

# Reading Assignments for CSE/ROB 543: Ethics for AI and Robotics (W24) Benjamin Kuipers Draft: April 4, 2024

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

- **Reading Assignment** (read before class)  
Course syllabus: <https://web.eecs.umich.edu/~kuipers/teaching/cse543-W24.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>  
UM CoE Honor Code [UM CoE Honor Code]  
“The Blind Men and the Elephant”: <https://web.eecs.umich.edu/~kuipers/opinions/Elephant.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

- **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

- **Reading Assignment** (read before class)
  - Leyton-Brown & Shoham, *Essentials of Game Theory*, ch.1-2 [Leyton-Brown+Shoham-08-ch.1-2]
  - Axelrod & Hamilton, The evolution of cooperation [Axelrod-science-81]
- **Other Readings** (valuable for many purposes)
  - Anatol Rapoport, The use and misuse of game theory [Rapaport-sciam-62]
  - Robert Axelrod, *The Evolution of Cooperation* [Axelrod, 1984]
  - . . . and more useful references in the slides.

## 1.3 Ethics, Trust, and Cooperation

- **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)
  - Kuipers, AI and Society: Ethics, Trust, & Cooperation, *CACM*, 2023.
    - <https://web.eecs.umich.edu/~kuipers/research/pubs/Kuipers-cacm-23.html>
  - Rousseau, et al, Not so different after all . . . [Rousseau et al., 1998]
  - Lee & See, Trust in automation [Lee and See, 2004]
  - Ethics Guidelines for Trustworthy AI* [on AI, 2019]
  - Jeannette Wing, Trustworthy AI [Wing, 2021]

## 1.4 Evolutionary Origins

- **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)
  - Bear & Rand, Intuition, deliberation, and the evolution of cooperation [Bear and Rand, 2016]
  - 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>
  - Koopman, SAE J3016 Users Guide, 2021. [Koopman, 2021]  
 Explains table: "SAE Autonomy Levels" [SAE J3016 levels 5-21]
  
- **Other Readings** (valuable for many purposes)
  - NHTSA Automated Driving Systems.  
<https://www.nhtsa.gov/vehicle-manufacturers/automated-driving-systems>
  - NHTSA Voluntary Safety Self-Assessments.  
<https://www.nhtsa.gov/automated-driving-systems/voluntary-safety-self-assessment>
  - Waymo Safety Publications: Methodology, Performance Data, Other. <https://waymo.com/safety/>
  - Waymo Safety Report, September 2021. [Waymo, 2021]
  - Rodney Brooks, roboticist: (focus on the predictions about self-driving cars)  
<https://rodneybrooks.com/predictions-scorecard-2023-january-01/>  
<https://rodneybrooks.com/edge-cases-for-self-driving-cars/>

### 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. [Kuipers, 2020]  
<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: Phil Koopman (CMU) guest lecture

- **Reading Assignment** (read before class)
  - Koopman & Widen, Safety ethics for design & test of automated driving features, *IEEE Design & Test*, 2024. [Koopman and Widen, 2024]
  - Koopman & Widen, Breaking the tyranny of net risk metrics for automated vehicle safety. [Koopman-ssrn-23].
  
- **Other Readings** (valuable for many purposes)
  - Widen & Koopman, Autonomous vehicle regulation & trust, *UCLA J. Law & Technology*, 2022. [Widen and Koopman, 2022]  
[https://en.wikipedia.org/wiki/Self-driving\\_car](https://en.wikipedia.org/wiki/Self-driving_car)
  - Cade Metz, NYTimes: [Metz-nyt-12-7-21] [Metz-nyt-6-8-22] [Metz-nyt-2-1-23]

## 2.4 AI Safety and Existential Threats

- **Reading Assignment** (read before class)
  - Vernor Vinge, The Technological Singularity, 1993/2003. [Vinge-wer-03]
  - Stuart Russell, It's not too soon to be wary of AI, *IEEE Spectrum*, 2019. [Russell-spectrum-19]
- **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>
  - Stuart Russell, *Human Compatible*, 2019.

## 3 Surveillance and Privacy

Information: 'I' vs. 'We' vs. 'They' [von Hanxleden, 2022]  
 Section 230 and a Tragedy of the Commons, *CACM* [Cusumano, 2021]  
 A legal challenge to algorithmic recommendations, *CACM* [Samuelson, 2023]  
 Bruce Schneier, Banning facial recognition isn't enough. [Schneier, 2020]

### 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].<sup>1</sup>
- **Other Readings** (valuable for many purposes)
  - Harkous, et al, Polisis: Automated analysis and presentation of privacy policies using deep learning. [Harkous et al., 2018]
  - Kumar, et al, Finding a choice in a haystack: Automating extraction of opt-out statements from 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>

---

<sup>1</sup>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.

### 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]
- **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]

### 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]
  - O'Neill, 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

The Bias Hunter, in *Science* [Starr, 2022]

The (im)possibility of fairness, 2021 [Friedler et al., 2021]

Actionable auditing revisited [Raji and Buolamwini, 2023, Conitzer et al., 2023]

### 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)

Barocas, Hardt & Narayanan, chap.3. <https://fairmlbook.org/pdf/classification.pdf>

[https://en.wikipedia.org/wiki/COMPAS\\_\(software\)](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].

### 4.3 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, *F&SF*, 1961. [Harrison Bergeron.rtf]

The Orange Story. <http://www.mediationtools.com/articles/smbj9605.html>

#### 4.4 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 Programmer as Woman is to Homemaker?" [Bolukbasi et al., 2016]

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

## 5 Jobs, Automation, and Existential Threats

Moshe Vardi, The winner-takes-all tech corporation, *CACM*, 2019 [Vardi, 2019]

review of Moritz Altenried, *The Digital Factory: The Human Labor of Automation*, 2022.

[doi:10.1126/science.abn1041]

### 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 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/>

## 5.4 Can AI be aligned with human values?

- **Reading Assignment** (read before class)  
 D. Hadfield-Menell and G. K. Hadfield, Incomplete contracting and AI alignment. *AIES*, 2019. [Hadfield-Menell-aies-19]  
 Ji, et al, AI alignment: a comprehensive survey, ArXiv, 2024. [Ji et al., 2024]
- **Other Readings** (valuable for many purposes)  
 Amodei, et al, Concrete problems in AI safety. [Amodei et al., 2016]  
 De Kai, Should AI accelerate? Decelerate? The answer is both. [Kai-nyt-12-10-23]  
 Pan, et al, Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark. ICML, 2023. [Pan et al., 2023]

## 6 Regulation of AI

Tragedy revisited, *Science* policy forum [Boyd-science-18] [Boyd et al., 2018]  
 review of Matthew Cobb, *As Gods*, 2022 [doi:10.1126/science.ade5848]



The AI ethicist's dirty hands problem [Sætra et al., 2023]  
Marc Rotenberg, Fair AI Practices (CACM blog) [Rotenberg, 2022]  
Marc Steen, Ethics as a participatory and iterative process [Steen, 2023]

## 6.1 Can we / Should we regulate AI?

- **Reading Assignment** (read before class)  
The EU AI Act. [EuroParl-aiact-12-9-23] [Parliament, 2023].  
Tamò-Larrieux, et al, Regulating for trust: Can law establish trust in artificial intelligence? [TamoLarrieux-rg-23] [Tamò-Larrieux et al., 2023].
- **Other Readings** (valuable for many purposes)  
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]  
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)  
Dan Crane, algorithmic collusion: [Canvas > Files/readings/Crane-algorithmic-collusion-20.pdf]; also explore Dan Crane's personal website (<https://profdancrane.com>)  
Bryant Walker Smith: explore his personal website (<https://newlypossible.org/wiki/Home>), and [https://sc.edu/study/colleges\\_schools/law/faculty\\_and\\_staff/directory/smith\\_bryant\\_walker.php](https://sc.edu/study/colleges_schools/law/faculty_and_staff/directory/smith_bryant_walker.php),
- **Other Readings** (valuable for many purposes)

## 6.3 Synthesis of this semester

- **Reading Assignment** (read before class)  
Tamara Broderick, et al, Toward a taxonomy of trust for probabilistic machine learning, *Science Advances*, 2023. [Broderick-sciadv-23] [Broderick et al., 2023]
- **Other Readings** (valuable for many purposes)

## 6.4 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. [Davis-22]  
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.

## 6.5 Next Challenges

- **Reading Assignment** (read before class)
  - Brianna Wessling, Cruise recalls 300 robotaxis in response to crash with bus. *The Robot Report*, 10 April 2023 [Wessling-robrept-4-10-23] [Wessling, 2023]
  - Kyle Vogt, Why we do AV software recalls. Cruise Blog, 7 April 2023. [Vogt-cruise-4-7-23] [Vogt, 2023]
  - Fatal Tesla collision with firetruck under federal investigation [Kolodny-cnbc-3-8-23] [Kolodny, 2023]
  - NHTSA Engineering Analysis of Autopilot & First Responder Scenes [NHTSA-ea-22] [NHTSA, 2022]
- **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>.
- [Amodei et al., 2016] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in AI safety. Technical report, ArXiv. arXiv:1606.06565v2.
- [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.
- [Bear and Rand, 2016] Bear, A. and Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proc. Nat. Acad. Sciences*, 113(4):936–941. [www.pnas.org/cgi/doi/10.1073/pnas.1517780113](http://www.pnas.org/cgi/doi/10.1073/pnas.1517780113).
- [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.
- [Boyd et al., 2018] Boyd, R., Richerson, P. J., Meinzen-Dick, R., Moor, T. D., Jackson, M. O., Gjerde, K. M., Harden-Davies, H., Frischmann, B. M., Madison, M. J., Strandburg, K. J., McLean, A. R., and Dye, C. (2018). Tragedy revisited. *Science*, 362(6420):1236–1241.
- [Broderick et al., 2023] Broderick, T., Gelman, A., Meager, R., Smith, A. L., and Zheng, T. (2023). Toward a taxonomy of trust for probabilistic machine learning. *Science Advances*, 9(eabn3999). <https://www.science.org/doi/pdf/10.1126/sciadv.abn3999>.
- [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>.

- [Conitzer et al., 2023] Conitzer, V., Hadfield, G. K., and Vallor, S. (2023). The impact of auditing for algorithmic bias. *CACM*, 66(1):100.
- [Cusumano, 2021] Cusumano, M. A. (2021). Section 230 and a tragedy of the commons. *CACM*, 64(10):16–18.
- [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](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.
- [Friedler et al., 2021] Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2021). The (im)possibility of fairness: different value systems require different mechanisms for fair decision making. *CACM*, 64(4):136–143.
- [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., Lebet, 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., Bolyanatz, 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.
- [Ji et al., 2024] Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K. Y., Dai, J., Pan, X., O’Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., McAleer, S., Yang, Y., Wang, Y., Zhu, S.-C., Guo, Y., and Gao, W. (2024). AI alignment: A comprehensive survey. Technical Report arXiv:2310.19852, ArXiv.
- [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>.
- [Kolodny, 2023] Kolodny, L. (2023). Fatal Tesla collision with firetruck under federal investigation. *CNBC*. <https://www.cnn.com/2023/03/08/fatal-tesla-collision-with-fire-truck-under-federal-investigation.html>.
- [Koopman, 2021] Koopman, P. (2021). SAE J3016 User Guide. <https://users.ece.cmu.edu/~koopman/j3016/>.
- [Koopman and Widen, 2024] Koopman, P. and Widen, W. (2024). Safety ethics for design & test of automated driving features. *IEEE Design & Test*, 41(1):17–24.
- [Kuipers, 2020] Kuipers, B. (2020). Perspectives on ethics of AI: Computer science. In Dubber, M., Pasquale, F., and Das, S., editors, *Oxford Handbook of Ethics of AI*, pages 421–441. Oxford University Press.
- [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-compass-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/>.
- [NHTSA, 2022] NHTSA (2022). Autopilot & first responder scenes. Technical Report EA 22-002, National Highway Traffic Safety Administration (NHTSA). <https://static.nhtsa.gov/odi/inv/2022/INOA-EA22002-3184.PDF>.
- [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>.
- [Pan et al., 2023] Pan, A., Chan, J. S., Zou, A., Li, N., Basart, S., Woodside, T., Zhang, H., Emmons, S., and Hendrycks, D. (2023). Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark. In *Int. Conf. Machine Learning (ICML)*.

- [Parliament, 2023] Parliament, E. (2023). Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI. Press release. <https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai>.
- [Raji and Buolamwini, 2023] Raji, I. D. and Buolamwini, J. (2023). Actionable auditing revisited – investigating the impact of publicly naming biased performance results of commercial AI products. *CACM*, 66(1):101–108.
- [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>.
- [Rotenberg, 2022] Rotenberg, M. (2022). Fair AI practices. <https://cacm.acm.org/blogs/blog-cacm/265535-fair-ai-practices/fulltext>.
- [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.
- [Sætra et al., 2023] Sætra, H. S., Coeckelbergh, M., and Danaher, J. (2023). The AI ethicist’s dirty hands problem. *CACM*, 66(1):39–41.
- [Samuelson, 2023] Samuelson, P. (2023). A legal challenge to algorithmic recommendations. *CACM*, 66(3):32–34.
- [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.
- [Schneier, 2020] Schneier, B. (2020). Banning facial recognition isn’t enough. *New York Times*. <https://www.nytimes.com/2020/01/20/opinion/facial-recognition-ban-privacy.html>.
- [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>.
- [Starr, 2022] Starr, D. (2022). The bias hunter. *Science*, 376(6594):686–690.
- [Steen, 2023] Steen, M. (2023). Ethics as a participatory and iterative process. *CACM*, 66(5):27–29.
- [Tamò-Larrieux et al., 2023] Tamò-Larrieux, A., Guitton, C., Mayer, S., and Lutz, C. (2023). Regulating for trust: Can law establish trust in artificial intelligence. *Regulation & Governance*. doi:10.1111/rego.12568.

- [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.
- [Vardi, 2019] Vardi, M. (2019). The winner-takes-all tech corporation. *CACM*, 62(11):7.
- [Vogt, 2023] Vogt, K. (2023). Why we do AV software recalls. Cruise Blog. <https://getcruise.com/news/blog/2023/why-we-do-av-recalls/>.
- [von Hanxleden, 2022] von Hanxleden, R. (2022). Information: ‘I’ vs. ‘We’ vs. ‘They’. *CACM*, 65(5):45–47. doi:10.1145/3491205.
- [Waymo, 2021] Waymo (2021). Waymo safety report. Technical report, Waymo.
- [Wessling, 2023] Wessling, B. (2023). Cruise recalls 300 robotaxis in response to crash with bus. *The Robot Report*. <https://www.therobotreport.com/cruise-recalls-300-robotaxis-in-response-to-crash-with-bus/>.
- [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](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.