Reading Assignments for
Ethics for AI and Robotics
(EECS 598/498 and ROB 599)

Benjamin Kuipers

Winter 2023

Draft: March 23, 2023

Recent progress in AI and Robotics has made the problem of Ethics increasingly urgent. Many people have started to approach many different aspects of this problem, but there is little consensus (yet) about the right questions, or their answers. Everyone participating in this class will help to formulate these questions and answers.

Some readings will be at URLs supplied. Some, with citations like [LeGuin-omelas-73] can be found in Canvas > Files/readings or Canvas > Files/news. Others, with citations like [Singer, 1981], are journal articles or books you can find in the library, and the full reference is at the end.

0 Course Introduction (1/4/23)

- **Reading Assignment** (read before class)
  Course syllabus: https://web.eecs.umich.edu/~kuipers/teaching/eecs598-W23.html
  "The Blind Men and the Elephant": https://web.eecs.umich.edu/~kuipers/opinions/Elephant.html
  ACM Code of Ethics and Professional Conduct: https://www.acm.org/code-of-ethics
  IEEE Code of Ethics: https://www.ieee.org/about/corporate/governance/p7-8.html

- **Other Readings** (valuable for many purposes)
  7000-2021 - IEEE Standard Model Process for Addressing Ethical Concerns during System Design: https://ieeexplore.ieee.org/document/9536679
  Consider other suggested readings on the syllabus web page.

1 Foundations

1.1 Philosophical Ethics (1/9/23)

- **Reading Assignment** (read before class)
  In the Stanford Encyclopedia of Philosophy (SEP), read enough from each of these articles to get the general idea, and to be able to go back for more depth as needed. The SEP is a valuable reference.

  Utilitarianism (https://plato.stanford.edu/entries/utilitarianism-history/)
  Consequentialism (https://plato.stanford.edu/entries/consequentialism/)
  Deontology (https://plato.stanford.edu/entries/ethics-deontological/)
  Virtue Ethics (https://plato.stanford.edu/entries/ethics-virtue/)
  Contractarianism (https://plato.stanford.edu/entries/contractarianism/)
  Contractualism (https://plato.stanford.edu/entries/contractualism/)

- **Other Readings** (valuable for many purposes)
  "Those who walk away from Omelas" by Ursula LeGuin [LeGuin-omelas-73].

Peter Singer, *The Expanding Circle* [Singer, 1981]
John Rawls, *A Theory of Justice* [Rawls, 1999]

1.2 The Prisoner's Dilemma (1/11/23)

- **Reading Assignment** (read before class)
  Leyton-Brown & Shoham, *Essentials of Game Theory*, ch.1-2 [Leyton-Brown+Shoham-08-ch.1-2]
  Anatol Rapaport, The use and misuse of game theory [Rapaport-sciam-62]

- **Other Readings** (valuable for many purposes)
  Axelrod & Hamilton, The evolution of cooperation [Axelrod-science-81]
  Robert Axelrod, *The Evolution of Cooperation* [Axelrod, 1984]
  . . . and more useful references in the slides.

1.3 Ethics, Trust, and Cooperation (1/18/23)

- **Reading Assignment** (read before class)
  Kuipers, Trust and Cooperation https://web.eecs.umich.edu/ kuipers/research/pubs/Kuipers-frai-22.html
  Mayer, Davis & Schoorman, An integrative model of organizational trust. [Mayer et al., 1995]

- **Other Readings** (valuable for many purposes)
  Ethics Guidelines for Trustworthy AI [on AI, 2019]
  Rousseau, et al, Not so different after all . . . [Rousseau et al., 1998]
  Lee & See, Trust in automation [Lee and See, 2004]
  Jeannette Wing, Trustworthy AI [Wing, 2021]
  Sreedhar & Gopal, Behind low vaccination rates . . . [Sreedhar-nyt-12-3-21]

1.4 Evolutionary Origins (1/23/23)

- **Reading Assignment** (read before class)
  Tomasello, et al, Two key steps in the evolution of human cooperation. *Current Anthropology*, 2012. [Tomasello et al., 2012]
  Boyd, Richerson & Henrich, The cultural niche: Why social learning is essential for human adaptation. *Proc. Nat. Acad. Sci. (PNAS)*. 2011. [Boyd et al., 2011]

- **Other Readings** (valuable for many purposes)
  Henrich, *The Secret of Our Success*, 2016. [Henrich, 2016]
  Henrich, et al, Markets, religion, community size, and the evolution of fairness and punishment. *Science*, 2010. [Henrich et al., 2010]
  Rand, et al, Social heuristics shape intuitive cooperation. *Nature Communication*, 2014. [Rand et al., 2014]
  Henrich & Muthukrishna, The origins and psychology of human cooperation. *Annual Review of Psychology*, 2021. [Henrich and Muthukrishna, 2021]

2 Safety and Autonomous Vehicles

2.1 Why should we build autonomous vehicles?

- **Reading Assignment** (read before class)
  NHTSA, Critical reasons for crashes . . . , 2018.
  https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812506
  SAE autonomy levels [SAE J3016 levels 5-21]
  Waymo Safety Report, September 2021. https://waymo.com/safety

- **Other Readings** (valuable for many purposes)
  NHTSA Voluntary Safety Self-Assessments, https://www.nhtsa.gov/automated-driving-systems/voluntary-safety-self-assessment
  Rodney Brooks, Blog. https://rodneybrooks.com/blog/

2.2 Moral dilemmas for autonomous vehicles

- **Reading Assignment** (read before class)
  Awad et al, The Moral Machine experiment. [Awad-nature-18]
  Kuipers, Perspectives on Ethics of AI: Computer Science.
  (https://web.eecs.umich.edu/ kuipers/research/pubs/Kuipers-oheai-20.html)

- **Other Readings** (valuable for many purposes)
  Judith Jarvis Thomson, The Trolley Problem. [Thomson-ylj-85]
  Philippa Foot, The problem of abortion and the doctrine of double effect. [Foot-or-67]
  Bonnefon et al, The social dilemma of autonomous vehicles. [Bonnefon-science-16]
  Awad et al, Crowdsourcing moral machines. [Awad-cacm-20]

2.3 AVs and regulations: Widen & Koopman guest lecture

- **Reading Assignment** (read before class)
  Koopman, SAE J3016 Users Guide, 2021. https://users.ece.cmu.edu/ koopman/j3016/
  Widen & Koopman, Autonomous vehicle regulation & trust, *UCLA J. Law & Technology*, 2022. [Widen and Koopman, 2022]
  Koopman & Widen, Ethical design & testing of automated driving features [Koopman-ssrn-1-24-23]

- **Other Readings** (valuable for many purposes)
  https://en.wikipedia.org/wiki/Self-driving_car

2.4 What are the actual ethical issues for AVs?

- **Reading Assignment** (read before class)
  Cade Metz, NYTimes: [Metz-nyt-12-7-21] [Metz-nyt-6-8-22] [Metz-nyt-2-1-23]
  Rodney Brooks, roboticist: (just read the predictions about self-driving cars)
  https://rodneybrooks.com/edge-cases-for-self-driving-cars/
  https://rodneybrooks.com/my-dated-predictions/ (2018)
  https://rodneybrooks.com/predictions-scorecard-2023-january-01/

- **Other Readings** (valuable for many purposes)
  http://rodneybrooks.com/predictions-scorecard-2019-january-01/
  http://rodneybrooks.com/predictions-scorecard-2020-january-01/
  http://rodneybrooks.com/predictions-scorecard-2021-january-01/
  https://rodneybrooks.com/predictions-scorecard-2022-january-01/

3 Surveillance and Privacy

3.1 Guest lecture: Florian Schaub, UM SI

- **Reading Assignment** (read before class)
  Schaub, Balebako & Cranor, Designing effective privacy notices and controls. [Schaub et al., 2017].
  Longer version: [Schaub-iptp-20].
  Acquisti, Brandimarti & Loewenstein, Secrets and Likes: The drive for privacy and the difficulty of achieving it in the digital age, 2020 [Acquisti et al., 2020].[1]

- **Other Readings** (valuable for many purposes)
  Harkous, et al, Polisis:Automated analysis and presentation of privacy policies using deeplearning. [Harkous et al., 2018]
  Kumar, et al, Finding a choice in a haystack: Automating extraction of opt-out statementsfrom privacy policy text. [Kumar-www-20]

3.2 Surveillance: Balancing the Good and the Bad

- **Reading Assignment** (read before class)
  VanBavel, et al, How social media shapes polarization. [VanBavel-tics-21]
  Rathje, et al, Out-group animosity drives engagement on social media. [Rathje-pnas-21]

- **Other Readings** (valuable for many purposes)
  Karen Hao, How Facebook got addicted to spreading misinformation. [Hao-tr-21]
  Ben Smith, Inside the information wars. [Smith-nyt-11-28-21]
  https://www.nytimes.com/series/new-york-times-privacy-project

3.3 How comprehensive is individual surveillance?

- **Reading Assignment** (read before class)
  NYT Editorial, Total surveillance is not what America signed up for. [NYT-Editorial-12-21-19]
  https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html
  Steinberger, Does Palantir see too much? [Steinberger-nyt-10-21-20]
  Stark, Facial recognition is the plutonium of AI. [Stark-xrds-19]
  Kashmir Hill, The secretive company that might end privacy as we know it. [Hill-nyt-1-18-20]
  Blatt, Some observations on the Clearview AI facial recognition system – from someone who has actually used it. [Blatt-cpomag-20]
  Kashmir Hill, Your face is not your own. [Hill-nyt-3-21-21]

[1] I can't provide you with a PDF copy of this paper, but you can read the paper online through the University Library's online journal collection.

- **Other Readings** (valuable for many purposes)
  Valentino, Your apps know where you were last night, and they're not keeping it secret. [Valentino-nyt-12-10-18]
  Warzel & Thompson, They stormed the Capitol. Their apps tracked them. [Warzel-nyt-2-5-21]
  Arthur Michel, There are spying eyes everywhere – and now they share a brain [Palantir]. [Michel-wired-2-4-21]

3.4 Surveillance capitalism

- **Reading Assignment** (read before class)
  Shoshana Zuboff, How Google discovered the value of surveillance. [Zuboff-longreads-19]
  Zuboff, Big other: surveillance capitalism and the prospect for an information civilization. [Zuboff, 2015]

- **Other Readings** (valuable for many purposes)
  https://safecomputing.umich.edu/privacy/history-of-privacy-timeline
  Shoshana Zuboff, *The Age of Surveillance Capitalism*, 2019. [Zuboff, 2019]
  Zuboff, You are now remotely controlled. [Zuboff-nyt-1-24-20]
  Zuboff, The coup we are not talking about. [Zuboff-nyt-1-29-21]
  Zuboff, You are the object of a secret extraction operation. [Zuboff-nyt-11-12-21]
  Bruce Schneier, Banning facial recognition isn't enough. [Schneier-nyt-1-20-20]

3.5 Regulating surveillance

- **Reading Assignment** (read before class)
  Helen Nissenbaum, A contextual approach to privacy online. [Nissenbaum, 2011].
  https://www.nytimes.com/series/new-york-times-privacy-project

- **Other Readings** (valuable for many purposes)
  Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life*, 2010.
  Gajda, *Seek and Hide: The Tangled History of the Right to Privacy* [Gajda, 2022].
  Isaac & Hsu, Meta plans to remove thousands of sensitive ad-targeting categories. [Isaac-nyt-11-9-21]
  ONeill, How facial recognition makes you safer. [ONeill-nyt-6-9-19]
  Friedman, China's bullying is becoming a danger to the world and itself. [Friedman-nyt-10-19-21]
  Mueller & Castro, The value of personalized advertising in Europe. [Mueller-cdi-21]
  Frank, The economic case for regulating social media. [Frank-nyt-2-11-21]

4 Bias and Fairness

4.1 Algorithmic bias

- **Reading Assignment** (read before class)
  Buolamwini & Gebru, Gender Shades: Intersectional accuracy disparities in commercial gender classification. [Buolamwini-fat*-18]
  Obermeyer, et al, Dissecting racial bias in an algorithm used to manage the health of populations. [Obermeyer-science-19]

- **Other Readings** (valuable for many purposes)
  Charette, Michigan's MiDAS unemployment system: Algorithm alchemy created lead, not gold. [Charette-spectrum-18]
  Raji, et al, Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. [Raji-aies-19]
  Raji, et al, Saving face: Investigating the ethical concerns of facial recognition auditing. [Raji-aies-20]
  "Face recognition performance . . . " [Klare et al., 2012].
  "Investigating bias in facial analysis systems: A systematic review, 2020" [Khalil et al., 2020].
  Barocas, Hardt & Narayanan, chap.1. https://fairmlbook.org/pdf/introduction.pdf

4.2 Formalizing Fairness

- **Reading Assignment** (read before class)
  Chouldechova, Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. [Chouldechova, 2017]
  Kleinberg, et al, Inherent trade-offs in the fair determination of risk scores. ITCS, 2017. [Kleinberg et al., 2017]

- **Other Readings** (valuable for many purposes)
  https://en.wikipedia.org/wiki/COMPAS_(software)
  "Machine Bias", *ProPublica*, 5-23-2016 [Angwin et al., 2016].
  "How we analyzed . . . ", *ProPublica*, 5-23-2016 [Larson et al., 2016].
  "COMPAS risk scales", *Northpointe, Inc.*, 7-8-2016 [Dieterich et al., 2016].
  "Bias in criminal risk scores", *ProPublica*, 12-30-2016 [Angwin and Larson, 2016].
  Barocas, Hardt & Narayanan, chap.2. https://fairmlbook.org/pdf/classification.pdf

4.3 Guest lecture: H. V. Jagadish, UM CSE

- **Reading Assignment** (read before class; available in Canvas > Files/readings/)
  Rodolfa, et al, *Nature Machine Intelligence*, 2021 [Rodolfa et al., 2021].
  Bolukbasi, et al, "Man is to Computer Prorgrammer as Woman is to Homemaker?" [Bolukbasi et al., 2016]

- **Other Readings** (valuable for many purposes)

4.4 Can trustworthy fairness be achieved?

- **Reading Assignment** (read before class)
  Lee, et al, Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Brookings report, 2019. [Lee-brookings-19]
  Comment collectively on these blog posts on bias from companies:
  https://www.toptal.com/artificial-intelligence/mitigating-ai-bias
  https://www.weforum.org/agenda/2022/10/open-source-data-science-bias-more-ethical-ai-technology/
  https://www.thomsonreuters.com/en-us/posts/legal/combating-ai-bias/
  https://www.oliverwyman.com/our-expertise/insights/2023/feb/manage-ai-bias-instead-of-trying-to-eliminate-it.html
  https://www.boozallen.com/s/insight/blog/algorithmic-bias.html

- **Other Readings** (valuable for many purposes)
  Relevant stories, science fiction and other:
  Gordon R. Dickson, Computers Don't Argue. *Analog Science Fiction*, September 1965.
  https://www.atariarchives.org/bcc2/showpage.php?page=133
  Kurt Vonnegut, Harrison Bergeron. [Harrison Bergeron.rtf]
  The Orange Story. http://www.mediationtools.com/articles/smbj9605.html

5 Jobs, Automation, and Existential Threats

5.1 The future of work

- **Reading Assignment** (read before class)
  What can machine learning do? [Brynjolfsson and Mitchell, 2017]
  Evaluating revolutions in AI [Forbus, 2021]

- **Other Readings** (valuable for many purposes)
  Brynjolfsson & McAfee, *The Second Machine Age* [Brynjolfsson and McAfee, 2014].
  Martin Ford, *The Rise of the Robots* [Ford, 2015].
  The Work of the Future [Autor-mittfwork-20]
  One day of employment a week is all we need for mental health benefits.
  https://www.sciencedaily.com/releases/2019/06/190618192030.htm
  Soon a robot will be writing this headline (NYT Book Review)
  https://www.nytimes.com/2020/01/14/books/review/a-world-without-work-daniel-susskind.html
  Can child care be a big business? Private equity thinks so. [Goldstein-nyt-12-16-22]
  Why you can't find child care. 100,000 workers are missing. [Goldstein-nyt-10-13-22]
  How other nations pay for child care. The U.S. is an outlier. [Miller-nyt-10-6-21]
  Policymakers used to ignore child care. Then came the pandemic. [Peck-nyt-5-9-21]
  "Would you let a robot take care of your Mom?" [Jackson-nyt-12-13-19]
  "The future of robot caregivers" [Aronson-nyt-7-19-14]

5.2 Economic inequality

- **Reading Assignment** (read before class)
  Kuipers, Perspectives on Ethics of AI: Computer Science. (Example 3; follow footnotes)
  https://web.eecs.umich.edu/ kuipers/research/pubs/Kuipers-oheai-20.html.
  McWilliams, "This political theorist predicted the rise of Trumpism. His name was Hunter S. Thompson." *The Nation*, 2016. [McWilliams, 2016]

- **Other Readings** (valuable for many purposes)
  [Leonhardt, 2019]
  [Appelbaum, 2019]
  [Edsall, 2021]
  [Sorkin, 2019]

5.3 Is superintelligent AI an existential threat?

- **Reading Assignment** (read before class)
  Vernor Vinge, The Technological Singularity, 1993/2003. [Vinge-wer-03]
  Hadfield-Menell, Dragan, Abbeel, Russell, Cooperative inverse reinforcement learning. *NIPS*, 2016. [Hadfield-Menell-nips-16]

- **Other Readings** (valuable for many purposes)
  Tim Urban, The AI revolution: the road to superintelligence, 2015.
  https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html
  Tim Urban, The AI revolution: our immortality or extinction, 2015.
  https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html
  D. Hadfield-Menell and G. K. Hadfield, Incomplete contracting and AI alignment. *AIES*, 2019. [Hadfield-Menell-aies-19]
  Dylan Hadfield-Menell, The principal-agent alignment problem in artificial intelligence. PhD thesis, EECS, UC Berkeley, 2021. [Hadfield-Menell-phd-21]

5.4 Corporations as intelligent agents

- **Reading Assignment** (read before class)
  Kuipers, An existing, ecologically-successful genus of collectively intelligent artificial creatures, Collective Intelligence, 2012. https://web.eecs.umich.edu/ kuipers/research/pubs/Kuipers-ci-12.html
  Milton Friedman, The social responsibility of business is to increase its profits. [Friedman-nytmag-70]

- **Other Readings** (valuable for many purposes)
  Richard Danzig, Machines, Bureaucracies and Markets as AIs. [Danzig-cset-22].
  Business Roundtable on Corporate Governance (8-19-2019)
  https://www.businessroundtable.org/business-roundtable-redefines-the-purpose-of-a-corporation-to-promote-an-economy-that-serves-all-americans
  https://opportunity.businessroundtable.org/ourcommitment/

6 Regulation of AI

6.1 Can we / Should we regulate AI?

- **Reading Assignment** (read before class)
  Brundage, et al, Toward trustworthy AI development: Mechanisms for supporting verifiable claims. Executive summary, Sections 1 & 5, the rest as needed. http://www.towardtrustworthyai.com
  Jobin, et al, The global landscape of AI ethics guidelines. [Jobin-nmi-19]

- **Other Readings** (valuable for many purposes)
  Leqi, Hadfield-Menell, Lipton, When curation becomes creation. CACM 64(12): 44-47, December 2021. [Leqi-cacm-21]

6.2 Guest lecture: Prof. Dan Crane, UM Law

- **Reading Assignment** (read before class)
- **Other Readings** (valuable for many purposes)

6.3 Guest lecture: Prof. Jerry Davis, UM Ross

- **Reading Assignment** (read before class)
  Gerald F. Davis, *Taming Corporate Power in the 21st Century*, 2022, chapters 1-2.
  Gerald F. Davis, *Taming Corporate Power in the 21st Century*, 2022, chapters 3-4.
- **Other Readings** (valuable for many purposes)
  Gerald F. Davis, *Taming Corporate Power in the 21st Century*, 2022, chapters 5-9.

6.4 Flex and surge

- **Reading Assignment** (read before class)
- **Other Readings** (valuable for many purposes)

6.5 Flex and surge

- **Reading Assignment** (read before class)
- **Other Readings** (valuable for many purposes)

References

[Acquisti et al., 2020] Acquisti, A., Brandimarti, L., and Loewenstein, G. (2020). Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age. *Journal of Consumer Psychology*, 30(4):736–758. https://doi.org/10.1002/jcpy.1191.

[Angwin and Larson, 2016] Angwin, J. and Larson, J. (2016). Bias in criminal risk scores is mathematically inevitable, researchers say. *ProPublica*. https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say.

[Angwin et al., 2016] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[Appelbaum, 2019] Appelbaum, B. (2019). Blame economists for the mess we're in. *The New York Times*. https://www.nytimes.com/2019/08/24/opinion/sunday/economics-milton-friedman.html.

[Axelrod, 1984] Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

[Bolukbasi et al., 2016] Bolukbasi, T., Chang, K.-W., Zou, J., Salgrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Technical Report 1607.06520, ArXiv.

[Boyd et al., 2011] Boyd, R., Richerson, P. J., and Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proc. Nat. Acad. Sciences (PNAS)*, 108:10918–10925.

[Brynjolfsson and McAfee, 2014] Brynjolfsson, E. and McAfee, A. (2014). *The Second Machine Age*. W. W. Norton & Co.

[Brynjolfsson and Mitchell, 2017] Brynjolfsson, E. and Mitchell, T. (2017). What can machine learning do? workforce implications. *Science*, 358:1530–1534. doi:10.1126/science.aap8062.

[Chouldechova, 2017] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Technical Report arXiv:1703.00056, arXiv. https://arxiv.org/pdf/1703.00056.pdf.

[Dieterich et al., 2016] Dieterich, W., Mendoza, C., and Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe Inc. https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

[Edsall, 2021] Edsall, T. B. (2021). Why Trump still has millions of Americans in his grip. *New York Times*. https://www.nytimes.com/2021/05/05/opinion/trump-automation-artificial-intelligence.html.

[Forbus, 2021] Forbus, K. D. (2021). Evaluating revolutions in artificial intelligence from a human perspective. In OECD, editor, *AI and the Future of Skills*, volume Volume 1: Capabilities and Assessments, pages 34–48. OECD Publishing, Paris. https://doi.org/10.1787/004710fe-en.

[Ford, 2015] Ford, M. (2015). *Rise of the Robots: Technology and the Threat of a Jobless Future*. Basic Books.

[Gajda, 2022] Gajda, A. (2022). *Seek and Hide: The Tangled History of the Right to Privacy*. Viking.

[Harkous et al., 2018] Harkous, H., Fawaz, K., Lebret, R., Schaub, F., and Shin, K. G. (2018). Polisis: Automated analysis and presentation of privacy policies using deep learning. In *Proc. 27th USENIX Security Symposium*, pages 531–548.

[Henrich, 2016] Henrich, J. (2016). *The Secret of Our Success*. Princeton University Press.

[Henrich et al., 2010] Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatx, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., and Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, 327:1480–1485. doi:10.1126/science.1182238.

[Henrich and Muthukrishna, 2021] Henrich, J. and Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual Review of Psychology*, 72:207–240.

[Khalil et al., 2020] Khalil, A., Ahmed, S. G., Khattak, A. M., and Al-Qirim, N. (2020). Investigating bias in facial analysis systems: A systematic review. *IEEE Access*, 8:130751–130761. doi:10.1109/ACCESS.2020.3006051.

[Klare et al., 2012] Klare, B. F., Burge, M. J., Klontz, J. C., Vorder Bruegge, R. W., and Jain, A. K. (2012). Face recognition performance: role of demographic information. *IEEE Trans. Information Forensics and Security*, 7(6):1789–1801.

[Kleinberg et al., 2017] Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proc. Innovations in Theoretical Computer Science (ITCS)*. https://arxiv.org/pdf/1609.05807.pdf.

[Larson et al., 2016] Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. Technical report, ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[Lee and See, 2004] Lee, J. D. and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, 46(1):50–80.

[Leonhardt, 2019] Leonhardt, D. (2019). How the upper middle class is really doing. *The New York Times*. https://www.nytimes.com/2019/02/24/opinion/income-inequality-upper-middle-class.html.

[Mayer et al., 1995] Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734.

[McWilliams, 2016] McWilliams, S. (2016). This political theorist predicted the rise of Trumpism. His name was Hunter S. Thompson. *The Nation*. https://www.thenation.com/article/this-political-theorist-predicted-the-rise-of-trumpism-his-name-was-hunter-s-thompson/.

[Nissenbaum, 2011] Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus*, 140(4):32–48.

[on AI, 2019] on AI, H. L. E. G. (2019). Ethics guidelines for trustworthy AI. Technical report, European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[Rand et al., 2014] Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., and Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, 5(3677). doi:10.1038/ncomms4677.

[Rawls, 1999] Rawls, J. (1999). *A Theory of Justice*. Harvard University Press, revised edition.

[Rodolfa et al., 2021] Rodolfa, K. T., Lamba, H., and Ghani, R. (2021). Empirical observation of negligible fairness-accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, 3:896–904. https://doi.org/10.1038/s42256-021-00396-x.

[Rousseau et al., 1998] Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. (1998). Not so different after all: a cross-discipline view of trust. *Academy of Management Review*, 23(3):393–404.

[Schaub et al., 2017] Schaub, F., Balebako, R., and Cranor, L. F. (2017). Designing effective privacy notices and controls. *IEEE Internet Computing*, 21(3):70–77.

[Singer, 1981] Singer, P. (1981). *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton University Press.

[Sorkin, 2019] Sorkin, A. R. (2019). Dear C.E.O.: Before you give to charity, look at your own workplace. *New York Times*. https://www.nytimes.com/2019/12/24/business/dealbook/income-inequality-corporate-response.html.

[Tomasello et al., 2012] Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., and Herrmann, E. (2012). Two key steps in the evolution of human cooperation: the interdependence hypothesis. *Current Anthropology*, 53(6):673–692.

[Widen and Koopman, 2022] Widen, W. H. and Koopman, P. (2022). Autonomous vehicle regulation & trust: the impact of failures to comply with standards. *UCLA J. of Law and Technology*, forthcoming. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3969214.

[Wing, 2021] Wing, J. M. (2021). Trustworthy AI. *Communications of the ACM*, 64(10):64–71.

[Zuboff, 2015] Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, 30:75–89. https://link.springer.com/article/10.1057/jit.2015.5.

[Zuboff, 2019] Zuboff, S. (2019). *The Age of Surveillance Capitalism*. PublicAffairs, New York.