Literature Summary of Principles for Safe AI

Last updated September 26, 2017

This summary was put together as part of the research for creating "4 Steps to Good Narrow AI". Special thanks to Louis-Félix La Roche Morin for his foundation help in gathering this research.

Please send additions, comments, or questions to sh@elementai.com

Table of Contents

Copenhagen Letter

***Author:** “To debate and draft The Copenhagen Letter, 150 practitioners came together for 48 hours during Techfestival in 2017.”*
***Date**: September 10, 2017*
Source: https://copenhagenletter.org/

Tech is not above us. It should be governed by all of us, by our democratic institutions. It should play by the rules of our societies. It should serve our needs, both individual and collective, as much as our wants.

Progress is more than innovation. We are builders at heart. Let us create a new Renaissance. We will open and nourish honest public conversation about the power of technology. We are ready to serve our societies. We will apply the means at our disposal to move our societies and their institutions forward.

Let us build from trust. Let us build for true transparency. We need digital citizens, not mere consumers. We all depend on transparency to understand how technology shapes us, which data we share, and who has access to it. Treating each other as commodities from which to extract maximum economic value is bad, not only for society as a complex, interconnected whole but for each and every one of us.

Design open to scrutiny. We must encourage a continuous, public, and critical reflection on our definition of success as it defines how we build and design for others. We must seek to design with those for whom we are designing. We will not tolerate design for addiction, deception, or control. We must design tools that we would love our loved ones to use. We must question our intent and listen to our hearts.

Let us move from human-centered design to humanity-centered design.

We are a community that exerts great influence. We must protect and nurture the potential to do good with it. We must do this with attention to inequality, with humility, and with love. In the end, our reward will be to know that we have done everything in our power to leave our garden patch a little greener than we found it.

Oren Etzioni’s 3 Rules for Regulating Artificial Intelligence

Author: *Oren Etzioni, CEO of the Allen Institute for AI, for* The New York Times Op-Ed
Date*: September 1, 2017*
Source: https://www.nytimes.com/2017/09/01/opinion/artificial-intelligence-regulations-rules.html?mcubz=0

1. An A.I. system must be subject to the full gamut of laws that apply to its human operator.
2. An A.I. system must clearly disclose that it is not human.
3. An A.I. system cannot retain or disclose confidential information without explicit approval from the source of that information.

Isaac Asimov's *Three Laws of Robotics*

Author: *Isaac Asimov in his novel* Robot s*eries*
Date*: December 23, 1940*
Directed to robots, more of a “sci-fi twist” than actual principles.
(https://en.wikipedia.org/wiki/Three_Laws_of_Robotics)

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws
0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

Asilomar AI Principles

Author: *Developed by the [Future of Life Institute](https://futureoflife.org) through a consensus of the attendees at the [Beneficial AI 2017](https://futureoflife.org) conference in Asilomar, California. To date, the list has been been signed by 1200 AI/Robotics researchers and 2342 others.*
Date*: January 8, 2017*
Source: https://futureoflife.org/ai-principles/ Discussion on the principles by signatories: https://futureoflife.org/2017/01/17/principled-ai-discussion-asilomar/

1. Research
    a. **Research Goal**: The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.
    b. **Research Funding**: Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies.
    c. **Science-Policy Link**: There should be constructive and healthy exchange between AI researchers and policy-makers.
    d. **Research Culture**: A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.
    e. **Race Avoidance**: Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.

2. Ethics and Values
    a. **Safety**: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.
    b. **Failure Transparency**: If an AI system causes harm, it should be possible to ascertain why.
    c. **Judicial Transparency**: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

d. **Responsibility**: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

e. **Value Alignment**: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.

f. **Human Values**: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

g. **Personal Privacy**: People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.

h. **Liberty and Privacy**: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.

i. **Shared Benefit**: AI technologies should benefit and empower as many people as possible.

j. **Shared Prosperity**: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

k. **Human Control**: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.

l. **Non-subversion**: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.

m. **AI Arms Race**: An arms race in lethal autonomous weapons should be avoided.

3. Long-term

a. **Capability Caution**: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.

b. **Importance**: Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.

c. **Risks**: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.

d. **Recursive Self-Improvement**: AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.

e. **Common Good**: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.

Criticism about Asilomar principles:

- **“It's time for some messy, democratic discussions about the future of AI”**
  Jack Stilgoe and Andrew Maynard, The Guardian
  February 1, 2017
  Source:
  https://www.theguardian.com/science/political-science/2017/feb/01/ai-artificial-intelligence-its-time-for-some-messy-democratic-discussions-about-the-future

  a. The principles are **short on accountability**, and there are notable absences, including the **need to engage with a broader set of stakeholders and the public.**

  b. At the early stages of developing new technologies, public concerns are often seen as an inconvenience. In a world in which populism appears to be trampling expertise into the dirt, it is easy to understand why scientists may be defensive. This can get messy. It involves **engaging with people who may not see the world the same way**. But without such grounded approaches to responsible innovation, the chances of beneficial AI becoming a reality begin to dwindle. [...]

  c. [The principles] **lack the sophistication and inclusivity that are critical to responsive and responsible innovation…** they should be treated as hypotheses – the start of a conversation around responsible innovation rather than the end. They now need to **be democratically tested**.

- **"The Asilomar AI Principles Should Include Transparency About the Purpose and Means of Advanced AI Systems"**
  Bill Hibbard, H+ Magazine February 2, 2017
  Source:
  http://hplusmagazine.com/2017/02/02/asilomar-ai-principles-include-transparency-purpose-means-advanced-ai-systems/

    a. 7 and 8, call for **transparency** about the reasons for harm caused by AI and transparency about explaining judicial decisions made by AI. While helpful, these two principles are much **too narrow**.
    b. 19 and 20 describe the unlimited potential of AI to transform life on Earth… Many actions of advanced **AI systems will be too complex and subtle for people to perceive**. The only way for the public to know how advanced AI systems are affecting their lives and families is for the **purpose and means of all advanced AI systems to be made known to everyone**.
    c. **Trusting a scientific elite to follow the principles is not a democratic way of doing AI safety and ethics.**
    d. The Asilomar AI Principles **do not include transparency about the purpose and means** of advanced AI systems.
    e. Says nothing about people **controlling how the data they generate is used**." which doesn't provide people a reliable solution for protecting themselves against an organization equipped with AI manipulating them.

Satya Nadella's Six Principles

Author: *Satya Nadella, CEO of Microsoft, for* Slate
Date*: June, 28, 2017*
Source:
http://www.slate.com/articles/technology/future_tense/2016/06/microsoft_ceo_satya_nadella_humans_and_a_i_can_work_together_to_solve_society.html

1. AI must be designed to assist humanity.
2. AI must be transparent.

3. AI must maximize efficiencies without destroying the dignity of people.
4. AI must be designed for intelligent privacy.
5. AI must have algorithmic accountability.
6. AI must guard against bias.

World Economic Forum's Top 9 Ethical Issues in Artificial Intelligence

Author: *Julia Bossman*
Date*: Oct, 21, 2016*
Source: https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/

1. Unemployment
2. Inequality
3. Effect on behaviour
4. AI's mistakes
5. Machine Bias
6. Security and cybersecurity
7. "Evil genies"
8. Control over "super-intelligent" AI
9. Robots rights?

Yves Béhar's Ten Principles for the Design in the Age of AI

Author: *Yves Béhar, interviewed by Julia Bossman*
Date*: Jan, 30, 2017*
Source: https://www.fastcodesign.com/3067632/10-principles-for-design-in-the-age-of-ai

1. Design solves human problem, solutions are anthropomorphic.
2. Design is not based on historical cliches, it is though out in a context-specific mindset.
3. Technology should be designed to enhance human capabilities, not replace human in their tasks.

4. Technology should help and profit everyone in a given context (ex.: new device in a home must not be hated by some members of the family and loved by others.)
5. Good design is discreet design.
6. Good design changes, adapts and learns from users. (ex.: AI)
7. Products should be designed for long-term relationships with customers without being emotionally sticky.
8. When able to learn, tech should be able to improve well-being through predictions.
9. Design accelerates new ideas
10. Design should reduce complexity before trying to improve simplicity.

MIRI: The Ethics of Artificial Intelligence

Author: *Nick Bostrom and Eliezer Yudkowsky*
Date*: 2011*
Source: https://intelligence.org/files/EthicsofAI.pdf

“[Safe AI] involves new programming challenges, but no new ethical challenges.”
“Transparency is not the only desirable feature of AI. It is also important that AI algorithms taking over social functions be predictable to those they govern.”

1. For machines less than generally intelligent:
    a. Responsibility,
    b. Transparency
    c. Auditability
    d. Incorruptibility
    e. Predictability

   f. Tendency to not make innocent victims scream with helpless frustration.

“[These are] all criteria that apply to humans performing social functions; all criteria that must be considered in an algorithm intended to replace human judgment of social functions”

2. For AGI, different methods will be necessary because:
   a. The local, specific behavior of the AI may not be predictable apart from its safety, even if the programmers do everything right;
   b. Verifying the safety of the system becomes a greater challenge because we must verify what the system is trying to do, rather than being able to verify the system’s safe behavior in all operating contexts;
   c. Ethical cognition itself must be taken as a subject matter of engineering.
3. We must consider moral status for generally intelligent systems for non-discrimination reasons:

   a. *Principle of Substrate Non-Discrimination:* If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status.

   b. *Principle of Ontogeny Non-Discrimination:* If two beings have the same functionality and the same consciousness experience, and differ only in how they came into existence, then they have the same moral status.
4. Minds with exotic properties must be analysed and looked upon taking into account our own *conditioned* normative precepts and truths. Context is relevant in applying ethics principles.

Partnership on AI

*The partnership has 32 partners signed, including 7 founding partners.

Tenets

(https://www.partnershiponai.org/tenets/)

1. We will seek to ensure that AI technologies **benefit and empower as many people** as possible.
2. We will **educate and listen to the public** and actively engage stakeholders to seek their feedback on our focus, inform them of our work, and address their questions.
3. We are committed to **open research** and dialogue on the ethical, social, economic, and legal implications of AI.
4. We believe that AI research and development efforts need to be actively engaged with and accountable to a **broad range of stakeholders**.
5. We will engage with and have representation from stakeholders in the business community to help **ensure that domain-specific concerns and opportunities are understood and addressed**.
6. We will work to maximize the benefits and address the potential challenges of AI technologies, by:
    1. Working to **protect the privacy and security** of individuals.
    2. Striving to **understand and respect the interests of all parties** that may be impacted by AI advances.
    3. Working to ensure that AI research and engineering communities **remain socially responsible, sensitive, and engaged directly** with the potential influences of AI technologies on wider society.
    4. Ensuring that **AI research and technology is robust**, reliable, trustworthy, and operates within secure constraints.

    5. **Opposing development and use of AI technologies that would violate international conventions or human rights**, and promoting safeguards and technologies that do no harm.

7. We believe that it is important for the operation of **AI systems to be understandable and interpretable by people**, for purposes of explaining the technology.
8. We strive to create a **culture of cooperation, trust, and openness** among AI scientists and engineers to help us all better achieve these goals.

7 Thematic Pillars

(https://www.partnershiponai.org/thematic-pillars/)

A. Safety-critical AIs:

Advances in AI have the potential to improve outcomes, enhance quality, and reduce costs in such safety-critical areas as healthcare and transportation. Effective and careful applications of pattern recognition, automated decision making, and robotic systems show promise for enhancing the quality of life and preventing thousands of needless deaths. However, where AI tools are used to supplement or replace human decision-making, we must be sure that they are safe, trustworthy, and aligned with the ethics and preferences of people who are influenced by their actions.

We will pursue studies and best practices around the fielding of AI in safety-critical application areas.

B. Fair, transparent and accountable AIs:

AI has the potential to provide societal value by recognizing patterns and drawing inferences from large amounts of data. Data can be harnessed to develop useful diagnostic systems and recommendation engines, and to support people in making breakthroughs in such areas as biomedicine, public health, safety, criminal justice, education, and sustainability. While such results promise to provide great value, we need to be sensitive to the possibility that there are hidden assumptions and biases in data, and therefore in the systems built from that data. This can lead to actions and recommendations that replicate those biases, and suffer from serious blindspots.

Researchers, officials, and the public should be sensitive to these possibilities and we should seek to develop methods that detect and correct those errors and biases, not replicate them. We also need to work to develop systems that can explain the rationale for inferences.

We will pursue opportunities to develop best practices around the development and fielding of fair, explainable, and accountable AI systems.

C. Collaborations between people and AI systems:

A promising area of AI is the design of systems that augment the perception, cognition, and problem-solving abilities of people. Examples include the use of AI technologies to help physicians make more timely and accurate diagnoses and assistance provided to drivers of cars to help them to avoid dangerous situations and crashes.

Opportunities for R&D and for the development of best practices on AI-human collaboration include methods that provide people with clarity about the understandings and confidence that AI systems have about situations, means for coordinating human and AI contributions to problem solving, and enabling AI systems to work with people to resolve uncertainties about human goals.

D. AI, labor, and the economy:

AI advances will undoubtedly have multiple influences on the distribution of jobs and nature of work. While advances promise to inject great value into the economy, they can also be the source of disruptions as new kinds of work are created and other types of work become less needed due to automation.

Discussions are rising on the best approaches to minimizing potential disruptions, making sure that the fruits of AI advances are widely shared and competition and innovation is encouraged and not stifled. We seek to study and understand best paths forward, and play a role in this discussion.

E. Social and societal influences of AI:

AI advances will touch people and society in numerous ways, including potential influences on privacy, democracy, criminal justice, and human rights. For example, while technologies that personalize information and that assist people with recommendations can provide people with valuable assistance, they could also inadvertently or deliberately manipulate people and influence opinions.

We seek to promote thoughtful collaboration and open dialogue about the potential subtle and salient influences of AI on people and society.

F. AI social good:

AI offers great potential for promoting the public good, for example in the realms of education, housing, public health, and sustainability. We see great value in collaborating with public and private organizations, including academia, scientific societies, NGOs, social entrepreneurs, and interested private citizens to promote discussions and catalyze efforts to address society’s most pressing challenges.

Some of these projects may address deep societal challenges and will be moonshots – ambitious big bets that could have far-reaching impacts. Others may be creative ideas that could quickly produce positive results by harnessing AI advances.

G. Special Initiatives:

Beyond the specified thematic pillars, we also seek to convene and support projects that resonate with the tenets of our organization. We are particularly interested in supporting people and organizations that can benefit from the Partnership’s diverse range of stakeholders.

We are open-minded about the forms that these efforts will take.

IEEE Global Initiative: *Ethically Aligned Design,* General Principles

(http://standards.ieee.org/develop/indconn/ec/ead_general_principles.pdf)

As of May 2017, over a hundred members to the committee took part in writing the document, including 14 initiative members. (http://standards.ieee.org/develop/indconn/ec/ec_bios.pdf)

3 pillars

Ethical concerns applying to all type of AI/[Autonomous Systems] AS that:

A. Embody the highest ideals of human rights.
B. Prioritize the maximum benefit to humanity and the natural environment.
C. Mitigate risks and negative impacts as AI/AS evolve as socio-technical systems.

Principles

1. Human Benefits
    a. AI/AS should be designed and operated in a way that respects human rights, freedoms, human dignity, and cultural diversity.
    b. AI/AS must be verifiably safe and secure throughout their operational lifetime.
    c. If an AI/AS causes harm it must always be possible to discover the root cause (traceability) for said harm (see also Principle 3 – Transparency).
2. Responsibility
    a. people and institutions need clarity around the manufacture of these systems to avoid potential harm. Additionally, manufacturers of these systems must be able to provide programmatic- level accountability proving why a system operates in certain ways

3. Transparency
    a. Operation must be transparent to a wide range of stakeholders for different reasons (noting that the level of transparency will necessarily be different for each stakeholder).
    b. Intelligent systems will eventually be included [and impactful] in the physical world. Meaning potential harm to the user (automated cars, medical diagnosis, etc.)
    c. Lack of transparency both increases the risk and magnitude of harm (users not understanding the systems they are using) and also increases the difficulty of ensuring accountability.
    d. Transparency is important to each stakeholder group for the following reasons:
        i. For users, it builds trust in the system, by providing a simple way for the user to understand what the system is doing and why.
        ii. For validation and certification of an AI/AS, it exposes the system's processes for scrutiny.
        iii. If accidents occur, the internal process that led to the to the outcome can be understood.
        iv. Following an accident, judges, juries, lawyers, and expert witnesses involved in the trial process require transparency to inform evidence and decision-making.
        v. For disruptive technologies, such as driverless cars, a certain level of transparency to wider society is needed in order to build public confidence in the technology.
4. Education and Awareness
    a. Need for new kind of education for citizens to be sensitized to risks associated with the misuse of AI/AS. Such risks might include hacking, “gaming,” or exploitation (e.g., of vulnerable users by unscrupulous manufacturers).