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
  - UM CoE Honor Code [UM CoE Honor Code]

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

### 1.2 The Prisoner’s Dilemma

- **Reading Assignment** (read before class)
  Axelrod & Hamilton, *The evolution of cooperation* [Axelrod-science-81]

- **Other Readings** (valuable for many purposes)
  Anatol Rapaport, *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)
  Mayer, Davis & Schoorman, *An integrative model of organizational trust*. [Mayer et al., 1995]

- **Other Readings** (valuable for many purposes)
  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)
  Boyd, Richerson & Henrich, *The cultural niche: Why social learning is essential for human adapta-

- **Other Readings** (valuable for many purposes)
  Bear & Rand, *Intuition, deliberation, and the evolution of cooperation* [Bear and Rand, 2016]
  *Science*, 2010. [Henrich et al., 2010]
  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)
  https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812506
  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.
  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)
  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-ssrm-23].

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

- **Other Readings** (valuable for many purposes)
  https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html
  Tim Urban, The AI revolution: our immortality or extinction, 2015.

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].

- **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)
  Rathjhe, 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]

---

1I 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]
  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
  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].

- **Other Readings** (valuable for many purposes)
  
  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. [O'Neill-nyt-6-9-19]
  Friedman, China’s bullying is becoming a danger to the world and itself. [Friedman-nyt-10-19-21]
  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].

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]

- **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].

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

- **Reading Assignment** (read before class; available in Canvas > Files/readings/)
  Bolukbasi, et al, “Man is to Computer Programmer 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)
  

- **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.boozallen.com/s/insight/blog/algorithmic-bias.html

- **Other Readings** (valuable for many purposes)
  
  Relevant stories, science fiction and other:
  
  
  Kurt Vonnegut, Harrison Bergeron. [Harrison Bergeron.rtf]

5 Jobs, Automation, and Existential Threats


  [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].
  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)
  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)

- **Other Readings** (valuable for many purposes)
  De Kai, Should AI accelerate? Decelerate? The answer is both. [Kai-nyt-12-10-23]

6 Regulation of AI

Tragedy revisited, *Science* policy forum [Boyd-science-18] [Boyd et al., 2018]
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].

- **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)
  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.

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

6.3 Guest lecture: Prof. Jerry Davis, UM Ross

- **Reading Assignment** (read before class)

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

6.4 Synthesis of this semester

- **Reading Assignment** (read before class)

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


