.”

11 P. Maes, “Agents That Reduce Work and Information Overload,” in *Readings in Human-Computer Interaction*, eds. R. M. Baecker, J. Grudin, W. A. S. Buxton, and S. Greenberg (Morgan Kaufmann, 1995), 811–821, <https://doi.org/10.1016/B978-0-08-051574-8.50084-4>;

P. Resnick and H. R. Varian, “Recommender Systems,” *Communications of the ACM* (March 1997), <https://dl.acm.org/doi/10.1145/245108.245121>.

12 M. Campbell, A. J. Hoane, and F. Hsu, “Deep Blue,” *Artificial Intelligence* 134 (2002): 57–83.

13 D. A. Ferrucci, “Introduction to ‘This Is Watson,’” *IBM Journal of Research and Development* 56, no. 3–4 (May–June 2012), <https://ieeexplore.ieee.org/abstract/document/6177724>;

R. High, “The Era of Cognitive Systems: An Inside Look at IBM Watson and How It Works,” *IBM Redbooks* (2012), <http://www.redbooks.ibm.com/abstracts/redp4955.html>.

14 D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, et al., “Mastering the Game of Go with Deep Neural Networks and Tree Search,” *Nature* 529 (2016): 484–489;

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, et al., “Mastering the Game of Go without Human Knowledge,” *Nature* 550 (2017): 354–359.

HOW HUMANS JUDGE MACHINES

- 15** N. Bostrom and E. Yudkowsky, “The Ethics of Artificial Intelligence,” in *Cambridge Handbook of Artificial Intelligence*, eds. K. Frankish and W. Ramsey (Cambridge University Press, 2014), 316–334, <https://doi.org/10.1017/CBO9781139046855.020>.
- 16** Rahwan et al., “Machine Behaviour.”
- 17** Rahwan et al., “Machine Behaviour.”
- 18** Lin et al., *Robot Ethics 2.0*;
D. J. Gunkel, “The Other Question: Can and Should Robots Have Rights?,” *Ethics and Information Technology* 20 (2018): 87–99;
D. J. Gunkel, *Robot Rights* (MIT Press, 2018);
G. McGee, “A Robot Code of Ethics,” *The Scientist*, 30 April 2017, <https://www.the-scientist.com/column/a-robot-code-of-ethics-46522>.
- 19** Bostrom and Yudkowsky, “The Ethics of Artificial Intelligence.”
- 20** A. Etzioni and O. Etzioni, “Incorporating Ethics into Artificial Intelligence,” *Journal of Ethics* 21 (2017): 403–418;
S. Torrance, “Ethics and Consciousness in Artificial Agents,” *AI & Society* 22 (2008): 495–521;
B. Friedman and P. H. Kahn, “Human Agency and Responsible Computing: Implications for Computer System Design,” *Journal of Systems and Software* 17 (1997): 7–14.
- 21** Gunkel, *Robot Rights*;
McGee, “A Robot Code of Ethics”;
E. Reynolds, “The Agony of Sophia, the World’s First Robot Citizen Condemned to a Lifeless Career in Marketing,” *Wired UK* (2018).
- 22** Gunkel, *Robot Rights*.
- 23** Gunkel, *Robot Rights*;

HOW HUMANS JUDGE MACHINES

J. Carpenter, *Culture and Human-Robot Interaction in Militarized Spaces: A War Story* (Routledge, 2016);

P. W. Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century* (Penguin, 2009);

J. Garreau, "Bots on the Ground," *Washington Post*, 6 May 2007.

24 O. Bendel, "Sex Robots from the Perspective of Machine Ethics," in *International Conference on Love and Sex with Robots* (Springer, 2016): 17–26;

K. Richardson, "Sex Robot Matters: Slavery, the Prostituted, and the Rights of Machines," *IEEE Technology and Society Magazine* 35 (2016): 46–53;

S. Nyholm and L. E. Frank, "It Loves Me, It Loves Me Not: Is It Morally Problematic to Design Sex Robots That Appear to Love Their Owners?," *Techné: Research in Philosophy and Technology* (2019), DOI: 10.5840/techne2019122110.

25 Gunkel, *Robot Rights*;

M. Scheutz, "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots," *Robot Ethics: The Ethical and Social Implications of Robotics* 205 (2011). Edited by Patrick Lin, Keith Abney, and George A. Bekey.

26 P. Foot, "The Problem of Abortion and the Doctrine of Double Effect," *Oxford Review* 5 (1967): 5–15;

Thomson, "Killing, Letting Die, and the Trolley Problem."

27 S. H. Seo, D. Geiskkovitch, M. Nakane, C. King, and J. E. Young, "Poor Thing! Would You Feel Sorry for a Simulated Robot? A Comparison of Empathy toward a Physical and a Simulated Robot," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (ACM, 2015), 125–132.

28 P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Re-make Our World* (Basic Books, 2015).

29 E. Turiel, *The Development of Social Knowledge: Morality and Convention* (Cambridge University Press, 1983);

HOW HUMANS JUDGE MACHINES

J. Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion* (Knopf Doubleday Publishing Group, 2012).

30 A. G. Greenwald, B. A. Nosek, and M. R. Banaji, “Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm,” *Journal of Personality and Social Psychology* 85 (2003): 197–216;

A. G. Greenwald, D. E. McGhee, and J. L. Schwartz, “Measuring Individual Differences in Implicit Cognition: the Implicit Association Test,” *Journal of Personality and Social Psychology* 74 (1998): 1464–1480.

31 Haidt, *The Righteous Mind*.

32 Haidt, *The Righteous Mind*.

33 Pinker, *The Blank Slate: The Modern Denial of Human Nature* (Penguin, 2003).

34 J. Haidt, S. H. Koller, and M. G. Dias, “Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?,” *Journal of Personality and Social Psychology* 65 (1993): 613–628;

R. A. Shweder, M. Mahapatra, and J. G. Miller, “Culture and Moral Development,” *The Emergence of Morality in Young Children* (1987): 1–83.

35 Buolamwini and Gebru, “Gender Shades”;

J. Guszczka, I. Rahwan, W. Bible, M. Cebrian, and V. Katyal, “Why We Need to Audit Algorithms,” *Harvard Business Review* (2018), <https://hbr.org/2018/11/why-we-need-to-audit-algorithms>;

K. Hosanagar and V. Jair, “We Need Transparency in Algorithms, But Too Much Can Backfire,” *Harvard Business Review* (2018), <https://hbr.org/2018/07/we-need-transparency-in-algorithms-but-too-much-can-backfire>;

A. P. Miller, “Want Less-Biased Decisions? Use Algorithms,” *Harvard Business Review* (2018), <https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms>.

36 F. Cushman, “Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment,” *Cognition* 108 (2008): 353–380;

HOW HUMANS JUDGE MACHINES

F. Cushman, R. Sheketoff, S. Wharton, and S. Carey, "The Development of Intent-Based Moral Judgment," *Cognition* 127 (2013): 6–21;

J. D. Greene, F. A. Cushman, L. E. Stewart, K. Lowenberg, L. E. Nystrom, and J. D. Cohen, "Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment," *Cognition* 111 (2009): 364–371;

B. F. Malle and J. Knobe, "The Folk Concept of Intentionality," *Journal of Experimental Social Psychology* 33 (1997): 101–121;

L. Young and R. Saxe, "When Ignorance Is No Excuse: Different Roles for Intent across Moral Domains," *Cognition* 120 (2011): 202–214.

37 Barrett et al., "Small-Scale Societies Exhibit Fundamental Variation"; McNamara et al., "Weighing Outcome vs. Intent."

38 S. Clifford, R. M. Jewell, and P. D. Waggoner, "Are Samples Drawn from Mechanical Turk Valid for Research on Political Ideology?," *Research & Politics* 2 (2015): 2053168015622072;

J. Kees, C. Berry, S. Burton, and K. Sheehan, "An Analysis of Data Quality: Professional Panels, Student Subject Pools, and Amazon's Mechanical Turk," *Journal of Advertising* 46 (2017): 141–155;

K. A. Thomas and S. Clifford, "Validity and Mechanical Turk: An Assessment of Exclusion Methods and Interactive Experiments," *Computers in Human Behavior* 77 (2017): 184–197.

39 A.J. Berinsky, G.A. Huber, and G.S. Lenz, "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk," *Political Analysis* 20 (2012): 351–368.

40 V. Amrhein, S. Greenland, and B. McShane, "Scientists Rise up against Statistical Significance," *Nature* 567 (2019): 305–307.

41 R. L. Wasserstein, A. L. Schirm, and N. A. Lazar, "Moving to a World beyond 'p<0.05,'" *American Statistician* 73 (2019): 1–19.

Chapter 2: Unpacking the Ethics of AI

- 1** B. Dietvorst, “Algorithm Aversion,” *Publicly Accessible Penn Dissertations* (2016);

B. J. Dietvorst, J. P. Simmons, and C. Massey, “Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err,” *Journal of Experimental Psychology: General* 144 (2015): 114–126.
- 2** H. Toivonen and O. Gross, “Data Mining and Machine Learning in Computational Creativity,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5 (2015): 265–275;

E. Francke and B. Alexander, “The Potential Influence of Artificial Intelligence on Plagiarism: A Higher Education Perspective,” in *ECIAIR 2019 European Conference on the Impact of Artificial Intelligence and Robotics* 131 (Academic Conferences and Publishing Limited, 2019);

A. Elgammal, “AI Is Blurring the Definition of Artist: Advanced Algorithms Are Using Machine Learning to Create Art Autonomously,” *American Scientist* 107 (2019): 18–22;

D. J. Gervais, “The Machine as Author,” SSRN (2019), <https://papers.ssrn.com/abstract=3359524>;

J. C. Ginsburg and L. A. Budiardjo, “Authors and Machines,” SSRN (2019), <https://papers.ssrn.com/abstract=3233885>;

K. Hristov, “Artificial Intelligence and the Copyright Dilemma,” *IDEA* 57 (2016), 431.
- 3** Denton et al., “Deep Generative Image Models”;

Radford et al., “Unsupervised Representation Learning”;

Isola et al., “Image-to-Image Translation”;

Odena et al., “Conditional Image Synthesis”;

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems* (2014): 2672–2680.
- 4** Gervais, “The Machine as Author”;

HOW HUMANS JUDGE MACHINES

Ginsburg and Budiardjo, “Authors and Machines”;

Hristov, “Artificial Intelligence and the Copyright Dilemma.”

5 F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, “Detection of GAN-Generated Fake Images over Social Networks,” in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (IEEE, 2018), 384–389, <https://doi.org/10.1109/MIPR.2018.00084>;

E. Gibney, “The Scientist Who Spots Fake Videos,” *Nature News* (2017), <https://doi.org/10.1038/nature.2017.22784>;

R. Chesney and D. K. Citron, “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,” SSRN (2018), <https://papers.ssrn.com/abstract=3213954>;

S. Tariq, S. Lee, H. Kim, Y. Shin, and S. S. Woo, “Detecting Both Machine and Human Created Fake Face Images In the Wild,” in *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security (MPS '18)* (Association for Computing Machinery, 2018), <https://doi.org/10.1145/3267357.3267367>.

6 F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, “Detection of GAN-Generated Fake Images over Social Networks,” in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (IEEE, 2018), 384–389, <https://doi.org/10.1109/MIPR.2018.00084>;

E. Gibney, “The Scientist Who Spots Fake Videos,” *Nature News* (2017), <https://doi.org/10.1038/nature.2017.22784>;

Tariq et al., “Detecting Both Machine and Human Created Fake Face Images.”

7 S. G. Sripada, E. Reiter, I. Davy, and K. Nilssen, “Lessons from Deploying NLG Technology for Marine Weather Forecast Text Generation,” *Proceedings of PAIS-2004* (2004), 760–764;

K. N. Dörr, “Mapping the Field of Algorithmic Journalism,” *Digital Journalism* 4 (2016): 700–722.

8 A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models Are Unsupervised Multitask Learners” OpenAI Blog 1.8 (2019): 9;

J. Seabrook, “Can a Machine Learn to Write for the New Yorker?,” *New Yorker* (14 October 2019).

HOW HUMANS JUDGE MACHINES

- 9** J. Jermsurawong and N. Habash, “Predicting the Structure of Cooking Recipes,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* 781–786 (Association for Computational Linguistics, 2015), <https://doi.org/10.18653/v1/D15-1090>.
- 10** K. Z. Hu, M. A. Bakker, S. Li, T. Kraska, and C. A. Hidalgo, “VizML: A Machine Learning Approach to Visualization Recommendation,” *arXiv:1808.04819 [cs]* (2018).
- 11** R. L. de Mantaras and J. L. Arcos, “AI and Music: From Composition to Expressive Performance,” *AI Magazine* 23 (2002), 43;
- G. Papadopoulos and G. Wiggins, “AI Methods for Algorithmic Composition: A Survey, a Critical View and Future Prospects” (AISB Symposium on Musical Creativity, 1999);
- B. L. Sturm, J. F. Santos, O. Ben-Tal, and I. Korshunova, “Music Transcription Modelling and Composition Using Deep Learning,” *arXiv:1604.08723 [cs]* (2016).
- 12** E. Francke and B. Alexander, “The Potential Influence of Artificial Intelligence on Plagiarism: A Higher Education Perspective,” in *ECIAIR 2019 European Conference on the Impact of Artificial Intelligence and Robotics* 131 (Academic Conferences and Publishing Limited, 2019);
- A. Elgammal, “AI Is Blurring the Definition of Artist: Advanced Algorithms Are Using Machine Learning to Create Art Autonomously,” *American Scientist* 107 (2019): 18–22;
- D. Lim, “AI & IP: Innovation & Creativity in an Age of Accelerated Change,” *Akron Law Review* 52 (2018): 813.
- 13** Bonnefon et al., “The Social Dilemma of Autonomous Vehicles”;
- Awad et al., “The Moral Machine Experiment”;
- Lin et al., *Robot Ethics 2.0*.
- 14** “Komatsu Outlines Past and Future of Its Autonomous Haulage System,” *International Mining* (2018), <https://im-mining.com/2018/01/29/komatsu-outlines-past-future-autonomous-haulage-system/>.
- 15** J. Vincent, “Self-Driving Truck Convoy Completes Its First Major Journey across Europe,” *The Verge* (2016), <https://www.theverge.com/2016/4/7/11383392/self-driving-truck-platooning-europe>.

HOW HUMANS JUDGE MACHINES

- 16** A. C. Madrigal, “Waymo’s Robots Drove More Miles than Everyone Else Combined,” *Atlantic* (2019), <https://www.theatlantic.com/technology/archive/2019/02/the-latest-self-driving-car-statistics-from-california/582763/>.
- 17** Bonnefon et al., “The Social Dilemma of Autonomous Vehicles”;
Awad et al., “The Moral Machine Experiment”;
Lin, “Why Ethics Matters for Autonomous Cars”;
J.-F. Bonnefon, *La voiture qui en savait trop: L’intelligence artificielle a-t-elle une morale?* (HUMANSCIENCES, 2019);
- 18** Bonnefon et al., “The Social Dilemma of Autonomous Vehicles.”
E. Awad, S. Levine, M. Kleiman-Weiner, S. Dsouza, J. B. Tenenbaum, A. Shariff, et al., “Drivers Are Blamed More than Their Automated Cars When Both Make Mistakes,” *Nature Human Behaviour* (2019): 1–10.
- 19** Awad et al., “The Moral Machine Experiment”;
Awad et al., “Universals and Variations in Moral Decisions Made in 42 Countries.”
- 20** Bonnefon et al., “The Social Dilemma of Autonomous Vehicles.”
- 21** A. Shariff, J.-F. Bonnefon, and I. Rahwan, “Psychological Roadblocks to the Adoption of Self-Driving Vehicles,” *Nature Human Behaviour* 1 (2017): 694–696.
- 22** “Three-Quarters of Americans ‘Afraid’ to Ride in a Self-Driving Vehicle,” *AAA NewsRoom* (2016), <https://newsroom.aaa.com/2016/03/three-quarters-of-americans-afraid-to-ride-in-a-self-driving-vehicle/>.
- 23** Awad et al., “Drivers Are Blamed More than Their Automated Cars.”
- 24** Awad et al., “Drivers Are Blamed More than Their Automated Cars.”
- 25** “Flag-Burning Amendment Fails by a Vote,” CNN.com, 28 June 2006, <http://www.cnn.com/2006/POLITICS/06/27/flag.burning/index.html>.

- 26** J. Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion* (Knopf Doubleday Publishing Group, 2012).

Chapter 3: Judged by Machines

- 1** V. Bilotkach, N. G. Rupp, and V. Pai, *Value of a Platform to a Seller: Case of American Airlines and Online Travel Agencies*, SSRN (2017), <https://papers.ssrn.com/abstract=2321767>;

B. Friedman and H. Nissenbaum, “Bias in Computer Systems,” *ACM Transactions on Information Systems* 14 (1996): 330–347.

- 2** B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, “Face Recognition Performance: Role of Demographic Information,” *IEEE Transactions on Information Forensics and Security* 7 (2012): 1789–1801;

Buolamwini and Gebru, “Gender Shades”;

A. Torralba and A. A. Efros, “Unbiased Look at Dataset Bias,” in *CVPR 2011* (IEEE, 2011): 1521–1528, <https://doi.org/10.1109/CVPR.2011.5995347>.

- 3** M. J. Kusner, J. Loftus, C. Russell, and R. Silva, “Counterfactual Fairness,” in *Advances in Neural Information Processing Systems*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Curran Associates, 2017), 4066–4076.

- 4** J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints,” *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (ACL, 2017), <https://doi.org/10.18653/v1/D17-1323>;

N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, “Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes,” *PNAS* 115 (2018): E3635–E3644;

L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, “Women Also Snowboard: Overcoming Bias in Captioning Models,” in *Computer Vision–ECCV 2018*, eds. V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss (Springer International Publishing, 2018), 793–811;

T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings,” in *Advances in Neural*

HOW HUMANS JUDGE MACHINES

Information Processing Systems 29, eds. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, 2016), 4349–4357.

5 Baeza-Yates, “Data and Algorithmic Bias in the Web.”

6 R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, “Fairness in Criminal Justice Risk Assessments: The State of the Art,” *Sociological Methods & Research* (July 2018), <https://doi.org/10.1177/0049124118782533>;

O. A. Osoba and W. Welsch IV, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence* (Rand Corporation, 2017);

Z. Lin, J. Jung, S. Goel, and J. Skeem, “The Limits of Human Predictions of Recidivism,” *Science Advances* 6 (2020): eaaz0652;

J. Dressel and H. Farid, “The Accuracy, Fairness, and Limits of Predicting Recidivism,” *Science Advances* 4 (2018): eaao5580.

7 A. Lambrecht and C. Tucker, “Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads,” *Management Science* 65 (2019): 2966–2981.

8 J. Koren, “What Does That Web Search Say about Your Credit?,” *Los Angeles Times* 17 July 2016, <https://www.latimes.com/business/la-fi-zestfinance-baidu-20160715-snap-story.html>.

9 M. Kearns and A. Roth, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (Oxford University Press, 2019);

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. A. Galstyan, “Survey on Bias and Fairness in Machine Learning,” *arXiv:1908.09635 [cs]* (2019).

10 Bilotkach et al., *Value of a Platform to a Seller*;

Baeza-Yates, “Data and Algorithmic Bias in the Web”;

M. G. Haselton, D. Nettle, and D. R. Murray, “The Evolution of Cognitive Bias,” in *Handbook of Evolutionary Psychology* 1–20 (American Cancer Society, 2015), <https://doi.org/10.1002/9781119125563.evpsych241>;

HOW HUMANS JUDGE MACHINES

T. M. Mitchell, *The Need for Biases in Learning Generalizations* (1980);

A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics Derived Automatically from Language Corpora Contain Human-Like Biases,” *Science* 356 (2017): 183–186;

S. Hajian, F. Bonchi, and C. Castillo, “Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016), 2125–2126, <https://doi.org/10.1145/2939672.2945386>.

11 Kearns and Roth, *The Ethical Algorithm*;

Mehrabi et al., “Survey on Bias and Fairness in Machine Learning.”

12 Kearns and Roth, *The Ethical Algorithm*;

Mehrabi et al., “Survey on Bias and Fairness in Machine Learning”;

M. Kearns, S. Neel, A. Roth, and S. Wu, “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness,” in *35th International Conference on Machine Learning, ICML 2018* (IMLS, 2018), 4008–4016.

13 M. Hardt, E. Price, and N. Srebro, “Equality of Opportunity in Supervised Learning,” in *Advances in Neural Information Processing Systems* (2016), 3315–3323.

14 Mehrabi et al., “Survey on Bias and Fairness in Machine Learning.”

15 Kearns and Roth, *The Ethical Algorithm*.

16 Bolukbasi et al., “Man Is to Computer Programmer.”

17 Zhao et al., “Men Also Like Shopping”;

Hendricks et al., “Women Also Snowboard”;

Bolukbasi et al., “Man Is to Computer Programmer”;

Caliskan et al., “Semantics Derived Automatically.”

HOW HUMANS JUDGE MACHINES

- 18** Zhao et al., “Men Also Like Shopping”;
Hendricks et al., “Women Also Snowboard”;
Bolukbasi et al., “Man Is to Computer Programmer.”
- 19** Bolukbasi et al., “Man Is to Computer Programmer.”
- 20** M. Turk and A. Pentland, “Eigenfaces for Recognition,” *Journal of Cognitive Neuroscience* 3 (1991), 71–86;
T. Kanade, Y. Tian, and J. F. Cohn, “Comprehensive Database for Facial Expression Analysis,” in *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000* (IEEE Computer Society, 2000), 46;
Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep Learning Face Attributes in the Wild,” (2015), 3730–3738;
O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep Face Recognition,” in *Proceedings of the British Machine Vision Conference 2015* (British Machine Vision Association, 2015), 41.1–41.12, <https://doi.org/10.5244/C.29.41>;
Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep Learning Face Representation by Joint Identification-Verification,” in *Advances in Neural Information Processing Systems 27*, eds. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Curran Associates, 2014), 1988–1996;
Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the Gap to Human-Level Performance in Face Verification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2014), 1701–1708.
- 21** Klare et al., “Face Recognition Performance”;
Buolamwini and Gebru, “Gender Shades.”
- 22** Torralba and Efros, “Unbiased Look at Dataset Bias”;
B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen. et al., “Pushing the Frontiers of Unconstrained Face Detection and Recognition: IARPA Janus Benchmark A,” in

HOW HUMANS JUDGE MACHINES

IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE, 2015), 1931–1939;

G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, “Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments,” *Workshop on Faces in ‘Real-Life’ Images: Detection, Alignment, and Recognition*. Marseille, France (2008).

23 M. Ngan and P. Grother, *Face Recognition Vendor Test (FRVT)—Performance of Automated Gender Classification Algorithms* (2015), <https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8052.pdf>, <https://doi.org/10.6028/NIST.IR.8052>;

I. D. Raji and J. Buolamwini, “Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products,” *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019).

24 Lin et al., “The Limits of Human Predictions of Recidivism”;

Dressel and Farid, “The Accuracy, Fairness, and Limits of Predicting Recidivism.”

25 Electronic Privacy Center, “Algorithms in the Criminal Justice System: Pre-Trial Risk Assessment Tools,” <https://epic.org/algorithmic-transparency/crim-justice/>.

26 J. Angwin, J. Larson, S. Mattu, and L. Kirchner, “Machine Bias,” *ProPublica* (23 May 2016).

27 D. Neumark, R. J. Bank, and K. D. Van Nort, “Sex Discrimination in Restaurant Hiring: An Audit Study,” *Quarterly Journal of Economics* 111 (1996): 915–941;

L. Kaas and C. Manger, “Ethnic Discrimination in Germany’s Labour Market: A Field Experiment,” *German Economic Review* 13 (2012): 1–20;

M. Bertrand and S. Mullainathan, “Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination,” *American Economic Review* 94 (2004): 991–1013;

P. Oreopoulos, “Why Do Skilled Immigrants Struggle in the Labor Market? A Field Experiment with Thirteen Thousand Resumes,” *American Economic Journal: Economic Policy* 3 (2011): 148–171;

D. Neumark, I. Burn, and P. Button, “Experimental Age Discrimination Evidence and the Heckman Critique,” *American Economic Review* 106 (2016): 303–308;

HOW HUMANS JUDGE MACHINES

C. Fershtman and U. Gneezy, "Discrimination in a Segmented Society: An Experimental Approach," *Quarterly Journal of Economics* 116 (2001): 351–377;

E. O. Arceo-Gomez and R. M. Campos-Vazquez, "Race and Marriage in the Labor Market: A Discrimination Correspondence Study in a Developing Country," *American Economic Review* 104 (2014): 376–380.

28 L. Kaas and C. Manger, "Ethnic Discrimination in Germany's Labour Market: A Field Experiment," *German Economic Review* 13 (2012): 1–20;

Bertrand and Mullainathan, "Are Emily and Greg More Employable?";

Oreopoulos, "Why Do Skilled Immigrants Struggle in the Labor Market?";

Neumark et al., "Experimental Age Discrimination Evidence."

29 Arceo-Gomez and Campos-Vazquez, "Race and Marriage in the Labor Market."

30 Dietvorst, "Algorithm Aversion."

31 G. Gigerenzer, *Gut Feelings: The Intelligence of the Unconscious* (Penguin, 2007);

G. Gigerenzer, "How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases,'" *European Review of Social Psychology* 2 (1991): 83–115;

G. Gigerenzer and H. Brighton, "Homo Heuristicus: Why Biased Minds Make Better Inferences," *Topics in Cognitive Science* 1 (2009): 107–143;

A. Tversky and D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science* 185 (1974): 1124–1131;

T. Gilovich, D. Griffin, and D. Kahneman, *Heuristics and Biases: The Psychology of Intuitive Judgment* (Cambridge University Press, 2002);

D. Kahneman, S. P. Slovic, P. Slovic, and A. Tversky, *Judgment under Uncertainty: Heuristics and Biases* (Cambridge University Press, 1982).

32 S. T. Fiske, A. J. C. Cuddy, P. Glick, and J. Xu, "A Model of (Often Mixed) Stereotype Content: Competence and Warmth Respectively Follow from Perceived Status and Competition," *Journal of Personality and Social Psychology* 82 (2002): 878–902;

HOW HUMANS JUDGE MACHINES

S. T. Fiske, A. J. C. Cuddy, and P. Glick, “Universal Dimensions of Social Cognition: Warmth and Competence,” *Trends in Cognitive Sciences* 11 (2007): 77–83.

33 Tversky and Kahneman, “Judgment under Uncertainty: Heuristics and Biases”;

Gilovich et al., *Heuristics and Biases*;

G. Gigerenzer and D. G. Goldstein, “Reasoning the Fast and Frugal Way: Models of Bounded Rationality,” *Psychological Review* 103 (1996): 650;

G. Gigerenzer and P. M. Todd, *Simple Heuristics That Make Us Smart* (Oxford University Press, 1999).

34 J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan, “Algorithmic Fairness,” *AEA Papers and Proceedings* 108 (2018): 22–27.

35 Kleinberg et al., “Algorithmic Fairness.”

36 Kleinberg et al., “Algorithmic Fairness.”

37 Title VII of the Civil Rights Act of 1964: Know Your Rights, *AAUW: Empowering Women since 1881*, <https://www.aauw.org/what-we-do/legal-resources/know-your-rights-at-work/title-vii/>.

38 *Griggs v. Duke Power Company*, Oyez, <https://www.oyez.org/cases/1970/124>.

39 R. Chowdhury and N. Mulani, “Auditing Algorithms for Bias,” *Harvard Business Review* (2018), <https://hbr.org/2018/10/auditing-algorithms-for-bias>.

40 Chowdhury and Mulani, “Auditing Algorithms for Bias”;

J. Guszczka, I. Rahwan, W. Bible, M. Cebrian, and V. Katyal, “Why We Need to Audit Algorithms,” *Harvard Business Review* (2018), <https://hbr.org/2018/11/why-we-need-to-audit-algorithms>.

Chapter 4: In the Eye of the Machine

- 1 “Japanese Hotel Apologizes for Robots That Allowed Video and Sound to Be Hacked,” *Security*, 25 October 2019, <https://www.securitymagazine.com/articles/91157-japanese-hotel-apologizes-for-robots-that-allowed-video-and-sound-to-be-hacked>.
- 2 Lance R. Vick on Twitter: ‘It has been a week, so I am dropping an 0day. The bed facing Tapia robot deployed at the famous Robot Hotels in Japan can be converted to offer anyone remote camera/mic access to all future guests. Unsigned code via NFC behind the head. Vendor had 90 days. They didn’t care. Twitter. <https://twitter.com/lrvick/status/1182823213736161280>.
- 3 John Oates, “Japanese Hotel Chain Sorry That Hackers May Have Watched Guests through Bedside Robots,” *The Register*, https://www.theregister.co.uk/2019/10/22/japanese_hotel_chain_sorry_that_bedside_robots_may_have_watched_guests/.
- 4 C. Dwyer, “Privacy in the Age of Google and Facebook,” *IEEE Technology and Society Magazine* 30 (2011): 58–63;

I. S. Rubinstein and N. Good, “Privacy by Design: A Counterfactual Analysis of Google and Facebook Privacy Incidents,” *Berkeley Technology Law Journal* 28 (2013): 1333;

E. Hargittai, “Facebook Privacy Settings: Who Cares?,” *First Monday* 15 (2010), <https://firstmonday.org/article/view/3086/2589>;

H. Chung, M. Iorga, J. Voas, and S. Lee, “Alexa, Can I Trust You?,” *Computer* 50 (2017): 100–104;

M. B. Hoy, “Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants,” *Medical Reference Services Quarterly* 37 (2018): 81–88;

Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, “Unique in the Crowd: The Privacy Bounds of Human Mobility,” *Scientific Reports* 3 (2013), 1376;

M. Z. Yao, R. E. Rice, and K. Wallis, “Predicting User Concerns about Online Privacy,” *Journal of the American Society for Information Science and Technology* 58 (2007): 710–722;

M. G. Hoy and G. Milne, “Gender Differences in Privacy-Related Measures for Young Adult Facebook Users,” (2010): 28–45.

HOW HUMANS JUDGE MACHINES

- 5 James Condliffe, “Chinese Cops Are Wearing Glasses That Can Recognize Faces,” *MIT Technology Review* (7 February 2018), <https://www.technologyreview.com/f/610214/chinese-cops-are-using-facial-recognition-specs/>.
- 6 P. Mozur, “In Hong Kong Protests, Faces Become Weapons,” *New York Times*, 26 July 2019.
- 7 L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye, “Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models,” *Nature Communications* 10 (2019): 1–9.
- 8 M. Kearns and A. Roth, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. (Oxford University Press, 2019);

“Advice to My Younger Self: Latanya Sweeney,” *Ford Foundation*, <https://www.fordfoundation.org/ideas/equal-change-blog/posts/advice-to-my-younger-self-latanya-sweeney/>.
- 9 de Montjoye et al., “Unique in the Crowd.”
- 10 L. Sweeney, “K-Anonymity: A Model for Protecting Privacy,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (2002): 557–570;

L. Sweeney, “Achieving K-Anonymity Privacy Protection Using Generalization and Suppression,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (2002): 571–588.
- 11 M. Kearns and A. Roth, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (Oxford University Press, 2019).
- 12 Kearns and Roth, *The Ethical Algorithm*; A. Hern, “Fitness Tracking App Strava Gives Away Location of Secret US Army Bases,” *The Guardian*, 28 January 2018.
- 13 Kearns and Roth, *The Ethical Algorithm*.
- 14 C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating Noise to Sensitivity in Private Data Analysis,” in *Theory of Cryptography*, eds. S. Halevi and T. Rabin (Springer Berlin Heidelberg, 2006): 265–284;

HOW HUMANS JUDGE MACHINES

C. Dwork, “Differential Privacy: A Survey of Results,” in *International Conference on Theory and Applications of Models of Computation* (Springer, 2008), 1–19.

15 Kearns and Roth, *The Ethical Algorithm*.

16 S. L. Warner, “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias,” *Journal of the American Statistical Association* 60 (1965): 63–69.

17 Kearns and Roth, *The Ethical Algorithm*.

18 Ú. Erlingsson, V. Pihur, and A. Korolova, “Rappor: Randomized Aggregatable Privacy-Preserving Ordinal Response,” in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2014): 1054–1067.

19 G. J. Lensvelt-Mulders, J. J. Hox, P. G. Van der Heijden, and C. J. Maas, “Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation,” *Sociological Methods & Research* 33 (2005): 319–348.

20 J. A. Landsheer, P. Van Der Heijden, and G. Van Gils, “Trust and Understanding, Two Psychological Aspects of Randomized Response,” *Quality and Quantity* 33 (1999): 1–12.

21 Erlingsson et al., “Rappor”;

K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, et al. “Practical Secure Aggregation for Privacy-Preserving Machine Learning,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2017), 1175–1191;

N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, “Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data,” *Proceedings of the 5th International Conference on Learning Representation* (ICLR, 2016);

R. C. Geyer, T. Klein, and M. Nabi, “Differentially Private Federated Learning: A Client Level Perspective,” *arXiv:1712.07557 [cs, stat]* (2017);

V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, “Federated Multi-task Learning,” in *Advances in Neural Information Processing Systems* (2017): 4424–4434;

P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, “Split Learning for Health: Distributed Deep Learning without Sharing Raw Patient Data,” *arXiv preprint arXiv:1812.00564* (2018).

HOW HUMANS JUDGE MACHINES

- 22 Lenvelt-Mulders et al., “Meta-Analysis of Randomized Response Research”;
Landsheer et al., “Trust and Understanding.”

Chapter 5: Working Machines

- 1 J. Kantor, “Working Anything but 9 to 5,” *New York Times*, 13 August 2014. <https://www.nytimes.com/interactive/2014/08/13/us/starbucks-workers-scheduling-hours.html>.
- 2 J. Kantor, “Times Article Changes a Starbucks Policy, Fast,” *New York Times*, 22 August 2014. <https://www.nytimes.com/times-insider/2014/08/22/times-article-changes-a-policy-fast/>.
- 3 M. Roosevelt, “Erratic Hours Are the Norm for Workers in Retailing. Can Los Angeles Buck the Trend?” *Los Angeles Times*, 2 March 2019. <https://www.latimes.com/business/la-fi-retail-scheduling-20190302-story.html>.
- 4 E. Brynjolfsson and A. McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (W. W. Norton & Company, 2016);
C. B. Frey and M. A. Osborne, “The Future of Employment: How Susceptible Are Jobs to Computerisation?,” *Technological Forecasting and Social Change* 114 (2017): 254–280;
J. Borland and M. Coelli, “Are Robots Taking Our Jobs?,” *Australian Economic Review* 50 (2017): 377–397;
W. E. Forum, “The Future of Jobs: Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution,” in *Global Challenge Insight Report*, World Economic Forum, Geneva (2016);
A. Smith and J. Anderson, “AI, Robotics, and the Future of Jobs,” *Pew Research Center* 6 (2014), <https://www.pewresearch.org/internet/2014/08/06/future-of-jobs/>.
- 5 J. Mokyry, C. Vickers, and N. L. Ziebarth, “The History of Technological Anxiety and the Future of Economic Growth: Is This Time Different?,” *Journal of Economic Perspectives* 29 (2015): 31–50;

HOW HUMANS JUDGE MACHINES

D. H. Autor, "Why Are There Still So Many Jobs? The History and Future of Workplace Automation," *Journal of Economic Perspectives* 29 (2015): 3–30.

6 C. Jara-Figueroa, A. Z. Yu, and C. A. Hidalgo, "How the Medium Shapes the Message: Printing and the Rise of the Arts and Sciences," *PLOS One* 14 (2019): e0205771.

7 E. L. Eisenstein, *The Printing Press as an Agent of Change: Communications and Cultural Trans* (Cambridge University Press, 1980).

8 Mokyr et al., "The History of Technological Anxiety."

9 *Time*, "The Automation Jobless," 24 February 1961, <http://content.time.com/time/magazine/0,9263,7601610224,00.html>.

10 Frey and Osborne, "The Future of Employment."

11 "The Four Global Forces Breaking All the Trends," McKinsey, <https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-four-global-forces-breaking-all-the-trends>.

12 Mokyr et al., "The History of Technological Anxiety";

Autor, "Why Are There Still So Many Jobs?";

D. Autor and A. Salomons, "Is Automation Labor Share-Displacing? Productivity Growth, Employment, and the Labor Share," *Brookings Papers on Economic Activity* (2018): 1–87;

M. Arntz, T. Gregory, and U. Zierahn, "The Risk of Automation for Jobs in OECD Countries," *OECD Social, Employment and Migration Working Papers*, No. 189 (2016), <https://doi.org/10.1787/5jlz9h56dvq7-en>.

13 H. J. Wilson and P. R. Daugherty, "Creating the Symbiotic AI Workforce of the Future," *MIT Sloan Management Review*, 21 October 2019, <https://sloanreview.mit.edu/article/creating-the-symbiotic-ai-workforce-of-the-future/>.

14 Autor, "Why Are There Still So Many Jobs?";

D. Acemoglu and P. Restrepo, "Artificial Intelligence, Automation and Work," NBER Working Paper no. 24196 (2018), <http://www.nber.org/papers/w24196>.

HOW HUMANS JUDGE MACHINES

- 15** Autor, “Why Are There Still So Many Jobs?”
- 16** Autor, “Why Are There Still So Many Jobs?”
- 17** Frey and Osborne, “The Future of Employment.”
- 18** Arntz et al., “The Risk of Automation for Jobs in OECD Countries.”
- 19** Mokyr et al., “The History of Technological Anxiety”;
Autor, “Why Are There Still So Many Jobs?”;
Autor and Salomons, “Is Automation Labor Share-Displacing?”;
Arntz et al., “The Risk of Automation for Jobs in OECD Countries”;
Acemoglu and Restrepo, “Artificial Intelligence, Automation and Work.”
- 20** Graetz and G. Michaels, “Robots at Work.” *Review of Economics and Statistics* 100 (2018): 753–768.
- 21** L. Barbieri, C. Mussida, M. Piva, and M. Vivarelli, “Testing the Employment Impact of Automation, Robots and AI: A Survey and Some Methodological Issues,” Institute for the Study of Labor (IZA) Research Paper 12612 (2019), <https://papers.ssrn.com/abstract=3457656>;
K. De Backer, T. DeStefano, C. Menon, and J. R. Suh, “Industrial Robotics and the Global Organisation of Production,” OECD Science, Technology and Industry Working Papers 3 (2018), https://www.oecd-ilibrary.org/industry-and-services/industrial-robotics-and-the-global-organisation-of-production_dd98ff58-en.
- 22** Frey and Osborne, “The Future of Employment.”
- 23** Arntz et al., “The Risk of Automation for Jobs in OECD Countries.”
- 24** A. Agrawal, J. Gans, and A. Goldfarb, *Prediction Machines: The Simple Economics of Artificial Intelligence* (Harvard Business Review Press, 2018).
- 25** G. Gereffi, J. Humphrey, and T. Sturgeon, “The Governance of Global Value Chains,” *Review of International Political Economy* 12 (2005): 78–104;

HOW HUMANS JUDGE MACHINES

J. Humphrey and H. Schmitz, “How Does Insertion in Global Value Chains Affect Upgrading in Industrial Clusters?,” *Regional Studies* 36 (2002): 1017–1027.

26 P.-A. Balland, C. Jara-Figueroa, S. G. Petralia, M. P. A. Steijn, D. L. Rigby, and C. A. Hidalgo, “Complex Economic Activities Concentrate in Large Cities,” *Nature Human Behaviour* 4 (2020): 1–7, <https://doi.org/10.1038/s41562-019-0803-3>.

27 M. Roser and E. Ortiz-Ospina, “Global Education—Our World in Data,” Our World in Data (2016), <https://ourworldindata.org/global-education>.

28 H. Rapoport, “Migration and Globalization: What’s in It for Developing Countries?,” *International Journal of Manpower* 37 (2016): 1209–1226.

29 R. Abbott and B. Bogenschneider, “Should Robots Pay Taxes: Tax Policy in the Age of Automation,” *Harvard Law and Policy Review* (2018): 145–176.

30 A. McAfee and E. Brynjolfsson, “Human Work in the Robotic Future: Policy for the Age of Automation Essays,” *Foreign Affairs* (2016): 139–150.

31 C. Jara-Figueroa, B. Jun, E. L. Glaeser, and C. A. Hidalgo, “The Role of Industry-Specific, Occupation-Specific, and Location-Specific Knowledge in the Growth and Survival of New Firms,” *PNAS* 115 (2018): 12646–12653.

32 S. D. Harris and A. B. Krueger, *A Proposal for Modernizing Labor Laws for Twenty-First-Century Work: The “Independent Worker”* (Hamilton Project, Brookings, 2015).

33 A Brief History of Basic Income Ideas, <https://ubi-europe.net/ubi/brief-history-basic-income-ideas/>.

34 McAfee and Brynjolfsson, “Human Work in the Robotic Future.”

35 A. Stern, *Raising the Floor: How a Universal Basic Income Can Renew Our Economy and Rebuild the American Dream* (PublicAffairs, 2016);

U. Colombino, “Is Unconditional Basic Income a Viable Alternative to Other Social Welfare Measures?” *IZA World of Labor* (2019), <https://wol.iza.org/uploads/articles/128/pdfs/is-unconditional-basic-income-viable-alternative-to-other-social-welfare-measures.pdf>;

HOW HUMANS JUDGE MACHINES

Matt Stevens and Isabella Grullón Paz, “Andrew Yang’s \$1,000-a-Month Idea May Have Seemed Absurd Before. Not Now,” *New York Times*, 18 March 2020, <https://www.nytimes.com/2020/03/18/us/politics/universal-basic-income-andrew-yang.html>

Chapter 6: Moral Functions

- 1 C. Efferson, R. Lalive, and E. Fehr, “The Coevolution of Cultural Groups and Ingroup Favoritism,” *Science* 321 (2008): 1844–1849;

M. Hewstone, M. Rubin, and H. Willis, “Intergroup Bias,” *Annual Review of Psychology* 53 (2002): 575–604;

T. Mussweiler and A. Ockenfels, “Similarity Increases Altruistic Punishment in Humans,” *PNAS* 110 (2013): 19318–19323.

- 2 J. Haidt, S. H. Koller, and M. G. Dias, “Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?” *Journal of Personality and Social Psychology* 65 (1993): 613–628;

J. Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion* (Knopf Doubleday Publishing Group, 2012).

- 3 F. Cushman, “Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment,” *Cognition* 108 (2008): 353–380;

F. Cushman, R. Sheketoff, S. Wharton, and S. Carey, “The Development of Intent-Based Moral Judgment,” *Cognition* 127 (2013): 6–21;

B. F. Malle and J. Knobe, “The Folk Concept of Intentionality,” *Journal of Experimental Social Psychology* 33 (1997): 101–121;

L. Young and R. Saxe, “When Ignorance Is No Excuse: Different Roles for Intent across Moral Domains,” *Cognition* 120 (2011): 202–214;

J. D. Greene, F. A. Cushman, L. E. Stewart, K. Lowenberg, L. E. Nystrom, and J. D. Cohen, “Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment,” *Cognition* 111 (2009): 364–371.

Chapter 7: Liable Machines

- 1 I. Asimov, *I, Robot*, Robot series (Bantam Books, 1950).
- 2 A. Jobin, M. Ienca, and E. Vayena, “The Global Landscape of AI Ethics Guidelines,” *Nature Machine Intelligence* 1 (2019): 389–399;

V. Dignum, “Ethics in Artificial Intelligence: Introduction to the Special Issue,” *Ethics and Information Technology* 20 (2018): 1–3.
- 3 “AI at Google: Our Principles,” Google (2018), <https://www.blog.google/technology/ai/ai-principles/>.
- 4 European Commission, “Ethics Guidelines for Trustworthy AI,” Digital Single Market—European Commission (2019), <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.
- 5 “OECD Principles on Artificial Intelligence,” Organisation for Economic Co-operation and Development (OECD), <https://www.oecd.org/going-digital/ai/principles/>.
- 6 K. Gödel, “Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I,” *Monatshefte für mathematik und physik* 38 (1931): 173–198;

J. Gleick, *The Information: A History, a Theory, a Flood* (Vintage, 2012).
- 7 P. M. Asaro, “A Body to Kick, But Still No Soul to Damn: Legal Perspectives on Robotics,” in *Robot Ethics*, eds. P. Lin, K. Abney, and G. A. Bekey (MIT Press, 2011), 169–186.
- 8 Asaro, “A Body to Kick.”
- 9 Asaro, “A Body to Kick.”
- 10 Asaro, “A Body to Kick.”
- 11 D. J. Calverley, “Android Science and Animal Rights, Does an Analogy Exist?,” *Connection Science* 18 (2006): 403–417;

E. Schaerer, R. Kelley, and M. Nicolescu, “Robots as Animals: A Framework for Liability and

HOW HUMANS JUDGE MACHINES

Responsibility in Human-Robot Interactions,” in *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication* (IEEE, 2009), 72–77.

- 12 M. Weber, *The Theory of Social and Economic Organization* (Simon and Schuster, 2009).
- 13 Hidalgo C. A. A Bold Idea to Replace Politicians. TED (2018), https://www.ted.com/talks/cesar_hidalgo_a_bold_idea_to_replacepoliticians?language=en.
- 14 S. Thrun and L Pratt, *Learning to Learn* (Springer Science & Business Media, 2012);
J. B. Tenenbaum, “Bayesian Modeling of Human Concept Learning,” in *Advances in Neural Information Processing Systems*, 59–68 (1999);
J. Feldman, “The Structure of Perceptual Categories,” *Journal of Mathematical Psychology* 41 (1997): 145–170.
- 15 T. M. Mitchell, R. M. Keller, and S. T. Kedar-Cabelli, “Explanation-Based Generalization: A Unifying View,” *Machine Learning* 1 (1986): 47–80.