Supplemental Resources for Learning about the Ethics of Artificial Intelligence

Professor Andrew A. Anda, Ph.D.

CSIT

St. Cloud State University

Author Note

WORKLOAD REASSIGNMENT, Spring 2024

Soon after papers describing successful experimentation with recombinant DNA were published in the early 1970s, scientists realized the potential existential biohazard risks this new technology posed. These concerns motivated the organization of the 1975 Asilomar Conference on Recombinant DNA to address these issues. The participants were not only biologists, but also lawyers and physicians. This conference generated a published set of principles, recommendations, and prohibitions – all with public transparency. The significant precedents this conference established are summarized here.

In 2010, based on the 1975 Asilomar Conference model, climate scientists organized the Asilomar International Conference on Climate Intervention Technologies to identify and minimize potentially existential risks concerning climate engineering. Again, the set of participants was broad, comprising scientists, engineers, environmentalists, disaster relief workers, and lawyers – in a transparent and open meeting. A set of recommendations followed.

In the first two decades of the twenty first century, rapid enabling advances in computing hardware and systems efficiencies and capacities driving big data capabilities, as well as deep learning algorithmic advances, led computer scientists, and the general public, to worry that pace and scope of AI innovations may be spiraling out of control. So, in 2017 the Asilomar Conference on Beneficial AI was convened to address ethical the potentially existential risks that artificial intelligence development and applications foreshadowed, and to promote the ethical development and application of artificial intelligence. Following the models, of the prior Asilomar conferences, attendees were experts comprising a diverse set of professions including, programmers, engineers, roboticists, physicists, economists, ethicists, and legal scholars. During the conference attendees participated in a principled AI discussion leading to the drafting of 23 principles of AI governance. These 23 principles were categorized and partitioned into a set of

research issues, a set of *ethics and values*, and a set of *longer-term issues*. Here is a nice journalistic overview of this Asilomar AI conference (including links to earlier efforts which were deemed "… either too narrow in scope, or far too generalized.").

Following are a set of categorized curated resource listings to better understand the ethical issues that artificial intelligence (AI) development and application raise.

Definitions

There are definitions we must establish:

- *Cognitive science* is not *artificial intelligence* --

  cognitive science is understanding how a human mind solves challenging problems, whereas artificial intelligence algorithms and systems are free to use any effective methodology to solve challenging problems;

  - Categories of AI problem solving;
    - Machine learning;
    - Knowledge representation and knowledge engineering;
    - Automated planning and automated decision making;
    - Natural language processing;
    - Machine perception;
    - Affective computing;
    - artificial general intelligence;
  - Heuristics of AI problem solving;
    - Outline of AI problem solving;
  - Applications of AI problem solving;
    - Generative AI;

        - Military AI;
    - Philosophy of AI problem solving;
    - Ethics of AI problem solving;
- An *automaton* is an autonomous entity with an etymological original meaning of "acting of one's own will".
    - Robotics glossary;
        - Nanorobotics;
    - Android;
        - Uncanny valley;
    - Cyborg;
    - Animatronic;
    - Self-replicating machine;
        - Gray goo;

AI History

How we got here…

- Automaton history;
    - Talos – Greek mythology;
    - Golem – Jewish folklore;
    - History of robots;
- History of artificial intelligence;
    - Timeline;
    - Progress;
    - Formal reasoning;
    - Automated reasoning;
    - Knowledge representation and reasoning;
    - Machine learning;
    - Dartmouth workshop & 50 year anniversary conference;

The history of consideration of thinking machines goes back to the nineteenth century in fiction literature, which we will cover in a separate section. In what we can consider modern computer science, we can refer to the Turing test, wherein a human tries to determine whether they are conversing with an artificial intelligence. Alan Turing included what he termed the "Imitation Game" in his seminal paper, "Computing Machinery and Intelligence". Interestingly, a recent article claims that ChatGPT broke the Turing test. Where an AI is tasked with determining whether it is conversing with a human, that is termed a reverse Turing test, and CAPTCHAs are examples.

The Butler did it – history of AI in fiction genres

Reacting to the epiphanies of Charles Darwin's publications on natural selection, Nineteenth Century author, Samuel Butler, (one quarter century after Charles Babbage designed his Analytical Engine, the first computer design considered to be *Turing complete*) wrote an article for newspaper publication, "Darwin among the Machines", wherein Butler speculated whether machines may eventually supplant humans. This essay was included in his *Book of the Machines* which became incorporated into his novel, *Erewhon*. The society in Erewhon has a distant history where a revolution destroyed most mechanical inventions – a more successful version of the machine-breaking in the Luddite movement in the early Nineteenth Century. This machine-destroying revolution in *Erewhon,* lent Butler's name to a backstory in Frank Herbert's *Dune* novel named the Butlerian Jihad, which is why there is no active AI in *Dune*.

There are other examples in Nineteenth Century speculative fiction, where artificial intelligences are described. One could consider Mary Shelly's Frankenstein's Monster to be an artificial intelligence. Prior references are found in "Ancient dreams of intelligent machines: 3,000 years of robots".

A principal sub-genre of speculative fiction where AI is often encountered is the science fiction genre. There is a rich history of AI in science fiction. Science fiction often extrapolates the development and implementation of a current or future technology then speculates its effect on people, society, or mankind – the technology we are considering is AI. An extensive and comprehensive survey of AI in fiction is in the article, "Artificial intelligence in fiction: between narratives and metaphors".

There is a long history of AI in film. Here's a listing of "The top 20 AI films in pictures". The AI *tropes* in film, and in television series, have been categorized (however, absent is the

transcendence trope where the AI evolves past the Singularity, transcending the need of physical media to exist, finding humans of no interest anymore, so the AI just leaves. The film, *Her*, exhibits this trope.). There are numerous television series which featured AI, also listed here.

The term "robot" was coined in the Czech play, *R.U.R* in 1921. One of the earliest and most memorable film robots was Maschinenmensch in Fritz Lang's *Metropolis* (1927). Here's a list of fictional robots and androids. Arguably, the most seminal set of stories about robots and artificial intelligence is Asimov's Robot Series, where he originates his Three Laws of Robotics in 1942 in his short story, "Runaround", which was added to his collection, *I Robot*. Common plot twists involving the three laws of robotics are that the robots interpret the laws differently than humans would. For instance, the primary law is "A robot may not injure a human being or, through inaction, allow a human being to come to harm", which might force the robot to imprison humans to keep them from hurting each other – this is analogous to the trolley problem for autonomous vehicles. The Robot series was then incorporated with Asimov's Foundation series to comprise the Foundation Universe.

One of the most seminal AI computers appeared in Kubrik's & Clark's *2001: A Space Odyssey* – HAL. In Clark's novelization of the film, we learn that HAL's deadly behavior was caused by an attempt to resolve conflicting directives. Here's a list of fictional computers, including HAL. A more humorous fictional computer is Deep Thought from the HHGttG by Douglas Adams. Another fictional trend is for an AI to emerge within a computer network – the most influential of these is Skynet, from the *Terminator* film, which proves to be an existential threat AI – here's a reference including other AI takeovers in popular culture. The problems with uncontrolled self-replicating is gray goo featured in the Stargate TV series. Von Neumann probes are the origin of the machine civilization in James P. Hogan's *Code of the Lifemaker*.

Existential Threats

An existential threat is a risk to the survival of the human race. Here are some categories:

- Global_catastrophic_risk
- Existential_risk_from_artificial_general_intelligence
- Technological_singularity
- Recursive_self-improvement
- Superintelligence
    - Superintelligence: Paths, Dangers, Strategies - Wikipedia
- AI_takeover
    - Skynet (*Terminator*) — Wikipedia
- Artificial_intelligence_arms_race
- Lethal_autonomous_weapon
- List_of_fictional_military_robots
- Slaughterbots
- Artificial_Intelligence_Cold_War
- Artificial intelligence arms race — Wikipedia
- Effective altruism: long-term future and global catastrophic risks — Wikipedia

AI Ethics

- Ethics_of_artificial_intelligence
- Machine_ethics
- AI_alignment
    - The AI Alignment Problem: Why It's Hard, and Where to Start — MIRI [video w. Webpage annotations]
    - *The Alignment Problem: Machine Learning and Human Values* by Brian Christian
    - *The Alignment Problem* — C-SPAN [video w. transcript]
    - *The Alignment Problem*, Linking Machine Learning And Human Values — Forbes {Review}
- Robot_ethics
    - Military robot: Ethical and legal concerns — Wikipedia
    - Is robotics about to have its own ChatGPT moment?
- AI_safety
- Moral_agency#Artificial_Moral_Agents
- Moral_responsibility#Artificial_systems
- science-fiction-ethics-and-the-human-condition
- *Human Compatible: Artificial Intelligence and the Problem of Control* — Wikipedia

Algorithm-Centric Ethics

- Rise of the machines: are algorithms sprawling out of our control? — Wired
- Statement on Algorithmic Transparency and Accountability — ACM US Public Policy Council
- Principles for Algorithmic Transparency and Accountability: A Provenance Perspective
- Rise of the machines: are algorithms sprawling out of our control? - Wired
- Regulation of algorithms — Wikipedia
    - Regulation_of_artificial_intelligence
- Algorithmic bias — Wikipedia
    - Why algorithms can be racist and sexist: A computer can make a decision faster. That doesn't make it fair — Vox
    - Stable Diffusion’s text-to-image model amplifies stereotypes about race and gender — here’s why that matters
    - Bias In AI Algorithms — towards data science
    - The Real Reason Tech Struggles With Algorithmic Bias — Wired
    - I Know Some Algorithms Are Biased — because I Created One — SciAm
    - Biased Algorithms Learn From Biased Data: 3 Kinds Biases Found In AI Datasets
    - Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms — Brookings
    - Algorithmic Bias In Health Care: A Path Forward — Health Affairs
    - The coming war on the hidden algorithms that trap people in poverty — MIT Technology Review
    - There's an Unsettling Racial Bias in How Well AI Faces Can Fool Us

    - How AI bias happens – and how to eliminate it - Healthcare IT News
    - Scientists Built an AI to Give Ethical Advice, But It Turned Out Super Racist - Futurism
    - Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models
    - Introducing the AI Equity Lab and the path toward more inclusive tech

- Behavioral ethics -- Wikipedia
    - Behavioral Ethics: Toward a Deeper Understanding of Moral Judgment and Dishonesty - Bazerman
    - Teaching Behavioral Ethics - Robert Prentice [PDF]
    - Behavioral Ethics - Herbert Gintis [PDF]
    - Behavioral ethics for Homo economicus, Homo heuristicus, and Homo duplex - Kluver
    - Behavioral Ethics lab (economics focus)
    - What is behavioral ethics? Why is it important for all companies?
    - The Ethics of Intracorporate Behavioral Ethics - California Law Review
    - Bounded Ethicality
    - Decision Making (& bounded ethicality)

Nation-state and pan-nation-state-level regulation

Nation-states and pan-nation-states (e.g. the EU) have been working towards developing regulatory frameworks individually and collectively to understand and regulate AI. These processes usually begin by forming a panel of experts. After some time, the panel of experts drafts a recommendation report. Laws and regulations are then enacted based on the report recommendations. Here's progress so far:

USA

- FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence -- The White House
    - Biden-Harris Administration Announces Key AI Actions 180 Days Following President Biden's Landmark Executive Order
- NIST: Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence
- NSF Statement on White House's Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence
    - NSF: NSF-led National Artificial Intelligence Research Resource Task Force Releases Final Report
    - NSF: Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource [PDF]
    - WH: National Artificial Intelligence Research Resource Task Force Releases Final Report
    - AI.GOV (U.S. government website for AI)
- U.S. Government Will Support Domestic PCB Manufacturing {*related* to AI Act}
- The US Just Unveiled The Most Ambitious Attempt to Control AI to Date - Science Alert {Australian perspective}

STATES

- The State of State AI Laws: 2023 -- EPIC.org
- Enacts the New York artificial intelligence bill of rights
- New York State Senate: Requires disclosure of the use of artificial intelligence in political communications

EU

- EU AI Act: first regulation on artificial intelligence
- EU: The Artificial Intelligence Act
- EU: THE AI ACT
- The EU AI Act: A Primer

UK

- GOV.UK: Proposed principles to guide competitive AI markets and protect consumers
    - AI Foundation Models: Initial report
- UK antitrust regulator lays out seven AI principles - The Verge
- UK's new AI principles target 'pro-innovation' edge over the EU -- TNW
- UK focuses on transparency and access with new AI principles -- Reuters

Canada

- Canada: Artificial Intelligence and Data Act
    - The Artificial Intelligence and Data Act (AIDA) – Companion document
    - Pan-Canadian Artificial Intelligence Strategy

International Cooperation

- The Global Partnership on Artificial Intelligence
- OECD: Recommendation of the Council on Artificial Intelligence
- Britain publishes 'Bletchley Declaration' on AI safety - Reuters
    - UK: The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023
- The US and 30 Other Nations Agree to Set Guardrails for Military AI - Wired

United Nations

- UN: Towards an Ethics of Artificial Intelligence
- UN: AI shows 'great promise for health' but regulation is key: WHO chief
- UN: Secretary-General Urges Security Council to Ensure Transparency, Accountability, Oversight, in First Debate on Artificial Intelligence
- UN-CEB: Principles for the Ethical Use of Artificial Intelligence in the United Nations System
- UN-CEB: Artificial Intelligence
- UN deadlocked over regulating AI -- Axios
- Artificial intelligence is on the world's mind. Is the UN the place to figure out what to do about it? -- PBS
- Can the U.N. really regulate the power — and danger — of AI? -- Washington Post

Industry Cooperation

- IBM, Meta form "**AI Alliance**" with 50 organizations to promote open source AI - ars Technica
- **AI Alliance** Launches as an International Community of Leading Technology Developers, Researchers, and Adopters Collaborating Together to Advance Open, Safe, Responsible AI

Risks

- CACM Inside Risks {links to every column}
- Privacy in an AI Era: How Do We Protect Our Personal Information?
    - Shaping the future: A dynamic taxonomy for AI privacy risks
    - Protecting privacy in an AI-driven world
    - Beware the Privacy Violations in Artificial Intelligence Applications
    - Privacy and responsible AI
- Cybersecurity
    - AI in Cyber Security: Risks of AI
    - AI and cyber security: what you need to know
    - NIST researchers warn of top AI security threats
    - The near-term impact of AI on the cyber threat
    - AI and cybersecurity: How to navigate the risks and opportunities
    - U.S. Department of the Treasury Releases Report on Managing Artificial Intelligence-Specific Cybersecurity Risks in the Financial Sector
    - How to improve cybersecurity for artificial intelligence
    - ARTIFICIAL INTELLIGENCE AND THE FUTURE OF RISK

- Work
    - Everyone in Your Organization Needs to Understand AI Ethics
    - AI: These are the biggest risks to businesses and how to manage them
    - AI-Related Risks Test the Limits of Organizational Risk Management
    - Employment Discrimination and AI for Workers- eeoc.gov [PDF]
    - The Risks of Artificial Intelligence to Security and the Future of Work
    - Automation and AI will disrupt the American labor force. Here's how we can protect workers
    - 5 Key Risks of Generative AI in the Workplace
    - AI's threat to diversity, equity and inclusivity
    - What are the potential risks of AI for equity and social justice?
    - A Blueprint for Equity and Inclusion in Artificial Intelligence [PDF]
    - In Reversal Because of A.I., Office Jobs Are Now More at Risk -- NYT
    - A.I.'s Threat to Jobs Prompts Question of Who Protects Workers -- NYT
- AI and Accessibility -- Pulitzer Center
    - Blind Internet Users Struggle With Error-Prone AI Aids
    - AI for Accessibility: Opportunities and Challenges -- EqualEntry
    - Accessibility and Artificial Intelligence: A More Diverse Future? -- IEEE
    - Digital accessibility in the era of artificial intelligence—Bibliometric analysis and systematic review
    - Accessibility of AI Interfaces

- Intellectual Property
    - Generative AI Has an Intellectual Property Problem
    - Copyright law is AI's 2024 battlefield
    - Artificial intelligence and intellectual property considerations
    - AI & Intellectual Property: Artificial Intelligence Legal Implications
    - IP Risks, Benefits and Ideal Use-Cases for AI: Best Practices When Drafting Generative AI Usage Policies
    - Generative Artificial Intelligence and Copyright Law [PDF]
    - Intellectual Property Protection for Artificial Intelligence
- Hallucination_(artificial_intelligence) – Wikipedia
    - AI hallucinates software packages and devs download them – even if potentially poisoned with malware
    - Opinion: The rise of deepfake pornography is devastating for women -- CNN
- Deepfake — Wikipedia
    - Artificial_intelligence_content_detection
    - A New Kind of AI Copy Can Fully Replicate Famous People. The Law Is Powerless. -- Politico
    - Deepfakes in the courtroom: US judicial panel debates new AI evidence rules

Transportation

- Autonomous Accidents: The Ethics of Self-Driving Car Crashes
- Ethical Decision Making during Automated Vehicle Crashes
- Trolley Problem: Implications for autonomous vehicles -- Wikipedia
    - Moral Machine – Wikipedia {poll people about how they would solve the Trolley Problem}

Politics

- AI and Democracy -- how AI-generated disinformation could threaten elections and democracies around the world -- Aljazeera [video]
- *Six ways that AI could change politics* by Bruce Schneier & Nathan Sanders -- MIT Technology Review
- Artificial Intelligence Enters the Political Arena -- Council on Foreign Relations
- How AI will transform the 2024 elections -- Brookings
- AI can strengthen U.S. democracy—and weaken it -- Brookings
- Can politicians catch up with AI? -- NPR
- Microsoft warns deepfake election subversion is disturbingly easy

Consciousness

- Consciousness in Artificial Intelligence: Insights from the Science of Consciousness - arXiv
- AI Consciousness: Scientists Say We Urgently Need Answers - Nature
- Association for Mathematical Consciousness Science (AMCS)
- If AI becomes conscious: here's how researchers will know - Nature

Advocacy

- Algorithmic_Justice_League
- Association_for_the_Advancement_of_Artificial_Intelligence
- https://www.eff.org/search/site/artificial%20intelligence
- https://www.google.com/search?q=site%3Aaclu.org+artificial+intelligence
- Centre for the Study of Existential Risk — Wikipedia
- Center for Security and Emerging Technology — Wikipedia
- Center for Human-Compatible Artificial Intelligence — Wikipedia
- Machine Intelligence Research Institute — Wikipedia
- Future of Life Institute — Wikipedia
- Partnership on AI — Wikipedia

Commentary Resources

- https://www.oneusefulthing.org/
- https://www.schneier.com/tag/artificial-intelligence/
- https://www.wbur.org/search?q=artificial%20intelligence&channel=onpoint
- AI & SOCIETY: Journal of Knowledge, Culture and Communication

Uncategorized

- It’s the End of the Web as We Know It
- Ex-Amazon exec claims she was asked to ignore copyright law in race to AI
- Tech brands are forcing AI into your gadgets—whether you asked for it or not
- Artificial Intelligence Is Now Smart Enough to Know When It Can't Be Trusted
- Democratizing harm: Artificial intelligence in the hands of nonstate actors - Brookings
- Artificial intelligence challenges what it means to be creative - ScienceNews
- How Artificial Intelligence enhances education
- The controversy behind a star Google AI researcher’s departure
    - Google fires another AI researcher who reportedly challenged findings (updated)
- ‘Obviously ChatGPT’ — how reviewers accused me of scientific fraud