

F.3. THE NEW POLITICS OF ARTIFICIAL INTELLIGENCE [preliminary notes]

MAIN MEMO

pp 3-14

- I. Introduction
- II. The Infrastructure: 13 key AI organizations
- III. Timeline: 2005-present
- IV. Key Leadership
- V. Open Letters
- VI. Media Coverage
- VII. Interests and Strategies
- VIII. Books and Other Media
- IX. Public Opinion
- X. Funders and Funding of AI Advocacy
- XI. The AI Advocacy Movement and the Techno-Eugenics Movement
- XII. The Socio-Cultural-Psychological Dimension
- XIII. Push-Back on the Feasibility of AI+ Superintelligence
- XIV. Provisional Concluding Comments

ATTACHMENTS

pp 15-78

ADDENDA

pp 79-95

REFERENCES

pp 96-99

DISCUSSION NOTES

pp 100-107

Richard Hayes
July 2018

DRAFT: NOT FOR CIRCULATION OR CITATION

ATTACHMENTS

- A. Definitions, usage, brief history and comments.
- B. Capsule information on the 13 key AI organizations.
- C. Concerns raised by key sets of the 13 AI organizations.
- D. Current development of AI by the mainstream tech industry
- E. Op-Ed: *Transcending Complacency on Superintelligent Machines* - 19 Apr 2014.
- F. Agenda for the invitational “Beneficial AI” conference - San Juan, Puerto Rico, Jan 2-5, 2015.
- G. An *Open Letter* on Maximizing the Societal Benefits of AI – 11 Jan 2015.
- H. Partnership on Artificial Intelligence to Benefit People and Society (PAI) – roster of partners.
- I. Influential mainstream policy-oriented initiatives on AI: Stanford (2016); White House (2016); AI NOW (2017).
- J. Agenda for the “Beneficial AI 2017” conference, Asilomar, CA, Jan 2-8, 2017.
- K. Participants at the 2015 and 2017 AI strategy conferences in Puerto Rico and Asilomar.
- L. Notes on participants at the Asilomar “Beneficial AI 2017” meeting.
- M. The “Asilomar AI Principles,” 11 Jan 2017.
- N. Key leadership in the AI+ superintelligence advocacy movement.
- O. interlocking relationships among key AI+ leaders and organizations.
- P. Op-Ed: Lord Martin Rees on his optimistic view of the human future.
- Q. Recent notable books and other publications and media addressing AI.
- R. Results of recent public opinion surveys addressing AI.
- S. Notes on funders and funding of AI advocacy
- T. The AI advocacy movement and the human techno-eugenics movement
- U. The socio-cultural-psychological dimensions of AI.
- V. Push-back on the feasibility of AI+ superintelligence.
- W. Provisional concluding comments: Discussion

ADDENDA

- A. AI+ Advocates’ Macro-Strategy and Short-Term Strategy
- B. Additional Topics for Possible Inclusion in these Attachments
- C. Update on Effective Altruism, Transhumanism, Longtermism and AI – June 2023
- D. Transhumanism as Precursor to and Major Influence on Development of EA and Longtermism – July 2023

REFERENCES

DISCUSSION NOTES

F.3. THE NEW POLITICS OF ARTIFICIAL INTELLIGENCE [preliminary notes]

I. INTRODUCTION

The last 4-5 years have seen an unprecedented surge of media, academic, commercial and philanthropic attention given to topics involving artificial intelligence, robotics and related technologies.¹ Press accounts herald the coming AI techno-utopia and warn of killer AI robot armies. New AI products are being deployed, new AI research initiatives are being lavishly funded, new AI advocacy organizations are being launched, new AI books appear almost weekly, governments are appointing AI study commissions and the United Nations has established an office to monitor AI developments worldwide.

Much of this accelerated public attention is being driven and guided by a small network of AI researchers, wealthy entrepreneurs and ideological advocates committed to a vision of the future in which human existence has been radically transformed, perhaps even overtaken and supplanted, by advanced AI and related technologies. As the impact of AI now begins to be felt more frequently and deeply in our daily lives, the AI techno-advocates know they need to do two things:

First, they need to ensure that nothing horrifically bad happens, e.g. the AI equivalent of a nuclear power plant meltdown, a lethal Bhopal disaster, a Jesse Gelsinger experimental tragedy, or anthropogenic climate change. Most AI advocates desire that the transformation of human existence proceeds safely.²

Second, they need to secure and maintain control of the public debate over the future of AI and over decisions regarding policy and governance both domestically and globally. The great majority of the world's peoples would oppose the radical transformation of human existence as envisioned by the AI advocates, were they made fully aware of what's planned and what's at stake. The AI researchers, wealthy entrepreneurs and ideological advocates are now taking steps to ensure that this doesn't happen, or if it does that it can be contained or overcome.

This memo is a first rough attempt to identify some of the main institutional and individual players in this new politics of artificial intelligence, outline the strategic thinking that guides their activity, note some of the major initiatives that they've established over the past several years, and very provisionally sketch some of the elements of a serious, pro-active response and alternative.^{3 4}

For notes on **DEFINITIONS, USAGE, BRIEF HISTORY AND COMMENTS** see **Attachment A**.

II. THE INFRASTRUCTURE: 3 Key AI Networks and 13 Key AI Organizations

There are hundreds of major tech companies, university labs, small private firms and independent advocacy organizations working on artificial intelligence. For these notes I distinguish three key AI networks:

A. The AI+ superintelligence advocacy network: This is a network of mostly smaller institutes, often affiliated with major universities, whose overriding objective is to help build a constituency and a movement in support of a future dominated by superintelligent AI entities. Most are aligned with transhumanist and singularitarian ideologies. Despite their small size they have recruited influential allies and have been framing and driving much of the recent public debate over the nature of and prospects concerning artificial intelligence.

B. The mainstream tech industry and academic AI research and advocacy network. This includes the market-dominant firms that are driving most of the real existing AI research, development and deployment: Google, Facebook, Apple, Amazon, Microsoft, IBM, Baidu, Alibaba and others, along with the top AI scientists and institutes at Stanford, MIT, University of California at Berkeley, Carnegie-Mellon and other key academic centers.

C. The liberal/progressive social accountability AI advocacy network: This is a set of mostly very new civil society NGOs established to ensure that AI and other I-tech develops in ways that promote, and don't undermine, social and economic justice, inclusion, diversity and civil and human rights. Right now they are few and mostly small. But they fill a key niche and are positioned to play an important role in the new politics of AI.

For concision I'll usually refer to these three networks as, respectively, the **AI+ advocates**, the **mainstream AI advocates**, and the **progressive AI advocates**. These names are provisional; see Endnote 5 for discussion.⁵

The AI future will be largely determined by the **mainstream AI advocates**. Thus, the strategy of both the **AI+ advocates** and the **progressive AI advocates** is to get the mainstream to adopt their visions, values and macro-strategy as its own. The vision, values and macro-strategy of the **AI+ advocates** are potentially the most dangerously consequential and are thus the main focus of this memo. A full understanding of the political terrain would call for a deeper treatment of the two other networks than this memo provides.

A. The AI+ Advocates

The eight AI+ advocacy groups shown here are mostly not directly involved in the development or application of AI hardware or software. Rather, they are engaged in academic studies, public communication, persuasion and advocacy, political strategy and macro-strategy, and ideological networking and base building.

1. [Machine Intelligence Research Institute](#) (2000) – Berkeley
2. [Future of Humanity Institute](#) (2005) – Oxford Martin School
3. [Centre for the Study of Existential Risk](#) (2012) - University of Cambridge
4. [The Future of Life Institute](#) (2014) - Boston
5. [Leverhulme Centre for the Future of Intelligence](#) (2015) – Oxford Martin School.
6. [Center for Human-Compatible Artificial Intelligence](#) (2016) – U.C. Berkeley
7. [Berkeley Existential Risk Institute](#) (2017) U.C. Berkeley
8. [OpenAI](#) (2015) – San Francisco

These are among the most consistently active groups, work in close collaboration, share common political perspectives and have multiple overlapping and intersecting leadership, staff, advisory personnel and funders. Most are formally associated with a university or are part of a university community. Most have fewer than a dozen full-time staff but have large teams of mostly grad student researchers, associates, post-docs, collaborators and advisers. Those associated with these eight organizations tend to share transhumanist, singularitarian or related values and beliefs.

The eighth organization listed, OpenAI, is of a different sort. It has a stated mission of developing the first Artificial General Intelligence (AGI) and releasing it as open source. The audacity of this mission and the size and source of its endowment (\$1 billion in pledges raised by Elon Musk and Sam Altman) suggest that aside from its stated mission OpenAI is in part intended to impact public and policy-maker understanding/attitudes/framing regarding AI. OpenAI could fail at creating AGI (most serious scholars and researchers believe that AGI is infeasible), yet still contribute to the triumph of an AI-dominant human future.⁶

The primary intended purposes of these eight AI+ organizations typically include:

- a) building a self-identified network of allies among academics, tech companies, funders and others, importantly including new cadres of young people;
- b) identifying ways in which AI technologies might, in fact, have objectively undesirable impacts, and working to build support of measures to avoid or counter these;

c) persuading the general public, policy makers and others that a future radically transformed by AI and other advanced technologies is a good thing, and is or can be made a safe thing;

d) influencing the political and policy-making process to ensure that measures contrary to the interests of the AI+ community are not adopted and that measures they call for or support are adopted.

Four of these organizations (1, 5, 6, 8) focus exclusively on AI. The other four organizations (2, 3, 4, 7) have a broader mandate of addressing *Existential Risks*, including risks posed by e.g. climate change, nuclear weapons and the threat of pandemics. For all of these four organizations, however, AI is currently the major focus.

B. The Mainstream AI Advocates

9. [Association for the Advancement of Artificial Intelligence](#) (1979)

10. [Partnership on Artificial Intelligence to Benefit People and Society](#) (2016)

The AAAI is the conventional mainstream nonprofit AI scientific society. It is “devoted to advancing the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines.” It had been eclipsed by aggressive, entrepreneurial AI and by AI+ advocates, and is working to catch up.

The PAI is a new association of 51 of the leading AI firms, research institutes and allied and other organizations. Its purpose is to address internal issues which these groups share as a community and to present a common face to the general public and the political world on both internal and external AI topics. Members include the mainstream AI groups, the AI+ advocates, the progressive AI groups and civil society groups such as the ACLU.

C. The Progressive AI Advocates

Three additional organizations differ from the ten listed above in that their priority objectives are to ensure that AI develops in a manner consistent with social and economic justice, inclusion, diversity and accountability. They are:

11. [Data and Society Research Institute](#) (2014) – New York City

12. [AI Now Institute](#) (2017) – New York University

13. [Upturn](#) (2017) – Washington D.C.

All three appear to take a somewhat more critical stance towards the AI prospect than do the ten AI groups just noted. The leadership, staff and advisors of these three organizations are notably more diverse and inclusive than are those of any of the other ten and include *none* of the ubiquitous AI+ ideologues, promoters and benefactors who overlap heavily among the AI+ advocacy groups.

There is, however, a major caveat. Most of the top leadership of these three progressive AI groups comes from the upper echelons of the tech/AI world: **Google, Microsoft, Apple, BuzzFeed**. That need not be a disqualification for critical leadership, of course, so long as the commitments to progressive values are strong and sincere.

[See DN 51.4 regarding a spate of very new NGOs and networks taking an even stronger “anti-Big Tech” position.]

For capsule descriptions of each of the thirteen AI organizations, see **Attachment B**.

The three categories of AI organizations each have sets of **priority concerns** that distinguish them from one another. For summaries of these concerns see **Attachments C.1** through **C.4**.

For capsule notes on the ways in which the mainstream AI groups are currently seeking to develop, use and commercialize AI, see **Attachment D**.

III. TIMELINE: 2005-present

Selected events from 2005-2012 are shown; more events are shown for the period 2013-present during which AI+ advocacy and controversy increased significantly.

2005-2012

May 2005: **Future of Humanity Institute** established as part of the new Oxford Martin School at Oxford University.

May 2006: Publication of *The Singularity is Near* by Ray Kurzweil.

Sep 2008: Singularity University established in Moffett Field, CA.

Feb 2009: Asilomar Conference on the Long-term Future of Artificial Intelligence (organized by the AAAI).

May 2011: IBM's Watson defeats two Jeopardy! Champions.

2013-present

Jan 2013: NYT piece by Huw Price: "Cambridge, Cabs and Copenhagen: My Route to Existential Risk."

Mar 2014: **Future of Life Institute** launched (Boston MA).

Apr 2014: *Huffington Post* blog post: "Transcending Complacency on Superintelligent Machines."
By Stephen Hawking, Max Tegmark, Stuart Russell and Frank Wilczek. See **Attachment E**.

May 2014: MIT Symposium on "Future of Technology" organized by the Future of Life Institute.
Moderator: Alan Alda; panel: George Church, Frank Wilczek, Jaan Tallinn, et al.

July 2014: Publication of *Superintelligence: Paths, Dangers, Strategies*, by Nick Bostrom.

Oct 2014: **Data and Society Research Institute** launched.

Jan 2015: 80 AI and AI+ researchers and advocates hold invite-only three-day meeting in San Juan, Puerto Rico, to plan research/media/political strategy; organized by Future of Life Institute. See **Attachment F**.

Jan 2015: *Open Letter on AI* published by Future of Life Institute. See **Attachment G**.

Jan 2015: Elon Musk announces \$10 million grant to Future of Life Institute for global program to "keep AI beneficial to humanity."

Apr 2015: Nick Bostrom TED talk: *What Happens When Our Computers Get Smarter Than We Are?*

Apr 2015: "Sophie the Robot" begins global round of media and event appearances [see below].

July 2015: *Open Letter* opposing AI in autonomous weapons published by Future of Life Institute.

Dec 2015: **Leverhulme Centre for the Future of Intelligence** launched at the Oxford Martin School.

Dec 2015: Elon Musk and Sam Altman launch **OpenAI**, with \$1 billion pledged endowment, to build "strong" AI and distribute it free to all as open source.

Mar 2016: Google's DeepMind AlphaGo AI beats world Go champion Lee Sedol, in Korea.

Sep 2016: **Center for Human-Compatible Artificial Intelligence** launched at U.C. Berkeley, with \$5.5 million from **Open Philanthropy Project**, the Leverhulme Trust and the Future of Life Institute.

Sep. 2016: **Partnership on AI to Benefit People and Society** launched by Google, Microsoft, Amazon, Facebook, IBM and Apple, with ~ 40+ partner organizations. See **Attachment H**.

Sep 2016: *Artificial Intelligence: The 100 Year Study*, released by Stanford University.

Oct 2016: White House Office on Science and Technology Policy releases its report, *Preparing for the Future of Artificial Intelligence*. See **Attachment I**.

- Jan 2017: “Beneficial AI 2017” conference at Asilomar (Jan 2-8), with Jan 11 release of “Asilomar AI Principles.” Organized by Future of Life Institute. See the agenda, a list of and notes on participants, and the “Principles” in **Attachments J, K, L and M**.
- Apr 2017: **Berkeley Existential Risk Institute** launched.
- Aug 2017: *Open Letter* calling on UN to ban lethal autonomous weapons published by Future of Life Institute.
- Sep 2017: United Nations creates office in The Hague to “monitor threats from AI and Robotics,” under the UN Interregional Crime and Justice Research Institute (UNICRI).
- Nov 2017: **AI Now Institute** launched.
- Dec 2017: **Upturn** launched.
- Feb 2018: Future of Humanity Institute and allied groups release *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*.

Timeline: Sophie the Robot

- Apr. 15, 2015: “Activated” by Hanson Robotics (Tokyo).
- Mar. 2016: Appears at SXSW, Austin, TX; first of many TV and event appearances in 2016-2017.
- Oct. 2016: Interviewed by Charlie Rose on 60 Minutes.
- Dec. 2016. Appears on the cover of *ELLE* magazine, Brazil.
- Apr. 2017: Appears on Jimmy Kimmel show. David Hanson tells Kimmel that Sophie is “basically, alive.”
- Oct. 2017: Saudi Arabia grants citizenship to Sophie the Robot.
- Nov. 2017: United Nations Development Program (UNDP) names Sophie an “Innovation Champion” for 2018.
- Jan. 2018: Facebook’s director of artificial intelligence, Yann LeCun, says Sophie is “complete bullshit” and slams media coverage and promotion of “Potemkin AI.”

IV. KEY LEADERSHIP

The 9 individuals listed below are playing key leadership roles in promoting the vision of humanity transformed by AI+ superintelligence as a solution to mounting civilizational crises. See **Attachment N** for brief biographical information. See **Attachment O** for their inter-locking associations within and among the 8 key AI+ advocacy organizations.

Nick Bostrom: philosopher and director of the Future of Humanity Institute at Oxford; author of *Superintelligence*.

Elon Musk: engineer/entrepreneur known for PayPal, Tesla, SpaceX, Solar City, Neuralink, OpenAI and more.

Seán Ó hÉigeartaigh: director of the Centre for the Study of Existential Risk at Cambridge; PhD in genomics.

Huw Price: philosopher; Academic Director of existential risk institutes CSER and LCFI at Cambridge.

Lord Martin Rees: noted Cambridge cosmologist; former Royal Society president, author of *Our Final Century?*

Stuart Russell: noted UC Berkeley computer scientist; founder, Center for Human-Compatible AI (CHAI).

Jaan Tallinn: engineer known for Skype and Kazaa; philanthropist; co-founder of CSER and FLI.

Max Tegmark: cosmologist at MIT, author of *Humanity 3.0*, co-founder of the Future of Life Institute in Boston.

George Church: professor of genetics at Harvard Medical School; also on faculty at Harvard University and MIT.

Comment 1: A key thing to know about the current AI debate is that those playing the most active public roles in warning society about “the dangers of AI” are in fact the AI+ advocacy groups, i.e., those most deeply committed to its development and, eventually, its foundational role in bringing about a superintelligent/singularitarian/post-human world. In playing this role the AI+ advocates establish themselves as the responsible leaders of the public debate and take control of its direction. The many recent reports and other media communications on the dangers

of AI that they've helped construct all convey a common message: that AI will be either the best or the worst thing that ever happens to humanity, and that by taking the proper actions now we can avoid the worst and realize the best. The worst is extinction. The best is the replacement of all humanity with immortal, omniscient, omnipotent Post-Humans. See **Attachment E**, Stephen Hawking et al's April 2014 op-ed, and **Attachment P**, Lord Martin Rees' August 2016 op-ed, as models of this messaging.

V. OPEN LETTERS

The Future of Life Institute organized and publicized a series of *Open Letters* advocating national and international policies regarding AI. These five letters helped build the network of politically active AI organizations, researchers and supporters, and generated much favorable press attention. All are posted on the FLI website.

1/1/15: ***Open Letter on maximizing the societal benefits of AI.*** 8,000 signers by Jan 2018.

6/4/15: ***Open Letter on addressing the economic impacts of AI.*** Initially signed by 70 economists; 1,000+ signers by Jan 2018.

7/28/15: ***Open Letter urging ban on offensive autonomous weapons beyond meaningful human control.*** Initially signed by ~ 50 AI/robotics researchers + 17 others. Signed by 3,722 researchers and 20,467 others by Jan 2018.

1/30/17: ***The Asilomar AI Principles to guide beneficial AI development.*** 3,817 signers by Jan 2018.

8/20/17: ***Open Letter calling for UN ban on lethal autonomous weapons.*** Signed by 82 founders/CEOs/Presidents of AI/Robotics companies.

VI. MEDIA COVERAGE

Beginning in ~ 2014 media coverage of the "dangers of AI" and related themes accelerated dramatically, driven by the apparent face-value import of the subject matter, the marquee value of those speaking out, and the intentional efforts of the AI+ advocacy community to generate media coverage:

12/2/14: **Stephen Hawking warns artificial intelligence could end mankind.** *BBC News.*

1/29/15: **Bill Gates on dangers of artificial intelligence: 'I don't understand why some people are not concerned.'** *Washington Post.*

7/27/15: **Musk, Hawking Warn of Artificial Intelligence Weapons.** *Wall Street Journal.*

7/28/15: **Tech Experts Warn of Artificial Intelligence Arms Race In Open Letter.** *NPR All Things Considered.*

2/15/16: **Scientists Warn that Robots and Artificial Intelligence Could Eliminate Work.** *Fortune.*

7/1/16: **Elon Musk-backed Group Awards Millions in Grants to Keep Us Safe from AI.** *NBC News.*

4/1/17: **Elon Musk's Billion-Dollar Crusade to Stop the AI Apocalypse.** *Vanity Fair.*

7/17/17. **AI is the biggest risk we face as a civilization, Elon Musk says.** *The Telegraph.*

11/14/17: **Killer robots are almost a reality and need to be banned, warns leading AI scientist.** [Stuart Russell, UC Berkeley]. *The Telegraph.*

2/21/18: **Control AI now or brace for nightmare future, experts warn.** *CNN.*

These warnings generated push-back. Many said that the apocalyptic scenarios were greatly exaggerated, completely non-credible, or could be avoided with appropriate policies and safeguards:

10/31/14: **Don't Fear Artificial Intelligence.** *Slate.*

10/24/15: **Don't Fear the Robots.** *New York Times*.

12/22/16: **Don't Fear the Robots:** they won't kill jobs. *Reuters*.

7/21/17: **Don't Fear the Robots:** Smart machines will replace some jobs, but they will create many more by generating new wealth and higher demand for products and services. *Wall Street Journal*.

1/29/18: **Why Will.i.am Isn't Afraid of Artificial Intelligence.** Report from Davos. *CNBC*.

Some AI+ advocates used the identical lead message to deliver precisely the *opposite* core message, saying that we needn't fear AI/robots precisely because they *will* overtake humanity, and we can take steps now to ensure that we merge with them or otherwise have them transform us into greatly enhanced immortal technological entities:

12/19/14: **Don't Fear Artificial Intelligence.** Ray Kurzweil, *Time*.

11/7/17: **Famous Futurist Explains Why We Shouldn't Fear AI:** Google engineer Ray Kurzweil says intelligent machines will enhance humans, not replace us. *Futurism*.

Much coverage was given in 2017 to strongly expressed differences between Mark Zuckerberg and Elon Musk:

7/24/17: **Mark Zuckerberg: Elon Musk's doomsday AI predictions are 'pretty irresponsible'.** *CNBC*. |

7/25/17: **Elon Musk says Mark Zuckerberg's understanding of AI is 'limited'.** *CNN*.

7/25/17: **Zuckerberg and Musk are both wrong about AI.** *Ars Technica*.

9/12/17: **Sam Altman: why Mark Zuckerberg and Elon Musk are both right about A.I.** *CNBC*.

Comment 2: There is something inherently amiss when debate over an important socio-political concern understands the range of contending opinion to be exhausted by Elon Musk on one end and Mark Zuckerberg on the other.

Comment 3: Most AI+ advocates agree that the “warning of AI dangers” strategy has paid off well. Leading with the “dangers” gets them far more media invitations, and a more attentive public audience, than they'd get if they sought to downplay the dangers and spoke only of the bright side. But as the Zuckerberg/Musk exchanges show, there are other assessments. Some worry that the “AI dangers” strategy runs the risk of unduly alarming the public and policy makers and inviting undesired regulations. Many mainstream AI advocates dislike the “AI dangers” strategy because they feel it's dishonest, which much of the time it is (see Section XIII below, and **Attachment V**). Other mainstream AI advocates acknowledge the exaggeration but argue that there *are* real risks that need to be addressed and that the publicity helps motivate action. Finally, a significant minority of the mainstream AI community aligns personally with the AI+ community, and has to navigate the resulting tension.

Comment 4: What do the AI+ advocates mean when they speak of ensuring AI safety? First, as noted, they know that an AI experiment gone wrong could set them back, and want to avoid that. Second, they want to solve the “value alignment problem.” This means figuring out how to program an AI so that any decision it makes will be in alignment with human values. No-one is quite sure what this means or if it's at all technically possible. Third, they know that the AI+ vision of the future generates public unease; they propose to address this with squads of AI+ ethicists, bioethicists, social ethicists, policy experts and others installed at every level of research, business and government, all young PhD philosophers, legal scholars and social scientists trained at the AI+ network institutes, and all adept at generating reports that appear to address public concerns without ever threatening the AI+ agenda itself.

VII. INTERESTS AND STRATEGIC IMPERATIVES

1. The **AI+ advocacy community** is motivated by a sincere belief that the development of human-level AI will quickly lead to Superintelligence, which in turn will quickly bring on to the Singularity and transform the planet into a world of health, wealth, happiness, sustainability, immortality, omniscience, omnipotence and expeditions to Alpha Centauri and beyond.

The strategic imperative of the AI+ advocacy community is three-fold. It needs to **1) continue to build its global cadre base; 2) get the AI mainstream to adopt the AI+ vision as its own; and 3) win the war of ideas at the level of the educated general public.** These imperatives are challenging but not outlandish. Significant overlap already exists among the AI+ and the mainstream AI communities; the model of small but well-funded institutes at leading universities is scalable; and the vision that the AI+ advocates are offering the world is, they believe, a straightforward extension of the Enlightenment vision that has dominated world events for the past three centuries and, despite periodic setbacks, continues to.⁷

2. The priority interest of those in the **mainstream AI community** is to become as personally wealthy as possible. Secondly they want to “make the world as awesome as possible” through AI and other I-tech. “Awesomeness” lies along a continuum. At one end is the conventional vision of technological progress and economic well-being held by the great majority of the world’s peoples. At the other end is the vision of the AI+ superintelligent/singulitarian/posthumanism.

The strategic imperative of the mainstream AI community is to stay alert and **not make any stupid mistakes.** They hold the cards and they own the table, the chairs and much of the rest of the house.

3. Those in the **progressive AI community** share a strong primary desire to help achieve a world of social and economic justice, inclusion and human rights. Many are also motivated by the conventional vision of technological progress and economic well-being held by the great majority of the world’s peoples. They don’t mind living materially comfortable lives, but they don’t feel driven to become billionaires.

The strategic imperative of the progressive AI groups is to **not fall off the tightrope** they have to walk vis-à-vis the mainstream tech community on the one side and the grassroots constituencies, the academy, the foundations, the media and the political class on the other. They need to be sensitive enough to the interests of the mainstream to be able to deal effectively when the time is right, and they need to be tough and confrontational enough with the mainstream to maintain the respect of their base and thus their legitimacy and clout.

In this schema the primary motivations of the three groups are distinct but there are overlaps between 1 and 2 and between 2 and 3. So far as I can see there’s little overlap between 1 and 3.

VIII. BOOKS AND OTHER MEDIA

Kevin Kelly’s 1994 *Out of Control: The New Biology of Machines, Social Systems and the Economic World* and Ray Kurzweil’s 1999 *The Age of Spiritual Machines: When Computers Exceed Human Intelligence* established a model of accessible writing in support of preposterous futuristic scenarios, grounded in vaguely center-left libertarian, market-friendly values, that has served as a template for scores of popular books on technology since. Following Nick Bostrom’s 2014 *Superintelligence: Paths, Dangers, Strategies*, and the growing presence of AI-enabled online features and consumer devices, the publication of such books on AI accelerated. By 2017, however, unease over Silicon Valley’s political, economic and cultural dominance, its role in the 2016 elections and successive data breach exposés reached a new high, and books taking strongly critical positions began to be published. See **Attachment Q** for a list of recent publications and other media.

IX. PUBLIC OPINION ON ARTIFICIAL INTELLIGENCE

I reviewed a collection of public opinion surveys addressing aspects of AI. See **Attachment R** for the full review. Selected results are summarized here.⁸

People have mixed feelings about AI. On one hand they generally express support, approval and optimism regarding AI and AI research overall. At the same time they express varying degrees of unease, disapproval and opposition regarding many aspects and applications of AI.

One survey reported 77% of respondents to be “optimistic” about the impact of AI on life and work in the future. Another asked if we should “increase or decrease our reliance on AI,” to which 39% of respondents said *increase*, 39% said *decrease*, and 23% *didn’t know*. Large majorities (~70%) support national and international AI regulation.

Respondents had a generally dismal assessment of the impact of AI on jobs and the economy. Majorities of 60-70% and higher said they believed that AI would increase unemployment and lead to greater economic inequality. At the same time, fewer respondents (20-30%) feared that AI would put their own jobs at risk.

Opinion was split ~ 53-46 in support of a Universal Basic Income as a response to permanent job loss generated by AI. However, this overall response concealed a stronger partisan divide: Democrats strongly supported UBI, by 65-35%, while Republicans even more strongly opposed it, by 28-72%.

Respondents seemed more accepting of AI for routine applications, such as “cleaning a house” (61%), or for applications unlikely to involve them personally, frequently or directly, such as “facial recognition computers to help catch criminals” (61%). They seemed less accepting of applications of potentially high personal consequence, such as “making a medical diagnosis” (27%), “driving a car” (24%), “flying an airplane” (23%), or “choosing a romantic partner” (23%).

Very generally, and not in all instances, younger, better educated, higher income respondents tended to be more approving of or comfortable with AI than were older, less educated, lower income respondents. But the margins of difference were often small. In some surveys better educated respondents more often replied “Don’t Know.”

Little variance was seen in responses from those identifying with different religious traditions, and likewise between religious and non-religious respondents. All roughly tracked population-wide responses.

Ideological liberals were generally more supportive of AI than were ideological moderates or conservatives. However, little difference of opinion was seen among Republicans, Democrats and Independents.

Across several questions Whites and African Americans appeared to be similarly uneasy with AI, while Hispanics and Others (presumably including Asians, Pacific Islanders, South Asians and Native Americans) appeared more receptive. On other questions e.g. whether AI is an existential threat racial/ethnic differences were not visible.

Among the few large differences that appeared repeatedly and consistently were the different responses from men and women:

	<u>Men</u>	<u>Women</u>
Feel we should increase our reliance on AI:	52	26
Feel that AI is safe:	54	29
Feel that AI will help the economy:	38	18

These gender differences transcended partisan identification. Republican, Democratic and Independent men agreed, by an average of 52%, that we **should** increase our reliance on AI, while Republican, Democratic and Independent women agreed that we **should not**, by an average of 28%.

It’s unclear what the few and thinly reported survey results reviewed here actually tell us. The topic of AI is new and technical and much of the media coverage has been sensationalistic. Should we read public opinion on AI in the same manner as we read it on longstanding topics such as, say, abortion, minimum wage and government spending? At a minimum, we might expect that current opinion on AI is thinly held and subject to change.

X. FUNDERS AND FUNDING OF AI ADVOCACY

See **Attachment S.1** for a review of funding strategies for the key AI organizations. Highlights are shown here.

A. Foundations and other institutions supporting one or more of the AI+ advocacy organizations:

Leverhulme Foundation	Ethereum	<i>The Open Philanthropy Project</i>
Global Challenges Foundation	Templeton World Charity Fdn	<i>Giving What We Can</i>
CITRIS	Hauser-Raspe Foundation	<i>Center for Effective Altruism</i>
Foundational Questions Institute	Kenneth Miller Trust	<i>GiveWell</i>
Berkeley Existential Risk Institute	Libra Foundation	<i>Effective Altruism Funds</i>
Future of Life Institute	Milner Foundation	<i>Effective Altruism Giving Fund</i>
Musk Foundation	Blavatnik Family Foundation	
The Thiel Foundation	The Grantham Foundation	

B. The funders shown in *italics* in the *third column above* are part of a growing philanthropic movement known as **effective altruism**. It purports to make philanthropic decisions using rational and scientific criteria free of emotional or subjective bias. It appeals to idealistic young professionals with a penchant for science and technology. Many “effective altruists” have come to believe that since Artificial Intelligence will be either the greatest or the worst thing that ever happens to humanity, a rational analysis suggests that efforts to avoid the second outcome and secure the first should receive high priority for charitable giving. See **Attachment S.2** for more.

C. Foundations helping support the three progressive AI advocacy organizations include Microsoft, MacArthur, MIT Media Lab, Gates, Sloan, Ford, Open Society, Robert Wood Johnson and Kellogg. These three groups also receive support from a new funder consortium, the [Ethics and Governance of Artificial Intelligence Fund](#), established by Knight, Omidyar Network, Hewlett, Reid Hoffman, Jim Pallota and others. See **Attachment S.3** for more.

XI. THE AI ADVOCACY MOVEMENT AND THE TECHNO-EUGENICS MOVEMENT

Many AI advocates are also strong advocates of human genetic modification, “designer babies” and human cloning. These cross-disciplinary advocates include scientists, leaders in both the I-tech and biotech industries, academics and major funders. For many of them the complementary nature of genetic modification and AI seems intuitively obvious: they both involve the use of powerful technological tools to greatly enhance human abilities and well-being. A more critical account would note that they both involve a deeply reductionist understanding of the human body, mind and person; generally libertarian social, cultural and political sensibilities; a premium placed on cognitive ability; and very often fantasized scenarios of personal immortality, omniscience and omnipotence. Many of the major I-tech leaders now moving heavily into AI are simultaneously committing funding and attention to topics involving human genetic modification. See **Attachment T** for more.

XII. THE SOCIO-CULTURAL-PSYCHOLOGICAL DIMENSION

Since the beginning of mass use of computers, the internet and then mobile devices and I-tech in general, critics have wondered if these dramatic changes in the media through which we communicate might not carry with them some unexpected and undesirable changes in our social, cultural and psychological lives as well. Efforts to document such impacts have been suggestive but inconclusive. Recently, however, solid evidence seems to link a sharp rise in personality and behavioral disorders among post-Millennials, those born after 1995, with the use of smartphones and related mobile devices and the concomitant rise of social media and screen time. These studies also show a sharp decrease in time spent in physical social interaction. For more see **list Q.2** in **Attachment Q**.

Initiatives are underway to discourage inordinate mobile device and social media use and screen time. Other initiatives go further and advocate radical unplugging from the ceaseless online, mediated world that constitutes the near entirety of many waking lives. It's possible to imagine that discontent with the impacts of today's hyper-technocentric civilizational paradigm, including its growing socio-cultural-psychological disorder, its undermining and sabotage of democratic political institutions, its accelerating technologically-generated growth of economic inequality and more could at some point reach critical mass and spark a transformational change quite different from that imagined by the AI+ superintelligence advocates. For more see **Attachment U**.

XIII. PUSH-BACK ON THE FEASIBILITY OF AI+ SUPERINTELLIGENCE

The frenzied attention given the prospect of AI+ Superintelligence, whether as existential threat or utopian salvation, typically glosses over the fact that the great majority of scientists, engineers and thoughtful others simply don't believe that AI+ Superintelligence is feasible. Real intelligence requires consciousness, volition, intentionality and desire. Even the most sophisticated AI is just software, and as such has no more consciousness, volition, intentionality or desire than does a pocket calculator or a doorknob.

That's not to say that AI doesn't have the potential to be transformationally consequential, for good or for ill, and that society doesn't need to take initiatives to ensure its proper development and use. It certainly does and we certainly do. But this is all the more reason to put the distracting fantasies aside and to focus on the real and difficult work that needs to be done. Credible AI scenarios can still be dystopic. Humans are capable of inflicting great harm on other humans, and those with access to AI will likely be able to inflict harms of unprecedented scope and scale on those who don't. See **Attachment V** for more on the case *against* AI+ Superintelligence as a real danger, but *for* action to ensure that AI develops in a manner consistent with widely held social values.

XIV. PROVISIONAL CONCLUDING COMMENTS

1. There is no danger of an advanced AI waking up one morning and deciding to enslave or destroy us. It can't and won't ever happen.
2. There is, however, great danger that AI and other emerging and converging technologies could be used by people with access to those technologies to greatly harm people without such access, perhaps catastrophically.
3. There is also the danger that AI and related advanced technologies might accidentally cause great harm.⁹
4. There is a further danger that what might at first appear to be benign and beneficial uses of AI could turn out to have malign and harmful impacts on society, culture, personhood, cognitive well-being or other valued human institutions and attributes.¹⁰
5. There is an additional danger that the incessant, reductionist and disempowering teaching that humans are nothing but machines, and that machines can be made that are equivalent or superior to humans, could contribute to an erosion of the sense of human autonomy and dignity that allows human flourishing.¹¹
6. To address the first real danger there needs to be strong legal oversight and control authority over the development and use of technologies that could be used in dangerous ways. At the limit this includes the authority to prevent selected technologies from being developed in the first place.
7. To address the second and third dangers we need a precautionary approach to all development and use of AI, and structures and processes to slow down, modify or in the limit relinquish, selected technologies when this appears warranted.
8. The fourth danger might be addressed in part in the course of addressing the first three. But we will also need independent initiatives intended to restore our understanding of what it means to be truly and fully human.

9. The claim that powerful AI technologies of the sort now being contemplated can't be regulated or prevented from being developed or used is specious. If enough people believe that such controls are warranted these can be imposed and enforced. Approaches to the governance of advanced technology will be discussed in more detail in the final working paper.¹²

10. Governance of AI technologies faces a unique challenge, however, in that their globalization, mass use and deep embeddedness could very well weaken or undermine the social, socio-psychological and political structures that would be necessary to formulate, implement and enforce such governance.

11. It is also possible that the potential dangers of certain AI technologies are so profoundly unacceptable, yet so *challenging* to oversee and control, that we would feel compelled to examine the conditions that gave rise to this quandary in the first place. This examination, in turn, could lead to a critical re-evaluation of the nature of modernity in all its forms and manifestations. This possibility is discussed further in Section IV of the working paper outline.

12. The fact that so many highly educated people appear to come so quickly and easily to the belief that an advanced AI could someday wake up and decide to enslave or destroy us is a concern in itself. This belief is generally associated with wider belief systems involving a superintelligent/singulatarian/post-humanist future. Such beliefs are serving to channel an influential sector of the human community towards anti-social, elitist, corrosively materialist/reductionist and very often nihilistic frames of thought. In the worst case such beliefs can serve to motivate and rationalize exclusionist, separatist and genocidal fantasies.

See **Attachment W** for discussion of these provisional concluding comments.

ATTACHMENTS

- A. Definitions, usage, brief history and comments.
- B. Capsule information on the 13 key AI organizations.
- C. Concerns raised by key sets of the 13 AI organizations.
- D. Current development of AI by the mainstream tech industry
- E. Op-Ed: *Transcending Complacency on Superintelligent Machines* (19 Apr 2014).
- F. Agenda for the invitational “Beneficial AI” convening - San Juan, Puerto Rico, Jan 2-5, 2015.
- G. An *Open Letter* on Maximizing the Societal Benefits of AI – January 11, 2015.
- H. Partnership on Artificial Intelligence to Benefit People and Society (PAI) – roster of partners.
- I. Influential mainstream policy-oriented initiatives on AI: Stanford (2016); White House (2016); AI NOW (2017).
- J. “Beneficial AI 2017” conference, Asilomar, CA, Jan 2-8, 2017.
- K. Participants at the 2015 and 2017 AI strategy conferences in Puerto Rico and Asilomar.
- L. Notes on participants at the Asilomar “Beneficial AI 2017” meeting.
- M. The “Asilomar AI Principles,” 11 Jan 2017.
- N. Key leadership in the AI+ superintelligence advocacy movement.
- O. The GRID: interlocking relationships among key AI+ leaders and organizations.
- P. Op-Ed: Lord Martin Rees on his optimistic view of the human future.
- Q. Recent notable books and other Publications and media addressing AI.
- R. Results of recent public opinion surveys addressing AI.
- S. Notes on funders and funding of AI advocacy.
- T. The AI advocacy movement and the human techno-eugenics movement
- U. The socio-cultural-psychological dimensions of AI.
- V. Push-back on the feasibility of AI+ superintelligence.
- W. Provisional concluding comments: Discussion

ADDENDA

- A. AI+ Advocates’ Macro-Strategy and Short-Term Strategy
- B. Additional Topics for Possible Inclusion in this Attachment
- C. Update on Effective Altruism, Transhumanism, Longtermism and AI – June 2023
- D. Transhumanism as Precursor to and Major Influence on Development of EA and Longtermism – July 2023

REFERENCES

DISCUSSION NOTES

ATTACHMENT A. DEFINITIONS, USEAGE, BRIEF HISTORY AND COMMENTS

BACKGROUND

Many reject the idea that machines can be intelligent in the sense that most people have long understood that word. On this view intelligence requires, at a minimum, such features as consciousness, self-awareness and volition. Many would add that the concept of intelligence would be meaningless in the absence of intension or desire as well. For many this leads to the affirmation that intelligence is an embodied, evolved property of animal life, and cannot be a property of machines.

Many others, including most scientists and engineers working on AI, are comfortable saying that machines can be intelligent. On inspection however, it's seen that most of them are using *intelligence* in a sense closer to what most others would call *computation*. All that computation requires is input, a set of rules for processing the input (an *algorithm*) and output. At no point is there any need to invoke consciousness, self-awareness or volition, much less intension or desire, for a machine to successfully compute.

There's no reason that machines using creative algorithms can't produce outputs that mimic human cognition and behavior across many output domains. But neither is there any reason to believe that these machines are any more conscious, self-aware or volitional than is a pocket calculator or a doorknob.¹³

Stanford AI instructor Jeffrey Kaplan (2016) says that "The essence of intelligence is the ability to make appropriate generalizations in a timely fashion based on limited data. The broader the domain, the quicker, the more intelligent." By this understanding intelligence doesn't require consciousness, self-awareness or volition.

So we have a choice. We can reserve the word *intelligence* as a quality unique to animal life and use *computation* for what machines do. (Animal brains can also compute but that's an appropriately mechanical process.) Or we can say that both machines and animal life can exhibit intelligent behavior, always being careful to distinguish between *human/animal intelligence* and *artificial or machine intelligence*.

The descriptor *artificial intelligence* was coined by Dartmouth professor of computer science John McCarthy in 1955 in a proposal for a 1956 summer conference intended to see if newly developing computer technology could be used to process abstract symbols in logical ways, and thus mimic certain patterns of human thought. McCarthy needed a term to distinguish his proposed use of symbolic logic and digital computation from the then better-known work of MIT professor Norbert Weiner, who called his largely analog-based field of research *cybernetics*.

In the ensuing few years *artificial intelligence* supplanted *cybernetics* within the research community and, appropriately, became widely used in reportage. But the term *artificial intelligence* further enabled the seemingly irresistible tendency among both researchers and journalists to misleadingly anthropomorphize any discussion of anything involving computers. They were said to be "electronic brains" or "artificial brains" that could "think like humans," but in truth they were nothing of the sort.

See Attachment V for more on use of the term *artificial intelligence*.

AI TODAY

With this background here are some key terms and the ways in which we'll be using them in these notes.¹⁴

Artificial Intelligence (AI) refers to a type of information technology that seeks to simulate some or many of the cognitive abilities human beings exhibit when solving problems and acting in the world. **Machine Intelligence** is a synonym.

Artificial General Intelligence (AGI) is AI technology that can simulate any and all human cognitive abilities. **Strong AI** and **AI-Complete** are synonyms for AGI. **Weak AI** is any AI more limited than AGI. Some authors take AGI,

Strong AI and AI-Complete to imply that the machines exhibiting these behaviors are **conscious**. Most AI researchers, however, following Turing (1950), maintain that the presence or absence of consciousness is irrelevant to AI, in part because neither state is verifiable.

Superintelligence is defined by Bostrom (1998) as “an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills.” As such it includes but goes beyond AGI. Bostrom notes that the definition is agnostic regarding subjective experience.¹⁵

As used in this memo, **Robots** are machines that use AI to act in the world. **Industrial** robots work on assembly lines or in other workplace situations. **Humanoid** robots are intended to look and act like humans. **Social** robots are intended to interact with humans, e.g. as care providers. The three types are not exclusive: an industrial robot may or may not be humanoid, a humanoid robot may be either social or industrial, etc. **Companion** robots are social robots designed to provide simulated emotional or personal relationship.

It’s important to recognize that AI is *not* a well-defined, coherent scientific or engineering discipline. There is no common body of AI theory. Kaplan characterizes AI as a “grab-bag” of techniques, approaches and applications.

McCarthy’s initial approach using symbolic logic sought to directly simulate the rational cognitive processes by which humans solve problems and act in the world. This is now referred to as **GOFAI**, for “good old fashioned artificial intelligence.” It proved to be too cumbersome and complicated for use beyond primitive applications.

A second set of approaches, referred to generally as **machine learning**, are very different from GOFAI. Machine learning is “the branch of AI computing that involves training algorithms to perform tasks by learning from previous data and examples rather than explicit commands programmed by humans.”¹⁶ An important attribute of a machine learning algorithm is thus the ability *to improve itself*. Machine learning accounts for the great bulk of the AI developments that are today drawing the most attention.

Among the most intriguing approaches to machine learning are those involving **artificial neural networks**. This approach is modeled on the ways in which neurons in the brain process information. As an artificial neural network is repeatedly exposed to images of, say, dogs, it builds up an embedded statistical record of patterns of pixels associated with “dog.” If it is then shown an image that it has never seen before, it can “decide” how likely that image is to include a dog. The more and varied the images it has been “trained” on – thousands, millions or billions of times – the more accurate will be its “decision.”

The idea of programming computers to act like neural networks is not a new one. It was originally proposed in the late 1940s but like McCarthy’s GOFAI it proved computationally difficult given the computer power available at the time. It developed in fits and starts for the next three decades and remained little more than an intriguing curiosity.¹⁷

Several developments that transpired during the 1990s and 2000s enabled artificial neural networks and other approaches to machine learning to gain traction and generate the research and applications that are now the focus of so much attention. These developments have included:

- large increases in computer processing power, i.e. in computer **speed and memory**;
- the gains in processing power and storage achieved by moving from physically to **electronically stored data**;
- easier **access to data** (mainly due to the internet);
- low-cost high-resolution **digital sensors**;
- the availability of huge, crowd-sourced data sets (**“Big Data”**) due to the internet/web/social media and electronic storage.

Some would add that the institutionalization of high-level engineering teams, such as those working on IBM’s Watson, have also been a contributing development.

Given these developments, artificial neural networks are now capable of enabling commercially practical operations and products involving:

- image recognition /computer vision (e.g., facial recognition for security and other purposes)
- voice/speech recognition (e.g. for personal home assistants)
- natural language processing, including language generation and translation
- autonomous vehicles
- robotics of many sorts (industrial, service, social)
- web-based financial transactions
- personalized website-user recommendation systems
- a wide range of internal organizational and business management procedures

A succinct generalized characterization of the common core of these recent AI developments might be: *machine learning contributions to pattern recognition*.

To repeat, this newly useful AI is *not* the result of any profound theoretical or engineering breakthroughs. Computational artificial neural networks were proposed over a half-century ago; increased processing speed and Big Data are conceptually *mundane*. Importantly, the newly empowered machine learning AI leads *away from*, not *closer to*, the ways in which human brains, minds and bodies actually solve problems and act in the world. Artificial neural networks were inspired by, but only superficially resemble, human neurons and neural structures; neural net AI is a statistical process of brute force pattern recognition. Amid all the hype, analysts have suggested that the new AI may be a “one-trick pony,” and have referred to the essential mundanity of today’s AI as AI’s “dirty little secret.” See e.g. Somers (2017).

Many AI proponents are declaring that AI going forward will be the most transformative event in human history. But even the more moderate claims that AI will be (simply) hugely consequential are uncertain. Autonomous vehicles could transform the global landscape profoundly, or they could prove incapable of guaranteeing the minimal level of safety we require, and end up limited to narrow, carefully controlled situations, e.g., golf courses and amusement parks. The voice activated home assistants now being marketed might evolve to change the entire meaning and purpose of “home,” or they could prove inherently disruptive (in the bad sense) and go the way of Googleglass.

At the 2017 “Beneficial AI” convening at Asilomar noted AI engineer Yoshua Bengio of the Montreal Institute for Learning Algorithms told the participants that “I think there is lots of hype in AI,” and predicted that future development of AI will be *more* difficult and not, as assumed by most participants, less difficult, because current AI is:

- based largely on *supervised learning* (and is thus computationally constrained)
- still based on superficial training clues, i.e. presence or absence of a distinct discrete feature such as color.
- still a long way from operating at the level of *abstractions*, which is fundamental to human-level intelligence.
- still relying on the lone, old recipe of *back propagation* (see following).

ADDITIONAL COMMON AI MACHINE LEARNING TECHNIQUES [see IBM eMarketeer (2017) for more]

Deep learning. The process through which a multi-level neural network “learns.” It does so by matching successive inputs with an identified model and building patterns of association that will enable it to identify some future input as either conforming to the model or not. Google/DeepMind’s AlphaGo used deep learning to win at Go in 2016.

Reinforcement learning. A type of deep learning in which the computer goes through the process of developing and trying out a series of algorithms instead of following a given algorithm. The results of each trial are presented to a human, who tells the computer if the result is closer to or farther away from the objective it seeks to realize. IBM’s Watson used reinforcement learning to win at Jeopardy in 2011.

Supervised learning: show it pictures with and without a cat in each one, and tagged accordingly; it will eventually be able to look at a new picture and tell if there is or is not a cat in it.

Unsupervised learning: show it millions of pictures with and without a cat in each one, but *don't* include a cat tag. It can still detect patterns and when you show it a new picture will be able to tell if there's a cat in it or not.

Backpropagation. This a process of reinforcement learning in which the act of judging how close a solution is to the desired solution is done by the machine itself, *not* a human judge. It's the powerful core of much current AI research and development.

RESOURCES

The leading introductory AI college text (lower division; presumes linear algebra and simple programming skills) is: Stuart Russell and Peter Norvig. 2011. *Artificial Intelligence: A Modern Approach*; 3rd ed. Essex, UK: Pearson.

Two useful non-technical introductions to AI are:

Jeffrey Kaplan. 2016. *Artificial Intelligence: What Everyone Needs to Know*. Oxford: Oxford University Press.

Kevin Warwick. 2012. *Artificial Intelligence: The Basics*. New York: Routledge.

A handout summary of introductory AI concepts and terms is:

[Artificial Intelligence: What's Now, What's New and What's Next](#) (IBM eMarketeer, 2017).

There are many introductory AI course materials on the web, including:

UC Berkeley [Computer Science 188 Introduction to AI](#) . Spring 2014.

MIT Open Courseware [Artificial Intelligence 6.034](#). Fall 2010.

ATTACHMENT B. CAPSULE INFORMATION ON THE 13 KEY AI ORGANIZATIONS [see also Appendix 1]

A. AI+ organizations advocating a human future deeply and radically transformed by AI:

- 1. Machine Intelligence Research Institute (MIRI; 2000)** – Berkeley. Founder Eliezer Yudkowsky is aligned with the Transhumanists and the “effective altruism” movement. Stated mission: to develop safe and beneficial AI. Team of ~ 15. Budget ~ \$ 2 million.
- 2. Future of Humanity Institute (FHI; 2005)** – Oxford University Martin School – founding Director Nick Bostrom is a key Transhumanist academic and organizer. Focused initially on “existential risk,” and now on building the AI+ movement. Team of 29. Shares office space and collaborates with the Center for Effective Altruism.
- 3. Centre for the Study of Existential Risk (CSER; 2012)** - University of Cambridge. Founded by philosopher Huw Price, cosmologist Martin Rees and funder Jaan Tallinn. Modeled on and collaborates with FHI at Oxford. Exec. Director Seán Ó hÉigeartaigh is former FHI staffer. Team of 32, + 24 Advisors.
- 4. The Future of Life Institute (FLI; 2014)** – Boston. Founded by MIT physicist Max Tegart and funder Jaan Tallinn as a key networking, organizing, base-building and PR hub for the AI advocacy movement. Has organized 6 major strategy conferences, 5 Open Letters, awards of 37 “safe AI” research grants (funded by Elon Musk), and more. No paid staff director. The work is done by Tegart, young volunteers and task-specific paid staff.
- 5. Leverhulme Centre for the Future of Intelligence (LCFI; 2015)** – University of Cambridge. Studies “opportunities and challenges” of AI. Sponsors collaborative efforts with FHI, FLI, CSER and MIRI. Team of 40.
- 6. Center for Human Compatible Artificial Intelligence (CHAI; 2016)** - UC Berkeley. Founded by UCB computer scientist Stuart Russel to refocus AI “towards the ability to generate provably beneficial behavior.” Team of 16. Partners with LCFI, FHI, BERI, MIRI. Holds an annual invitational conference.
- 7. Berkeley Existential Risk Institute (BERI; 2017)** – Berkeley. Research, papers, students committed to identifying and proposing solutions to existential risks. Team of 12. Currently focused on collaborations with CHAI, CSER, FHI, MIRI. Major funding from Open Philanthropy Project.
- 8. OpenAI (2015)** – San Francisco. A non-profit R&D firm whose mission is to develop open-source AGI, “the single most important technology ever developed by humans.” Launched by Elon Musk and Sam Altman with pledges of \$ 1 billion in funding. Team of 60 full-time research staff in place by 2017. Open-source intent is controversial.

B. Mainstream academic/corporate AI associations

- 9. Association for the Advancement of Artificial Intelligence (AAAI; 1979)** – The major, nonprofit, international scientific society devoted to promoting research and applications of AI. It has a journal, an annual conference, etc.
- 10. Partnership on Artificial Intelligence to Benefit People and Society (PAI; 2016)** – NYC. PAI is the presumptive mainstream consensus voice on AI. Founding partners: Amazon, Facebook, Google, IBM, Microsoft & Apple. Total of 49 partners to date. New ED Terah Lyons is formerly of Pres. Obama’s OSTP.

C. Organizations motivated by social responsibility/accountability/ justice values

- 11. Data and Society Research Institute (DSRI; 2014)** - NYC. Non-profit research group focused on social & cultural issues of data-centric and automated technologies. Team of 112, with 18 staff, 26 researchers, etc. Diverse & inclusive. Funding: Sloan, Gates, Ford, Knight, Kellogg, Open Society, etc.
- 12. AI Now Institute (2017)** – NYU. An interdisciplinary research center focused on social implications of AI. Team so far of 5 staff, 45 Advisors. Co-founders/directors Kate Crawford and Meredith Whittaker are NYU faculty and formerly of Microsoft and Google. Funding: MacArthur, Knight, Hewlett, etc.
- 13. UPTURN (2017)** – Washington DC. Activism/advocacy so that “technology serves the dignity and well-being of all people.” 5 staff, 3 board members. E.D. Harlan Yu: Princeton computer sci PhD, formerly at Google in engineering and in policy. Funding: Ford, MacArthur, Open Society.

ATTACHMENT C. CONCERNS RAISED BY KEY SETS OF THE 13 AI ORGANIZATIONS

C.1. CONCERNS RAISED BY THE 8 AI+ SUPERINTELLIGENCE ADVOCACY ORGANIZATIONS

1. The possibility of a Superintelligence explosion that goes wrong and destroys humanity.

* **Proposed solution:** Develop the algorithms necessary to ensure “human values alignment” and install these in all AIs, so that if and when the Superintelligence explosion happens, the results (whatever they might be) will be in alignment with human values.

Note: Much of the new research funded by the Future of Life Institute is directed at this “values alignment” problem. The topic was discussed at length at the 2017 Asilomar event. Some wondered how an AI could be programmed to optimize a single human values utility function, as different people hold different values. Ray Kurzweil argued that this would not be difficult, since most people hold the same *ultimate* values: they want to be healthy, financially successful, etc. Others were still unsure, and urged that experts in the social sciences be consulted.

2. The firm that first develops Superintelligence will be able to achieve global domination within a matter of hours, in not minutes. There is thus an incentive for firms to *race to the finish*, and in doing so to cut corners on AI safety, thus increasing the risks of a catastrophically bad Superintelligence explosion.

* **Proposed solution:** The major firms working on AI should agree that regardless of who gets there first, product lines and other commercial opportunities will be divided up in some manner that is felt to be fair (although not exactly equal). In addition, there should be agreement that certain results of the research conducted by all AI firms will be put in the public domain (although some will also remain proprietary). With such agreements all AI firms will still have an incentive to move quickly, although not so quickly that safety and the public good is endangered.

3. Religious, ultra-Luddite, political, lone wolf or other terrorists develop an AI capable of destroying humanity.

* **Proposed solution:** Ensure that AI R&D skills have been disseminated as widely as possible, ideally universally, so that in the event of a terrorist attack counter-AIs could be devised and deployed within minutes or even seconds.

Note: This is the motivation given by Elon Musk and Sam Altman for establishing OpenAI, their \$1 billion effort to build an AGI and put its source code in the public domain. It’s identical to the motivation given in 2009 by synthetic biologist Drew Endy of MIT for encouraging wide dissemination of DIY synbio kits.

4. Mass, permanent technological unemployment and its social/political repercussions (“pitchforks”).

* **Proposed solutions (partial list):**

- Some form of the *Universal Basic Income*.
- A *Conditional Basic income*, i.e., you only get \$\$\$ if you are in a skills re-training program.
- *Networked market solutions*: use AI and other technologies to empower individuals. Example:
 - * Airbnb now lets people advertise themselves as *tour guides* for unique, personalized tours for those visiting their cities, e.g, a “Motown music history” tour of Detroit, or an architectural tour of select San Francisco neighborhoods.

5. AI will *not develop rapidly enough*, allowing existential dangers to happen and humanity to go extinct.

* **Proposed solution:** work hard to make sure that the four risk noted above, and similar negative possibilities, do not come to pass, while continuing to educate key constituencies and the general public concerning the future of abundance that AI holds for humanity.

For further discussion see e.g. Bostrom (2014), Tegmark (2016), Hason (2016), Yampolskiy (2016), Chase (2016, 2015).

C.2a. MAINSTREAM CONCERNS IDENTIFIED BY THE AAAI (2009) AND BY DIETTERICH & HORVITZ (2015)

A. AAAI President Eric Horvitz (2008-2009) organized the [Presidential Panel on Long-Term AI Futures](#), which prepared reports and held a workshop at Asilomar in February 2009. Key findings included:

1. Neither the utopian vision of Kurzweil nor the dystopian vision of “Robot Takeover” are credible AI scenarios. The likelihood of an intelligence explosion or large-scale loss of control of AI systems is vanishingly small. Academics, journalists and others who promote fear of such dangers are distracting society from real concerns that need to be addressed now.

2. These real, shorter-term and actionable concerns include:

- * the need to ensure that AI does not compromise people’s privacy
- * the need to enhance mechanisms that allow people and machine intelligences to work collaboratively.
- * the need to make the processes of machine learning more transparent, both to lay and expert personnel (eg, give the machine the ability to explain what it is doing and why).
- * There are many opportunities, and will be more, for malicious use of AI. This doesn’t pose “existential risks” but could still be extremely dangerous. We need to be proactive about this. Example: research on user modeling and intrusion detection directed at potential threats, in advance of evidence of criminal efforts.

3. We need research on a) ethical and legal responsibility for action recommended by, or taken directly by, AI systems; and b) ethical and psychological issues involved when AI systems are built that have believable affect, feelings and personality, which do not exist in reality.

B. By the mid-2010s the AAAI had been eclipsed in the public debate by the A+ Superintelligence Advocates, and especially so after Musk, Hawking and others with money and/or celebrity amplified the A+ message. Dietterich and Horvitz (2015) published an opinion piece that laid out the mainstream reaction and positions.

It said that “Superintelligence” was not a credible threat, but that the possibility of large-scale harmful developments could not be ruled out, and that some sort of monitoring was appropriate. Of higher priority were more immediate risks, including:

1. BUGS - These risks stem from software errors, for both AI and non-AI systems. Such errors can cause significant economic loss, and even fatalities. Especially with advanced systems, e.g. autonomous vehicles and robot surgery, very high levels of quality control are necessary. This may require new self-monitoring architectures

2. CYBERSECURITY – We need to consider new attack surfaces that AI exposes, especially when charged with high-stakes decisions. And AI itself can contribute to new cybersecurity systems. For example, machine learning can learn to recognize the fingerprints of malware.

3. THE SORCERER’S APPRENTISE – If we tell an autonomous car to “get us to the airport ASAP”, will it run other cars off the road on the way? This and similar scenarios result from human failure to properly specify instructions. AIs will also need to know what to do when faced with novel situations, e.g. when to ask for further instructions.

4. SHARED AUTONOMY - Systems using human/machine collaboration can be highly efficient and productive, but are difficult to construct. Aircraft under 100% human control or 100% AI control can work fine, but joint control can be highly risky, because pilot and AI have to monitor each other, in addition to the craft and the environment.

5. SOCIOECONOMIC IMPACTS - AI systems could very well contribute to growing economic inequality. We need a much better understanding of how and exactly why this happens, and what can be done to mitigate it or compensate for it.

C.2b. GOALS, WORK AND TENETS OF THE PARTNERSHIP ON ARTIFICIAL INTELLIGENCE (PAI)

[source: summarized from [PAI website](#)]

A. GOALS of the Partnership on AI

1. Develop and Share Best Practices: in areas of research, development, testing, fielding, fairness & inclusivity, explanation & transparency, security & privacy, values & ethics, collaboration, interoperability, trustworthiness, reliability, containment, safety and robustness.
2. Provide an Open and Inclusive Platform for Discussion and Engagement: between AI researchers and stakeholders in law, policy, government, civil liberties, and the greater public.
3. Advance Public Understanding: by educating the public, answering questions, acting as a trusted contact.
4. Identify and Foster Aspirational Efforts in AI for Socially Beneficial Purposes: including promising technologies and applications not being explored by academia.

B. THE WORK of the Partnership on AI

1. Engagement of Experts: from e.g. psychology, philosophy, economics, finance, sociology, public policy, and law to discuss and provide guidance on emerging issues related to the impact of AI on society.
2. Engagement of Other Stakeholders, including AI users, developers, representatives of industry (healthcare, financial services, transportation, commerce, manufacturing, telecomms, media, etc.) to support best practices.
3. Support of objective third-party studies on best practices for ethics, safety, fairness, inclusiveness, trust and robustness. Support of aspirational projects in AI that would greatly benefit people and society.
4. Development of informational materials on the current and future likely trajectories of research and development in core AI and related disciplines.

C. TENETS of the Partnership on AI

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:
 - a. Working to protect the privacy and security of individuals.
 - b. Striving to understand and respect the interests of all parties that may be impacted by AI advances.
 - c. 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.
 - d. Ensuring that AI research and technology is robust, reliable, trustworthy, and operates within secure constraints.
 - e. 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.

C.3. CONCERNS RAISED BY THE PROGRESSIVE / SOCIAL RESPONSIBILITY AI ORGANIZATIONS

[sources: summarized from the websites of [Data & Society](#), [AI Now](#) and [Upturn](#)]

- 1. Safety & Justice:** Ensure that AI technology in the criminal justice system supports civil rights.
- 2. Markets & Opportunity:** Ensure that AI technology expands, rather than diminishes, opportunities for consumers and workers in the digital age. Prevent predatory new marketplace practices. Promote inclusion.
- 3. Open & Secure Communication:** Ensure a free, open, and secure Internet. Develop and deploy anti-censorship software. Support policies ensuring privacy, digital security, and content moderation.
- 4. Decisions, Automation and Power:** Ensure that when people's lives are shaped by high tech predictions and automated decisions, the results are fair and the process is accountable.
- 5. Fairness in Precision Medicine:** Critically assess the potential for bias and discrimination in data-driven health care, which relies on the collection, sharing, and interpretation of medical records, genetic information and more.
- 6. Libraries and Privacy:** Improve the ability of local libraries to support their communities with regard to data-centric technological development, by addressing privacy in libraries, facilitating safe research data sharing worldwide, and new roles for librarians as data scientists.
- 7. Law & Ethics in Computational Social Science:** Ensure non-discrimination, due process and understandability in decision-making.
- 8. Enabling Connected Learning:** As young people embrace technology to learn, play, and socialize, the boundaries between education, the home, and society get increasingly blurred. We need to support the Cities of Learning movement to maximize learning experience and to decide where and when student data can and should be used.
- 9. Data, Human Rights and Human Security:** Cell phones, drones, Big Data, AI networks and many more systems can both greatly help and seriously hurt large numbers of people. How can these technologies be used to assist people in need, prevent abuse, and protect from harm? How should data analytics be used to make decisions in human rights and security domains? Should values like privacy fluctuate in the face of immediate threats? How will data collection and monitoring shape the practices of those that repress or advocate for human rights? How can data regulation, policy, and standards effectively govern and provide accountability? How must the human rights and human security fields interact with data tools and techniques that can identify specific individuals at risk? Will private sector data sharing (or data philanthropy) become a positive and sustainable movement?
- 10. Media manipulation:** Provide news organizations, civil society, platforms and policymakers with insights into new forms of media manipulation, to ensure a close and informed relationship between technical research and socio-political outcomes.
- 11. Intelligence and autonomy:** We need to connect the dots between robots, algorithms and automation, and reframe the debates around the rise of machine intelligence.
- 12. Future of labor:** Data-centric technology is disrupting, destabilizing, and transforming the labor force, fostering structural inequalities, employment discrimination, workplace surveillance and more. How do we protect laborers from abuse, poor work conditions, and discrimination, in a networked age in which unions are no longer suited?
- 13. Rights and Liberties:** AI and related technologies are used to make determinations and predictions in high stakes domains such as criminal justice, law enforcement, housing, hiring and education. They have the potential to impact basic rights and liberties in profound ways. How can we ensure that these impacts help rather than hurt?

ATTACHMENT D. CURRENT DEVELOPMENT OF AI BY THE MAJOR MAINSTREAM I-TECH FIRMS

IBM built the AI supercomputer Watson, which uses natural language processing, machine learning, and pattern recognition (images, videos, text, speech). In 2011 it used Watson to beat a human at 'Jeopardy!' It used the resulting publicity to promote its Watson-based AI products world-wide. It now has over 100 commercialized AI products and services and is working with clients in 45 countries across 20 different industries.

Google acquired DeepMind in 2014 as part of a major expansion of its AI R&D. Google/DeepMind garnered much favorable publicity in 2016 when its AlphaGo machine beat the world champion Go player. Google has been incorporating machine learning into its core search operations, autonomous vehicles, the Google Now virtual assistance, a chat app (Allo), and TensorFlow, an open-source machine learning software library.

Apple continues to develop its iPhone-based Siri personal assistant. It is integrating IA-driven features, including facial and image recognition, content recommendations, word choice prediction and usage pattern recognition into its full product and service line. It has opened Siri and other AI-driven apps to third-party developers.

Facebook is using AI to recognize and tag faces in photographs, curate News Feeds and manage ad placements. It has developed a text understanding engine, DeepText, to help extract intent and sentiment from posts, and help remove objectionable content. Facebook is testing an intelligent agent, M, as part of Facebook Messenger, which communicates via text and helps with customer service features such as "booking travel and ordering flowers."

Microsoft developed Skype Translator to allow translation in real time, and is now applying it to many of its operations, including search. It recently purchased Swiftkey, which analyzes users' typing history and learns usage patterns. It also offers its digital personal assistant Cortana; Microsoft Cognitive Services, which consults with firms on use of AI; and its Azure cloud platform which helps developers create bots using text, type and video.

Amazon's AI developers are working on models that predict how likely customers are to click on products shown in search queries, and to click on related links and make purchases. Amazon Machine Learning Service provides cloud products. Amazon's deep learning software, DSSTNE, is available to developers. The Amazon Echo home speaker device includes virtual assistant Alexa, which can answer questions, read books/articles, play music, control home devices, and order products.

Salesforce has acquired MetaMind, which develops natural language processing, computer vision and predictive analytics tools. In 2016 Salesforce introduced Salesforce Einstein, a platform that integrates AI into a company's sales, service and marketing platform and enables customers to build AI-powered apps.

Alibaba joined IBM, Microsoft and Amazon in the 2015 launch of Aliyun, a turnkey cloud-based machine learning service that can predict user behavior. In 2016 Alibaba released a suite of AI solutions to increase efficiency, lower costs and monitor risks.

Baidu is integrating AI features, including image recognition, video analysis and augmented reality, into its core search and web offerings. It is using AI to find malware on its network and to target ads, and is developing AI systems for autonomous vehicles.

Samsung made a major strategic investment in 2015 in Vicarious, a machine learning startup focused on integrating AI software and hardware.

[Comment: Compared with the grandiose rhetoric over the past ~ 8 yrs by the AI community and the media, the real, existing AI R&D and product, as noted here, are limited to a small number of engineering approaches (e.g. backpropagation) and products. The potential is real but the magnitude of the eventual impact is uncertain.]

Sources: eMarketeer (2017), firm websites and other websites.

ATTACHMENT E. OP-ED: TRANSCENDING COMPLACENCY ON SUPERINTELLIGENT MACHINES

By Stephen Hawking, Max Tegmark, Stuart Russell, and Frank Wilczek

Huffington Post blog post - 19 April 2014

As the Hollywood blockbuster *Transcendence* debuts this weekend with Johnny Depp, Morgan Freeman and clashing visions for the future of humanity, it's tempting to dismiss the notion of highly intelligent machines as mere science fiction. But this would be a mistake, and potentially our worst mistake ever.

Artificial intelligence (AI) research is now progressing rapidly. Recent landmarks such as self-driving cars, a computer winning at *Jeopardy!*, and the digital personal assistants Siri, Google Now and Cortana are merely symptoms of an IT arms race fueled by unprecedented investments and building on an increasingly mature theoretical foundation. Such achievements will probably pale against what the coming decades will bring.

The potential benefits are huge; everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of war, disease, and poverty would be high on anyone's list. Success in creating AI would be the biggest event in human history.

Unfortunately, it might also be the last, unless we learn how to avoid the risks. In the near term, for example, world militaries are considering autonomous weapon systems that can choose and eliminate their own targets; the UN and Human Rights Watch have advocated a treaty banning such weapons. In the medium term, as emphasized by Erik Brynjolfsson and Andrew McAfee in *The Second Machine Age*, AI may transform our economy to bring both great wealth and great dislocation.

Looking further ahead, there are no fundamental limits to what can be achieved: there is no physical law precluding particles from being organized in ways that perform even more advanced computations than the arrangements of particles in human brains. An explosive transition is possible, although it may play out differently than in the movie: as Irving Good realized in 1965, machines with superhuman intelligence could repeatedly improve their design even further, triggering what Vernor Vinge called a "singularity" and Johnny Depp's movie character calls "transcendence." One can imagine such technology outsmarting financial markets, out-inventing human researchers, out-manipulating human leaders, and developing weapons we cannot even understand. Whereas the short-term impact of AI depends on who controls it, the long-term impact depends on whether it can be controlled at all.

So, facing possible futures of incalculable benefits and risks, the experts are surely doing everything possible to ensure the best outcome, right? Wrong. If a superior alien civilization sent us a text message saying, "We'll arrive in a few decades," would we just reply, "OK, call us when you get here — we'll leave the lights on"? Probably not — but this is more or less what is happening with AI. Although we are facing potentially the best or worst thing ever to happen to humanity, little serious research is devoted to these issues outside small non-profit institutes such as the [Cambridge Center for Existential Risk](#), the [Future of Humanity Institute](#), the [Machine Intelligence Research Institute](#), and the [Future of Life Institute](#). All of us — not only scientists, industrialists and generals — should ask ourselves what can we do now to improve the chances of reaping the benefits and avoiding the risks.

Stephen Hawking is Director of Research at the Centre for Theoretical Physics at Cambridge and a 2012 Fundamental Physics Prize laureate for his work on quantum gravity. **Stuart Russell** is a computer science professor at Berkeley and co-author of "Artificial Intelligence: a Modern Approach." **Max Tegmark** is a physics professor at M.I.T. and the author of "Our Mathematical Universe." **Frank Wilczek** is a physics professor at M.I.T. and a 2004 Nobel laureate for his work on the strong nuclear force.

ATTACHMENT F. AGENDA OF THE INVITATIONAL “BENEFICIAL AI” CONVENING – San Juan, Puerto Rico, Jan 2-5, 2015

Primary organizing role was played by the [Future of Life Institute](#). The PDFs for all the presentations are available on the FLI website.

Friday January 2:

1600-late: Registration

1930-2130: Welcome reception (Las Olas Terrace)

Saturday January 3:

0800-0900: Breakfast

0900-1200: Overview (one review talk on each of the four conference themes)

- Welcome – Max Tegmark, Future of Life Institute
- Ryan Calo (Univ. Washington): AI and the law
- Erik Brynjolfsson (MIT): AI and economics (pdf)
- Richard Sutton (Alberta): Creating human-level AI: how and when? (pdf)
- Stuart Russell (Berkeley): The long-term future of (artificial) intelligence (pdf)

1200-1300: Lunch; 1300-1515: Free play/breakout sessions on the beach; 1515-1545: Coffee & snacks

1545-1600: Breakout session reports

(A typical 3-hour session consists of a few 20-minute talks followed by a discussion panel where the panelists who haven't already given talks get to give brief introductory remarks before the general discussion ensues.)

1600-1900: Optimizing the economic impact of AI. What can we do now to maximize the chances of reaping the economic bounty from AI while minimizing unwanted side-effects on the labor market?

Speakers:

- Andrew McAfee, MIT (pdf)
- James Manyika, McKinsey (pdf)
- Michael Osborne, Oxford (pdf)

Panelists: Ajay Agrawal (Toronto), Erik Brynjolfsson (MIT), Robin Hanson (GMU), Scott Phoenix (Vicarious)

1900: dinner

Sunday January 4:

0800-0900: Breakfast

0900-1200: Creating human-level AI: how and when? Will it happen, and if so, when? Via engineered solution, whole brain emulation, or other means? (We defer until the 4pm session questions regarding what will happen, about whether machines will have goals, about ethics, etc.)

Speakers:

- Demis Hassabis, Google/DeepMind
- Dileep George, Vicarious (pdf)
- Tom Mitchell, CMU (pdf)

Panelists: Joscha Bach (MIT), Francesca Rossi (Padova), Richard Mallah (Cambridge Semantics), Richard Sutton (Alberta)

[cont.]

[cont.]

1200-1300: Lunch; 1300-1515: Free play/breakout sessions on the beach; 1515-1545: Coffee & snacks

1545-1600: Breakout session reports

1600-1900: Intelligence explosion: science or fiction? If an intelligence explosion happens, then what are likely outcomes? What can we do now to maximize the probability of a positive outcome? Containment problem? Is “friendly AI” possible? Feasible? Likely to happen?

Speakers:

- Nick Bostrom, Oxford (pdf)
- Bart Selman, Cornell (pdf)
- Jaan Tallinn, Skype founder (pdf)
- Elon Musk, SpaceX, Tesla Motors

Panelists: Shane Legg (Google/DeepMind), Murray Shanahan (Imperial), Vernor Vinge (San Diego), Eliezer Yudkowsky (MIRI)

1930: banquet (outside by beach)

Monday January 5:

0800-0900: Breakfast

0900-1200: Law & ethics: Improving the legal framework for autonomous systems. How should legislation be improved to best protect the AI industry and consumers? If self-driving cars cut the 32000 annual US traffic fatalities in half, the car makers won’t get 16000 thank-you notes, but 16000 lawsuits. How can we ensure that autonomous systems do what we want? And who should be held liable if things go wrong? How tackle criminal AI? AI ethics? AI ethics/legal framework for military systems & financial systems?

Speakers:

- Joshua Greene, Harvard (pdf)
- Heather Roff Perkins, Univ. Denver (pdf)
- David Vladeck, Georgetown

Panelists: Ryan Calo (Univ. Washington), Tom Dietterich (Oregon State, AAAI president), Kent Walker (General Counsel, Google)

1200: Lunch, depart

- - - -

ATTACHMENT G. An Open Letter on Maximizing the Societal Benefits of AI – January 11, 2015

[This *Open Letter* was prepared and released following the January 2015 meeting in Puerto Rico. It was signed by dozens of AI researchers and advocates, most of them presumably being among the January 2-5 meeting participants. By January 2018 the *Open Letter* had received upwards of 8,000 signers on-line.]

Research Priorities for Robust and Beneficial Artificial Intelligence

Artificial intelligence (AI) research has explored a variety of problems and approaches since its inception, but for the last 20 years or so has been focused on the problems surrounding the construction of intelligent agents – systems that perceive and act in some environment. In this context, “intelligence” is related to statistical and economic notions of rationality – colloquially, the ability to make good decisions, plans, or inferences. The adoption of probabilistic and decision-theoretic representations and statistical learning methods has led to a large degree of integration and cross-fertilization among AI, machine learning, statistics, control theory, neuroscience, and other fields. The establishment of shared theoretical frameworks, combined with the availability of data and processing power, has yielded remarkable successes in various component tasks such as speech recognition, image classification, autonomous vehicles, machine translation, legged locomotion, and question-answering systems.

As capabilities in these areas and others cross the threshold from laboratory research to economically valuable technologies, a virtuous cycle takes hold whereby even small improvements in performance are worth large sums of money, prompting greater investments in research. There is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase. The potential benefits are huge, since everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of disease and poverty are not unfathomable. Because of the great potential of AI, it is important to research how to reap its benefits while avoiding potential pitfalls.

The progress in AI research makes it timely to focus research not only on making AI more capable, but also on maximizing the societal benefit of AI. Such considerations motivated the AAAI 2008-09 Presidential Panel on Long-Term AI Futures and other projects on AI impacts, and constitute a significant expansion of the field of AI itself, which up to now has focused largely on techniques that are neutral with respect to purpose. We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial: our AI systems must do what we want them to do. The attached research priorities document gives many examples of such research directions that can help maximize the societal benefit of AI. This research is by necessity interdisciplinary, because it involves both society and AI. It ranges from economics, law and philosophy to computer security, formal methods and, of course, various branches of AI itself.

In summary, we believe that research on how to make AI systems robust and beneficial is both important and timely, and that there are concrete research directions that can be pursued today.

Comment: *As open letters go this one is remarkably understated. However, it accomplished its major purposes of affirming both the huge promise of AI (“...the eradication of disease and poverty...”), and of presenting the leadership of the AI research and advocacy community as trustworthy societal stewards of this technology and its promise. [This communication](#) from Popular Science writer Eric Sodge includes the suggestion that the subdued tone of the letter was in reaction to the extended frenzied tone taken by much of the press to AI/robotics topics at this time. The FLI organizers had exaggerated the nature of the threat of AI, and the press took that and exaggerated it even further, beyond a level with which even the FLI organizers were comfortable. - RH*

ATTACHMENT H. PARTNERSHIP ON ARTIFICIAL INTELLIGENCE TO BENEFIT PEOPLE AND SOCIETY (PAI) – Roster of Partners

The PAI was established in 2016. As of March 2018 it had 51 Partners and Founding Partners. Co-Chairs are **Eric Horovitz** (Microsoft) and **Mustafa Suleyman** (DeepMind). Executive Director is **Terah Lyons**, former Policy Advisor to the U.S. Chief Technology Officer at the White House Office of Science and Technology Policy (OSTP), under President Barak Obama. See the [PAI website](#) for further partnership information.

FOUNDING PARTNERS

APPLE

DEEPMIND (acquired by Google, 2017)

FACEBOOK

GOOGLE

IBM

MICROSOFT

PARTNERS (as of March 2018)

ASSOCIATION FOR THE ADVANCEMENT OF AI

ACCENTURE

ACLU

AFFECTIVA

AI FORUM NEW ZEALAND

AI NOW INSTITUTE

THE ALLEN INSTITUTE FOR ARTIFICIAL INTELLIGENCE

AMNESTY INTERNATIONAL

ARTICLE 19

ASSOCIATION FOR COMPUTING MACHINERY

CENTER FOR DEMOCRACY & TECHNOLOGY (CDT)

CENTER FOR HUMAN-COMPATIBLE AI

CENTER FOR INFORMATION TECHNOLOGY POLICY

CENTRE FOR INTERNET AND SOCIETY, INDIA (CIS)

LEVERHULME CENTRE FOR THE FUTURE OF
INTELLIGENCE

COGITAI

DATA & SOCIETY RESEARCH INSTITUTE

DIGITAL ASIA HUB

DOEVERYONE

eBAY

ELEMENT AI

ELECTRONIC FRONTIER FOUNDATION (EFF)

FRAUNHOFER IAOTHE

FUTURE OF HUMANITY INSTITUTE

FUTURE OF LIFE INSTITUTE

THE FUTURE OF PRIVACY FORUM

THE HASTINGS CENTER

HONG KONG UNIVERSITY OF SCIENCE AND

TECHNOLOGY DEPARTMENT OF ELECTRONIC &
COMPUTER ENGINEERING

HUMAN RIGHTS WATCH

INTEL

MARKKULA CENTER FOR APPLIED ETHICS -
SANTA CLARA UNIVERSITY

MCKINSEY & COMPANY

NVIDIA

OMIDYAR NETWORK

OPENAI

OXFORD INTERNET INSTITUTE

SALESFORCE

SAP

SONY

TUFTS UNIVERSITY HRI LAB

UCL ENGINEERING

UNICEF

UNIVERSITY OF WASHINGTON TECH POLICY LAB

UPTURN

XPRIZE

ZALAN

[Comments in progress. The diversity of the partner organizations is novel.]

ATTACHMENT I. INFLUENTIAL MAINSTREAM POLICY-ORIENTED AI INITIATIVES: STANFORD (2016), WHITE HOUSE (2016), and AI NOW (2017)

The heightened attention to AI that began ~ 2013 was sparked by the AI+ advocates and in the immediate run capitalized upon most successfully by them. The mainstream academic and political community had to respond and did so, with conferences and reports. Highlights of three of the most influential of these are summarized here.

I. Stanford University: [*The One Hundred Year Study on Artificial Intelligence*](#) (September 2016)

AI100 was initiated by Microsoft AI developer and Stanford alumnus **Eric Horvitz**. The plan is to prepare a report on important developments in AI at five-year intervals for at least the next 100 years. AI100 is funded by an endowment to Stanford University from Horvitz and his wife. It is overseen by a 7-member standing committee; each 5-year report is to be prepared by a Study Panel of ~ 12-15 members selected by the standing committee.

The first report [see Stone et al (2016)] begins with a brief summary of the state of AI research. It continues by reviewing current and expected near-term AI developments regarding healthcare, education, transportation, “low-resource communities,” home/service robots, employment/workplace, and entertainment.

It concludes with three policy recommendations:

- 1) All levels of government should “define a path towards accruing technical expertise in AI;”
- 2) Remove impediments to research “on the fairness, security, privacy and social impacts of AI systems.”
- 3) “Increase public and private funding for interdisciplinary studies of the societal impacts of AI.”

II. White House: [*Preparing for the Future of Artificial Intelligence*](#) (October 2016)

This was one of the last reports issued by the White House Office of Science and Technology Policy (OSTP) before the end of the Obama administration. It was prepared by the National Science and Technology Council Subcommittee on Machine Learning and Artificial Intelligence, with representation from ~ 30 federal agencies. Five public workshops, a symposium and other outreach activities helped inform the report.

The report surveys a wide range of AI topics, including “Applications for the Public Good,” the state of regulation, economic impacts, current research plans and needs, global and military considerations, “Fairness, safety and governance,” and the role of the Federal Government in all of these.

The tone of the report is one of heavy boosterism. Although it acknowledges many questions that need to be addressed regarding, eg, impact on jobs, inequality and privacy, its strongest language is used to hail the manifold positive transformations that AI will bring to society and the world. It cautions against regulation that might discourage innovation. And it affirms the need for federal funding to support AI research and development.

The report strongly rejects the “dystopian” view that “super-intelligent machines would exceed the ability of humanity to understand or control them.” It offers “A more positive view... held by many researchers... [in which AI works] ... as helpers, assistants, trainers, and teammates of humans, and are designed to operate safely and ethically.” It concludes, “The NSTC Committee on Technology’s assessment is that long-term concerns about super-intelligent General AI should have little impact on current policy.”

III. [*AI NOW 2017 Report*](#) (2017)

AI Now is a spin-off of the White House AI initiative; its founders, **Meredith Whittaker** and **Kate Crawford** of NYU, had organized the July 2016 White House symposium that helped inform the October 2016 White House report. AI Now is now an independent institute affiliated with NYU and held its own symposium and workshop in 2017. The resulting report focused on near-term social and economic implications of AI in four priority areas:

1. Labor and Automation: Attention is rightly being given to the possibility of future mass unemployment, but AI is having a negative impact right now on the balance of workplace power. The gig economy inherently leaves workers at a disadvantage; powerful workplace surveillance technology is being introduced; and AI-assisted management is blurring accountability. These and other problematic developments call for more attention now.

2. Bias and Inclusion: AI can potentially reduce subjective bias in e.g., hiring, judicial or medical decisions, but if the AI is “trained” using biased data it will reproduce and even exacerbate those biases. Those designing AI systems tend to be male, highly educated and very well paid, and tend to use data with which they are familiar and comfortable. More diversity is needed at all levels of AI if it is to live up to its potential.

3. Rights and Liberties: Governments and private firms are using AI in problematic ways, eg, by using police body camera footage to train machine vision algorithms for law enforcement. And privacy rights are being compromised by the need for huge data banks for AI training, eg, medical record data banks to train diagnostic AI.

4. Ethics and Governance: Many AR firms and researchers are now promulgating ethical codes and principles, but these are still inadequate. Many are inconsistent with one another. Enforcement and accountability is lacking. They often focus on individual behavior and let big institutions off the hook. The military development of lethal autonomous weapons raises additional challenging questions.

ATTACHMENT J. “[BENEFICIAL AI 2017](#)” Conference, Asilomar, CA, January 2-8, 2017

Primary organizing role was played by the Future of Life Institute.

Workshops were held Mon-Wed Jan 2-4. The sessions Thurs-Sun Jan 5-8 were all full plenary sessions.

Thursday January 5

All afternoon: registration open; come chill & meet old and new friends!

1800-2100: Welcome reception

Friday January 6

0730-0900: Breakfast

0900-1200: Opening keynotes on AI, economics & law: Progress since Puerto Rico 2015

- Welcome Remarks by Max Tegmark ([video](#))
- Talks:
 - Erik Brynjolfsson (MIT) ([pdf](#)) ([video](#))
 - Yoshua Bengio (Montreal) ([pdf](#)) ([video](#))
 - Viktoriya Krakovna (DeepMind/FLI) ([pdf](#)) ([video](#))
 - Stuart Russell (Berkeley) ([pdf](#)) ([video](#))
 - Ryan Calo (U. Washington) ([video](#))
- Group photo
- 1200-1300: Lunch;
- 1300-1500: Breakout sessions

1500-1800: Economics: How can we grow our prosperity through automation without leaving people lacking income or purpose?

- Talk: Daniela Rus (MIT) ([video](#))
- Panel with Daniela Rus (MIT), Andrew Ng (Baidu), Mustafa Suleyman (DeepMind), Moshe Vardi (Rice) & Peter Norvig (Google): How AI is automating and augmenting work ([video](#))
- Talks:
 - Andrew McAfee (MIT) ([pdf](#)) ([video](#))
 - Jeffrey Sachs (Columbia) ([pdf](#)) ([video](#))
- Panel with Andrew McAfee (MIT), Jeffrey Sachs (Columbia), Eric Schmidt (Google) & Reid Hoffman (LinkedIn): *Implications of AI for the Economy and Society* ([video](#))
- Fireside chat with Daniel Kahneman: *What makes people happy?* ([video](#))

1800-2100: Dinner

Saturday January 7

0730-0900: Breakfast

0900-1200: Creating human-level AI: Will it happen, and if so, when and how? What key remaining obstacles can be identified? How can we make future AI systems more robust than today's, so that they do what we want without crashing, malfunctioning or getting hacked?

- Talks:
 - Demis Hassabis (DeepMind) ([video](#))

- Ray Kurzweil (Google) ([video](#))
- Yann LeCun (Facebook/NYU) ([pdf](#)) ([video](#))
- Panel with Anca Dragan (Berkeley), Demis Hassabis (DeepMind), Guru Banavar (IBM), Oren Etzioni (Allen Institute), Tom Gruber (Apple), Jürgen Schmidhuber (Swiss AI Lab), Yann LeCun (Facebook/NYU), Yoshua Bengio (Montreal) ([video](#))

1200-1300: Lunch

1300-1500: Breakout sessions

1500-1800: Superintelligence: Science or fiction? If human level general AI is developed, then what are likely outcomes? What can we do now to maximize the probability of a positive outcome? ([video](#))

- Talks:
 - Shane Legg (DeepMind)
 - Nick Bostrom (Oxford) ([pdf](#)) ([video](#))
 - Jaan Tallinn (CSER/FLI) ([pdf](#)) ([video](#))
- Panel with Bart Selman (Cornell), David Chalmers (NYU), Elon Musk (Tesla, SpaceX), Jaan Tallinn (CSER/FLI), Nick Bostrom (FHI), Ray Kurzweil (Google), Stuart Russell (Berkeley), Sam Harris, Demis Hassabis (DeepMind): **If we succeed in building human-level AGI, then what are likely outcomes? What would we like to happen?**
- Panel with Dario Amodei (OpenAI), Nate Soares (MIRI), Shane Legg (DeepMind), Richard Mallah (FLI), Stefano Ermon (Stanford), Viktoriya Krakovna (DeepMind/FLI): **Technical research agenda: What can we do now to maximize the chances of a good outcome?** ([video](#))

1800-2200: Banquet

SUNDAY JANUARY 8

0730-0900: Breakfast

0900-1200: Law, policy & ethics: How can we update legal systems, international treaties and algorithms to be more fair, ethical and efficient and to keep pace with AI?

- Talks:
 - Matt Scherer ([pdf](#)) ([video](#))
 - Heather Roff-Perkins (Oxford)
- Panel with Martin Rees (CSER/Cambridge), Heather Roff-Perkins, Jason Matheny (IARPA), Steve Goose (HRW), Irakli Beridze (UNICRI), Rao Kambhampati (AAAI, ASU), Anthony Romero (ACLU): **Policy & Governance** ([video](#))
- Panel with Kate Crawford (Microsoft/MIT), Matt Scherer, Ryan Calo (U. Washington), Kent Walker (Google), Sam Altman (OpenAI): **AI & Law** ([video](#))
- Panel with Kay Firth-Butterfield (IEEE, Austin-AI), Wendell Wallach (Yale), Francesca Rossi (IBM/Padova), Huw Price (Cambridge, CFI), Margaret Boden (Sussex): **AI & Ethics** ([video](#))

1200-1300: Lunch

1300: Depart

ATTACHMENT K. Participants at the "Beneficial AI 2017" Conference - Asilomar, CA. Jan 2-8, 2017				
<i>"X" = affiliated with an AI+ or "effective altruism" organization</i>				
1	Anthony	Aguirre	UC Santa Cruz; Future of Life Institute	X
2	Sam	Altman	Y Combinator	X
3	Dario	Amodei	OpenAI	X
4	Amara	Angelica	office of Ray Kurzweil	
5	Stuart	Armstrong	Future of Humanity Institute	
6	Peter	Asaro	The New School	
7	Kareem	Ayoub	DeepMind	
8	Guru	Banavar	IBM	
9	Yoshua	Bengio	University of Montreal	
10	Nicolas	Berggruen	Berggruen Institute	
11	Irakli	Beridze	United Nations	
12	Margaret	Boden	University of Sussex	
13	Gregory	Bonnet	University of Caen Normandy	
14	Nick	Bostrom	Future of Humanity Institute	X
15	Erik	Brynjolfsson	MIT Institute on the Digital Economy	
16	Vitalik	Buterin	Ethereum	
17	Craig	Calhoun	Berggruen Institute	
18	Ryan	Calo	University of Washington Law	
19	Stephen	Cave	Leverhulme Centre for the Future of Intelligence	X
20	David	Chalmers	NYU Philosophy; Australian National University	
21	Nancy	Chang	Google	
22	Meia	Chita-Tegmark	Future of Life Institute	X
23	Paul	Christiano	UC Berkeley; Future of Humanity Institute	X
24	Jack	Clark	OpenAI	X
25	Vincent	Conizer	Duke University; Computer Science, Econ, Philosophy	
26	Ariel	Conn	Future of Life Institute	X
27	Owen	Cotton-Barratt	Future of Humanity Institute	X
28	Kate	Crawford	Microsoft Research	
29	Andrew	Critch	Machine Intelligence Research Institute	X
30	Allan	Dafoe	Yale Political Science; Future of Humanity Institute	X
31	Tucker	Davey	Future of Life Institute	X
32	Abram	Demski	USC PhD student - computer sci	
33	Daniel	Dewey	Open Philanthropy Project	X
34	Thomas	Dieterich	Oregon State U - computer science	
35	Anca	Dragan	UC Berkeley - Electrical Engineering	
36	Eric	Drexler	Future of Humanity Institute	X
37	Stefano	Ermon	Stanford U - Computer Science	
38	Oren	Etzioni	Allen Institute for Artificial Intelligence	
39	Owain	Evans	Future of Humanity Institute	X
40	Sebastian	Farquhar	Oxford Global Priorities Project	
41	Chelsea	Finn	UC Berkeley grad student - AI	
42	Kay	Firth-butterfield	AI-Austin.org	
43	Dikeep	George	Vicarious Systems Inc.	
44	Ian	Goodfellow	OpenAI	X
45	Stephen	Goose	Human Rights Watch - arms division	
46	Joseph	Gordon-Levitt	actor & filmmaker	
47	Katja	Grace	Machine Intelligence Research Institute	X

48	Joshua D.	Greene	Harvard - Psychology	
49	Tom	Gruber	Apple	
50	Marta	Halina	Univeristy of Cambridge - Cognitive Science	
51	Verily	Harding	DeepMind	
52	Sam	Harris	author	
53	Demis	Hassabis	DeepMind	
54	Nick	Hay	Vicarious Systems Inc.	
55	John	Hering	Lookout	
56	Jose	Hernandez-Orallo	University of Valencia - Info Systems	
57	Reid	Hoffman	LinkedIn	
58	ShaoLan	Hsueh	entrepreneur/ Chineasy	
59	Tim	Hwang	Google	
60	Daniel	Kahneman	Princeton - Psychology, emeritus	
61	Rao	Kambohampali	Arizona State U - computer Science	
62	Angela	Kane	Vienna Centre for Disarmament and Non-Proliferation	
63	Holden	Karnofsky	GiveWell	X
64	Viktoriya	Krakovna	DeepMind; Future of Life Institute	X
65	Janos	Kramar	researcher - AI safety and deep learning	
66	Lawrence M.	Kraus	Arizona State U - physics	
67	Ramana	Kumar	Commonwealth Scientific and Industrail Research Org.	
68	Martina	Kunz	Phd cand., Cambridge; Future of Humanity Institute	X
69	Ray	Kurzweil	Google; Singularity U; author	X
70	Neil	Lawrence	Amazon Research Cambridge	
71	David	Leake	Indiana U - info science	
72	Yann	LeCun	Facebook	
73	Sean	Legassick	DeepMind	
74	Shane	Legg	DeepMind	
75	Jan	Leike	DeepMind	X
76	Sergey	Levine	UC Berkeley - Computer Science	
77	Fuxin	Li	Oregon State - EECS	
78	Patrick	Lin	California Polytechnic State U	
79	Moshe	Looks	Google	
80	William	MacAskill	Center for Effective Altruism, Oxford	
81	Richard	Mallah	Future of Life Institute	
82	Jason	Matheny	Intelligence Advanced Research Projects Activity	
83	Yutaka	Matsuo	University of Tokyo	
84	Andrew	Maynard	Arizona State U - School for the Future of Innovation	
85	Andrew	McAfee	MIT Institute on the Digital Economy	
86	Tasha	McCauley	GeoSim Systems	X
87	Tom	Mitchell	Carnegie Mellon - Machine Learning	
88	Elon	Musk	SpaceX, Tesla, OpenAI	X
89	Andres	Ng	Baidu	
90	Peter	Norvig	Google	
91	Sean	OhEigartaigh	Center for the Study of Existential Risk, Cambridge	X
92	Catherine	Olsson	OpenAI	X
93	Steve	Omohundro	Possibility Research / Self Aware Systems	
94	Toby	Ord	Oxford - Philosophy; Giving What We Can	X
95	Laruent	Orseau	DeepMind	X
96	Pedro	Ortega	DeepMind	
97	Long	Ouyang	independent AI researcher	
98	Claudia	Passos-Ferreira	Columbia/NYU - consciousness	

99	Lucas	Perry	Future of Life Institute	X
100	Tomaso	Poggio	MIT - Brain & Cognitive Sciences	
101	Gill	Pratt	Toyota Research Institute	
102	Hew	Price	Oxford - Phil.; Center for the Study of Existential Risk	X
103	Martin	Rees	UK Astronomer Royal	X
104	Heather	Roff	Oxford - global security	
105	Anthony	Romero	ACLU	
106	Francesca	Rossi	IBM Watson Center	
107	Jonathan	Rothberg	"NextGen" DNA sequencing	
108	Daniela	Rus	MIT - EECS	
109	Stuart	Russell	UC Berkeley - Computer Science	X
110	Jeffrey	Sachs	Columbia - Economics	
111	Anna	Salamon	Center for Applied Rationality	X
112	David	Sanford	Office of Reid Hoffman, LinkedIn	
113	Matt	Scherer	author; info tech & jobs	
114	Jurgen	Schmidhuber	Swiss AI Lab IDSIA	
115	Eric	Schmidt	CEO, Alphabet/Google	
116	Bart	Selman	Cornell U - computer Science	
117	Andrew	Serazin	Templeton World Charity Foundation	
118	Carl	Shulman	Future of Humanity Institute	X
119	Scott	Siskind	psychiatrist; blogger	
120	Andrew	Snyder-Beattie	Future of Humanity Institute	X
121	Nate	Soares	Machine Intelligence Research Institute	X
122	Marin	Soljatic	MIT - Physics	
123	Jacob	Steinhardt	Stanford grad student; Open Philanthropy Project	X
124	Bas	Steunebrink	Swiss AI Lab IDSIA	X
125	Mustafa	Suleyman	DeepMind	
126	Ilya	Sutskever	OpenAI	X
127	Richard	Sutton	U of Alberta, computer science	
128	Jaan	Tallinn	Skype, Kazaa, Future of Life Institute	X
129	Alexander	Tamas	Vy Capital	X
130	Jessica	Taylor	Machine Intelligence Research Institute	X
131	Max	Tegmark	MIT - Physics; Future of Life Institute	X
132	Sam	Teller	SpaceX, Tesla, Open AI	X
133	Marty	Tenenbaum	Cancer Commons	
134	Joshua	Tenenbaum	MIT - Cognitive Science	
135	Kristinn	Thorisson	Reykjavik U - Computer Science	
136	Helen	Toner	Open Philanthropy Project	X
137	Moshe	Vardi	Rice University - Computer Science	
138	Manuela	Veroso	Carnegie Mellon - somputer science	
139	Wendell	Wallach	Yale - technology & Ethics	
140	Toby	Walsh	U of New South Wales - AI	
141	Kelly	Walter	Google	
142	David	Weld	U Washington - Computer Science	
143	Adrian	Weller	U Cambridge - Computational & Biological Learning; Leverkuhn Cent	X
144	Michael	Wellman	U of Michigan - computer science	
145	Meredith	Whittaker	Google	
146	Roman	Yampolskiy	U of Louisville - computer engineering; MIRI	X
147	Eliezer	Yudkowsky	Machine Intelligence Research Institute	X
148	Brian	Ziebart	U of Illinois-Chicago - computer science	
149	Shivon	Zillis	venture capitalist; Open AI advisor	X

ATTACHMENT L. NOTES ON PARTICIPANTS AT THE ASILOMAR “BENEFICIAL AI 2017” MEETING

The “[Beneficial AI 2017](#)” meeting at Asilomar was the follow-up to the “Beneficial AI” meeting held in 2015 in Puerto Rico. The main meeting was held from Thurs-Sunday Jan 5-8. It was preceded by a series of workshops held Mon-Wed Jan 2-4.

The 2017 meeting drew 149 participants.

An eyeball scan of the 149 participants identified many of likely significant wealth. These included:

<i>participant</i>	<i>affiliation</i>	<i>est. net worth 2017</i>
Elon Musk	Paypal, Tesla, SpaceX, etc.	\$ 20 billion
Eric Schmidt	former CEO, Google /Alphabet	15 billion
Demis Hassabis	Co-founder/CEO DeepMind	6 billion
Sam Altman	Loopt, Y-Combinator	5 billion
Reid Hoffman	LinkedIn, Paypall	3.3 billion
Nicholas Berggruen	Berggruen Institute	2 billion
Jaan Tallinn	Skype, Kazaa	?
Alexander Tamas	Vy Capital	?
DiLeep George	Vicarious	?
John Hering	Lookout	?

These 10 participants alone may qualify the “Beneficial AI 2017” meeting as having had the single greatest concentration of wealth of any conference ever held at the Asilomar conference center since its founding in 1913.

Noted non-AI academics who presented at the Jan 2017 conference included:

Jeffrey Sachs	economics	Columbia University
Daniel Kahneman	psychology	Princeton University emeritus; Nobel Laureate 2002
Lord Martin Rees	physics	Cambridge University ; UK Astronomer Royal
Wendell Wallach	bioethics	Yale; The Hastings Center

Participants from government included:

Irakli Beridze	United Nations Interregional Crime and Justice Research Institute
Angela Kane	former United Nations High Representative for Disarmament Affairs

Participants from non-AI civil society organizations included:

Stephen Goose	Director, Human Right Watch Arms Control Division
Anthony Romero	Executive Director, American Civil Liberties Union (ACLU)

Of the 149 participants 42 were affiliated with one of the AI+ advocacy groups and 8 with one of the Effective Altruism groups, for a total of 50 (34%) participants from the AI+ base. The two organizations sending the largest number of participants were the Future of Life institute (11) and the Future of Humanity Institute (10).

All six founding partners of the Partnership on AI – Google, DeepMind, Facebook, Apple, IBM and Microsoft – sent at least one participant.

The two co-founders/co-directors of AI NOW - Kate Crawford and Meredith Whittaker - participated.

Of the 149 participants ~ 27 (18%) were women.

ATTACHMENT M. “THE ASILOMAR AI PRINCIPLES” - 11 JANUARY 2017.

Released after the conclusion of the “[Beneficial AI 2017](#)” Conference, Jan 3-8, 2017, held at Asilomar CA.

Artificial intelligence has already provided beneficial tools that are used every day by people around the world. Its continued development, guided by the following principles, will offer amazing opportunities to help and empower people in the decades and centuries ahead.

RESEARCH ISSUES

- 1) Research Goal: The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.
- 2) 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, such as:
 - How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked?
 - How can we grow our prosperity through automation while maintaining people’s resources and purpose?
 - How can we update our legal systems to be more fair and efficient, to keep pace with AI, and to manage the risks associated with AI?
 - What set of values should AI be aligned with, and what legal and ethical status should it have?
- 3) Science-Policy Link: There should be constructive and healthy exchange between AI researchers and policy-makers.
- 4) Research Culture: A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.
- 5) Race Avoidance: Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.

ETHICS AND VALUES

- 6) Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.
- 7) Failure Transparency: If an AI system causes harm, it should be possible to ascertain why.
- 8) Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.
- 9) 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.
- 10) 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.
- 11) Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.
- 12) 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.

- 13) Liberty and Privacy: The application of AI to personal data must not unreasonably curtail people’s real or perceived liberty.
- 14) Shared Benefit: AI technologies should benefit and empower as many people as possible.
- 15) Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
- 16) Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.
- 17) 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.
- 18) AI Arms Race: An arms race in lethal autonomous weapons should be avoided.

LONG-TERM ISSUES

- 19) Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.
- 20) 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.
- 21) Risks: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.
- 22) 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.
- 23) 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.

- - - - -

[Comments are pending. The February 1975 **Asilomar Conference on Recombinant DNA** resulted in a set of agreements among university and corporate biotechnology researchers about short-term research procedures intended to prevent potentially risky experimentation from being conducted. The meeting and the agreements were motivated by a desire to assure both the general public and the federal government that scientists could conduct themselves in such a manner that unwanted governmental oversight would be unnecessary. Since that time many “Asilomar Agreements” have been prepared by groups of scientists and industry leaders involved in controversial research, and for the same reasons. Many of these documents are intended to serve a largely public relations function. - RH]

ATTACHMENT N. KEY LEADERSHIP IN THE AI+ SUPERINTELLIGENCE ADVOCACY MOVEMENT

Max Tegmark is professor of cosmology at MIT and the author of *Humanity 3.0*, in which he speculates on human life and society transformed by a technological Superintelligence. He played a central role in establishing and leading the Future of Life Institute in building a network of AI advocates calling for development of “friendly AI”.

Elon Musk is an engineer and entrepreneur known for his role in developing PayPal, and his roles in founding and leading Tesla Inc, SpaceX, Solar City, Neuralink and other ventures. He co-founded OpenAI and has funded many artificial intelligence initiatives. His 2017 net worth was estimated at ~ \$20 billion.

Nick Bostrom is founding director of the Future of Humanity Institute at Oxford Martin School and Professor of Philosophy at Oxford. He is known for his work on global catastrophic risk, and on risks possibly posed by AI. He is the author of *Superintelligence: Paths, Dangers, Strategies*. He identifies as a Transhumanist, and believes that we are very likely living in a computer simulation.

Seán Ó hÉigeartaigh is Executive Director of Cambridge University’s Centre for the Study of Existential Risk (CSER). He worked previously at the Future of Humanity Institute. His current research focuses on the risks of AI, emerging technologies, technology policy, synthetic biology and evolutionary theory. He has a PhD in genomics.

Stuart Russell is a professor of Computer Science, and of Engineering, at U.C. Berkeley, adjunct professor of Neurological Surgery at U.C. San Francisco, and Vice-Chair of the World Economic Forum’s Council on AI and Robotics. He is co-author of the leading university textbook on artificial intelligence. He founded and helps lead the Center for Human-compatible Artificial Intelligence (CHAI) at UC Berkeley.

Huw Price is Academic Director of the Centre for the Study of Existential Risk (CSER), as well as of the Leverhulme Centre for the Future of Intelligence (LCFI), both at Cambridge. Earlier he was professor of Philosophy at the University of Sydney. He is a Fellow of the British Academy and Past President of the Australasian Association of Philosophy

Jaan Tallinn is a founding engineer of Skype and Kazaa. He co-founded the Centre for the Study of Existential Risk (CSER) and the Future of Life Institute (FLI) and philanthropically supports these and other existential risk research organizations. He is a partner at Ambient Sound Investments, an active angel investor, and has served on the Estonian President’s Academic Advisory Board.

Lord Martin Rees is Astronomer Royal for the UK, Emeritus Professor of Cosmology and Astrophysics at University of Cambridge, and former President of the Royal Society. He is known for major contributions regarding the cosmic microwave background radiation, the origin of black holes, and more. He co-founded the Centre for the Study of Existential Risk (CSER), and is author of *Our Final Century?*, which focuses on existential risk.

George Church is professor of Genetics at Harvard Medical School and of Health Science and Technology at Harvard University and MIT. He’s known as a leader in developing techniques of synthetic biology and of human genetic modification. He is attempting to “de-extinct” the Woolly Mammoth. He is a key link between the AI and the human genetic enhancement movements.

- - -

Stephen Hawking was Director of Research at the Centre for Theoretical Physics at Cambridge and at the Centre for Theoretical Cosmology within the University of Cambridge. He was known for his study of black holes, his advocacy of the many-worlds interpretation of quantum physics, and his popular book on cosmology. For over a decade prior to his passing in March 2018, Hawking spoke frequently in support of AI+ superintelligence and extra-terrestrial colonization.

ATTACHMENT P. OP-ED: LORD MARTIN REES ON HIS OPTIMISTIC VIEW OF THE HUMAN FUTURE

The Anthropocene epoch could inaugurate even more marvelous eras of evolution

By Martin Rees

The darkest prognosis is that bio, cyber or environmental catastrophes could foreclose humanity's potential. But there is an optimistic option

THE GUARDIAN. Monday 29 August 2016

On Christmas Eve 1968, the Apollo 8 astronaut William Anders took a photograph of the view outside the window as his spaceship orbited the moon. The now iconic Earthrise image shows our half-moon blue planet under a decoration of clouds rising from the blackness of space over the lunar surface.

The picture encapsulated Earth's precariousness in the cosmos and, for many, contained a message of humility and stewardship for our home.

We've had Earthrise and images like it from the Apollo missions for half a century now. But suppose some aliens had been viewing our planet for its entire 4.5bn-year history. What would they have seen?

Over nearly all that immense time, changes would have been very gradual: continents drifted; the ice cover waxed and waned; successive species emerged, evolved and became extinct during a succession of geological eras.

But visible change has accelerated rapidly in the past few thousand years – a tiny sliver of the Earth's history. Now geologists have decided those changes have been so profound, so global and so permanent that our catalogue of the Earth's history needs to change accordingly. Since the last ice age, around 11,000 years ago, human civilisation has flourished in the climatically benign Holocene. Now they believe that epoch has come to an end and we have entered a new human-influenced age, the Anthropocene.

The changes that our aliens could observe from space are not hard to spot. In just the last few thousand years, the patterns of vegetation altered much faster than before. These human-induced changes signalled the start of agriculture.

And human activity manifested itself in other ways that will leave traces in the geological record. Constructs of concrete and metal sprawled across the continents; domesticated vertebrates numerically overwhelmed wild ones; the carbon dioxide in the atmosphere rose anomalously fast; traces appeared of plutonium and other "unnatural" substances.

The imaginary aliens watching our world would have noticed something else unprecedented in geological history. Rockets launched from the planet's surface escaped the biosphere completely. Some were propelled into orbits around the Earth; some journeyed to the moon and planets.

What do these trends portend? Should we be optimistic or anxious? It's surprising how little we can confidently predict – indeed, we can't predict as far ahead as our forebears could. Our medieval ancestors thought the Earth was only a few thousand years old, and might only last another thousand. But they didn't expect their children's lives to be very different from theirs. They built cathedrals that wouldn't be finished in their lifetime.

Our time horizons, both past and future, now stretch billions of years, not just thousands. The sun will keep shining for about another 6bn years. But ironically we can't forecast terrestrial trends with as much confidence as our ancestors could. Their lives and environment changed slowly from generation to generation. For us, technological change is so fast that scenarios quickly enter the realm of wild conjecture and science fiction.

But some things we can predict, at least a few decades ahead. By mid-century, the world will be more crowded, and our collective footprint will be heavier. World population is now 7.2 billion and is forecast to rise to around 9 billion by 2050. Experts predict continuing urbanisation – and huge growth of megacities such as Lagos, São Paulo and Delhi. Population trends later this century depend largely on what happens in Africa, where some UN predictions foresee a further doubling between 2050 and 2100.

Moreover, if humanity’s collective impact on nature pushes too hard against what Johan Rockstrom calls “planetary boundaries”, the resultant “ecological shock” could irreversibly degrade our biosphere. And if global warming reaches a tipping point that triggers melting of Greenland’s ice, coastlines a millennium hence would be drastically different. Extinction rates are rising. We’ve only identified about two million of the (estimated) 10 million living species: we’re destroying the book of life before we’ve read it. To quote the great ecologist EO Wilson, “mass extinction is the sin that future generations will least forgive us for”.

The darkest prognosis for the next millennium is that bio, cyber or environmental catastrophes could foreclose humanity’s immense potential, leaving a depleted biosphere. Darwinian selection would resume, perhaps leading, in some far-future geological era, to the re-emergence of intelligent beings. If this happens, or if there are aliens out there who actually visit and study the Earth, then, digging through the geological record (and applying archaeological techniques as well) they would uncover traces of a distinctive transient epoch, and ponder the all-too-brief flourishing of a species that failed in its stewardship of “spaceship Earth”.

But there is an optimistic option.

Human societies could navigate these threats, achieve a sustainable future, and inaugurate eras of post-human evolution even more marvelous than what’s led to us. The dawn of the Anthropocene epoch would then mark a one-off transformation from a natural world to one where humans jumpstart the transition to electronic (and potentially immortal) entities, that transcend our limitations and eventually spread their influence far beyond the Earth.

Even in a cosmic time-perspective, therefore, the 21st century is special. It marks our collective realization that the Anthropocene has begun – and it’s a century when human actions will determine how long that epoch lasts.

- - - - -

*Comment: The first 14 paragraphs could almost have been written by Greenpeace or Bill McKibben. The 15th paragraph, which I’ve highlight in **bold**, is evidence, in my opinion, of sociopathy. It’s a stunning example of where the scientific and technological mind almost necessarily leads if not formatively grounded in a broader and deeper set of human values. – RH*

ATTACHMENT Q. RECENT NOTABLE BOOKS AND OTHER PUBLICATIONS ADDRESSING IA

Q.1. Books that mostly celebrate the techno/AI future are shown without preceding marks. More critical books are marked by an *. Those judged to be somewhat in between are marked with a ^. Comments follow in Q.4.

1994. Out of Control: The New Biology of Machines, Social Systems and the Economic World.	Kevin Kelly
1999. The Age of Spiritual Machines: When Computers Exceed Human Intelligence.	Ray Kurzweil
2003. Our Final Century: Will the Human Race Survive the Twenty-first Century?	Lord Martin Rees
2006. The Singularity Is Near: When Humans Transcend Biology.	Ray Kurzweil
*2010. The Shallows: What the Internet is Doing to Our Brains.	Nicholas Carr
2011. What Technology Wants.	Kevin Kelly
2012. How to Create a Mind: The Secret of Human Thought Revealed.	Ray Kurzweil
*2013. To Save Everything, Click Here. The Folly of Technological Solutionism.	Evengy Morozov
2014. Superintelligence: Paths, Dangers, Strategies.	Nick Bostrom
2014. Smarter than Us: The Rise of Machine Intelligence.	Stuart Armstrong
2014. Abundance: The Future Is Better Than You Think Paperback.	Peter Diamandis
2014. Regenesi: How Synthetic Biology Will Reinvent Nature and Ourselves.	George Church, Ed Regis
2014. The Artificial Intelligence Revolution: Will Artificial Intelligence Serve Us or Replace Us?	Louis Del Monte
*2014. The Glass Cage: Automation and Us.	Nicholas Carr
2015. What to Think about Machines That Think.	John Brockman, ed.
2015. Surviving AI: The Promise and Peril of Artificial Intelligence -	Calum Chace
2015. The Second Intelligent Species: how humans will become as irrelevant as cockroaches.	Marshall Brain
^2015. Machines of Loving Grace: The Quest for Common Ground between Humans and Robots.	John Markoff
2016. The Age of Em: work, love and life when robots rule the world.	Robin Hanson
2016. Life 3.0: Being human in the age of artificial intelligence.	Max Tegmark
2016. The Economic Singularity: Artificial intelligence and the death of capitalism.	Calum Chace
2016. Our Final Invention: artificial intelligence and the end of the human era:	James Barrat
2016. Augmented: Life in the Smart Lane.	Bret King
2016. Rise of the Robots: Technology and the Threat of a Jobless Future.	Martin Ford
2016. Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies.	E. Brynjolfsson, A. McAfee
*2016. Chaos Monkeys: Obscene Fortune and Random Failure in Silicon Valley.	Antonio G. Martinez
2017. Homo Deus: A Brief History of Tomorrow.	Yuval Harari
2017. The Inevitable: Understanding the 12 Technological Forces That Will Shape Our Future.	Kevin Kelly
^2017. Valley of the Gods: A Silicon Valley Story.	Alexandra Wolf
*2017. Move Fast and Break Things.	Jonathan Tapliin
*2017. World Without Mind: The existential threat of Big Tech.	Franklin Foer
*2017. The Know-It-Alls: the Rise of Silicon Valley as a Political Powerhouse.	Noam Cohen
*2017. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.	Cathy O'Neil
*2017. The Four: The Hidden DNA of Amazon, Apple, Facebook, and Google.	Scott Galloway
*2017. Dawn of the New Everything: Encounters with Reality and Virtual Reality.	Jaron Lanier
*2017. Technically Wrong: Sexist Apps, Biased Algorithms, and other Threats of Toxic Tech.	Sara Wachter-Boettcher
*2018. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.	Virginia Eubanks
*2018. Algorithms of Oppression: How Search Engines Reinforce Racism.	Safiya Umoja Noble
*2018. Brotopia: Breaking Up the Boys' Club of Silicon Valley.	Emily Chang
*2018. Antisocial Media: How Facebook Disconnects Us and Undermines Democracy.	Siva Vaidhyanathan

Q.2. I list separately books that focus critically on the impact of online and related technologies on personal, family and community lives. See discussion of these in **Attachment U**.

2011. Alone Together: Why We Expect More from Technology and Less from Each Other.	Sherry Turkle
2013. To Save Everything, Click Here. The Folly of Technological Solutionism.	Evengy Morozov
2016. Disconnected: How To Reconnect Our Digitally Distracted Kids.	Thomas J Kersting
2017. Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked.	Adam Alter
2017. Glow Kids: How Screen Addiction Is Hijacking Our Kids - and How to Break the Trance.	Nicholas Kardara
2017. iGen: Why Today's Super-Connected Kids Are Growing Up Less Rebellious, More Tolerant, Less Happy--and Completely Unprepared for Adulthood--and What That Means for the Rest of Us.	Jean Twenge

Q.3. Books not included. In the foregoing lists I haven't included several categories of books:

- * popular introductory books on artificial intelligence limited mostly to nuts/bolts/applications.
- * textbooks or specialized academic texts.
- * the many how-to-succeed-in-the-new- AI-world books.
- * biographies of noted people: Turing, Jobs, Musk, Gates et al.
- * books focused on a single firm: Apple, Google, Facebook etc.
- * histories of Silicon Valley, the Internet, etc.
- * fiction, science fiction or science fantasy.

Q.4. Blogs and journalism

[*Comments in preparation.*] Two consistently smart authors are Nicholas Carr, whose blog [Rough Type](#) covers a wide range of technology/culture/economics topics, with a focus on digital/web/AI, and Sue Halpern, who appears regularly in *The New York Review of Books* and other outlets: see e.g. Halpern 2016a, 2016b, 2015, 2014, 2011. [MIT Technology Review](#) is tech-friendly but not reluctant to publish strongly critical pieces.

Q.5. Comments

In the 1990s and 2000s many books that reflected hyper techno-utopian themes were published, as were a lesser number of critical books, but the focal technologies of both were more commonly biotechnological: genetic engineering, human cloning, designer babies, neo-eugenics and the like. Kurzweil was exceptional in focusing on computer technology and the transcendent image of the Singularity, although he was fully supportive of human genetic modification as well.

In the 2010s the focus began to shift, and books celebrating the AI+ advocacy future dominated during the three years 2014, 2015 and 2016. Over this period both the deployment of artificial intelligence devices and features and the enthusiasm for AI as the dominating technology of the human future grew rapidly.

By 2017, however, serious questions began to be acknowledged regarding the Silicon Valley mentality and its impact on the larger society. People began challenging the effective-monopoly status of the big 5 tech corporations (Apple, Google/Alphabet, Facebook, Amazon, Microsoft), and the impact of tech on economic inequality, democratic governance, community life, our children and our relationships and more. Books addressing these topics appear to have supplanted, at least for the moment, further paeans to the cyberculture digerati, the Superintelligent robotic AI futurists, et al. By 2018 books taking the techno-dominant mentality even more strongly to task from racial, gender and class perspectives began being published and promoted.

It notable, however, that neither the 2017 wave of liberal tech criticism, nor the 2018 wave of race/gender/class tech criticism, as important as they are, directly challenge the two mega-claims of the AI+ advocacy movement. One claim is that, *like it or not*, humanity will soon be made obsolete by advanced, Superintelligent AIs, and will either be violently exterminated, coldly pushed aside, or lovingly assimilated by them. The second, more moderate, claim is that even though the robot takeover scenario is adolescent sci-fi, we can still expect, *like it or*

not, that advanced AI will eliminate the need for the great majority of manual, service and blue collar jobs, and many white collar and professional jobs as well. On this scenario, perhaps 15% of the world's population will be needed to tend the machines. They'll reap the cornucopia of financial reward, while the remaining 85% of the world's 10 billion people live on food stamps and Facebook. For many, the only remaining question is whether the hand-outs are to be meager or generous. What is missing is an account of a human future of ecological integrity, economic justice and technological responsibility that could actually be embraced by large sectors of all the world's peoples.

Q6. FILM - A great number of movies over the past half-century have depicted futures dominated by AI, robots and futuristic technology more generally. The great majority portray these worlds dystopically. *2001: A Space Odyssey* (1968), *Blade Runner* (1982), *Gattaca* (1997), and *the Matrix* (1999), among others, have become recurring cultural referents in discussions of the human future. There is active debate over whether viewing these films leaves people more concerned or less concerned about the impact of these technologies on human life and society.

[Further notes in preparation - RH]

ATTACHMENT R. PUBLIC OPINION AND ARTIFICIAL INTELLIGENCE

I. INTRODUCTION

II. THE DATA

I found 12 published public opinion surveys, most of them conducted over the last three years. I selected 8 of these as best suited for this brief review:¹⁸

US1. 60 Minutes/Vanity Fair (CBS News): Artificial Intelligence	March 2016
US2. YouGov/Omnibus Research: Robots	July 2016
US3. Morning Consult National Tracking Poll	April 2017
US4. Pew Research Center: Automation in Everyday Life	October 2017
US5. Gallup/Northeastern University: Optimism and Anxiety	March 2018
US6. Fast and Horvitz: Long-Term Trends in the Public Perception of Artificial Intelligence	December 2016
UK1. British Science Association: Survey for British Science Week	March 2016
UK2. Royal Society/Ipsos-Mori: "Public Views on Machine Learning"	April 2017

Item US6 is not an opinion survey *per se*; it's an analysis of newspaper coverage of AI, structured as a measure of public perceptions of AI, over the three decades 1985-2015.

Six of the surveys are of Americans and two are of Britons. Most contain brief, incomplete descriptions of their methodology. Some report only general population results and others include reports by gender, ethnicity, income, geography, education and other demographic variables.

The surveys asked a range of questions, most of them falling within these categories:

- A. What's your general sentiment** about AI overall? Do you support/oppose it? Are you optimistic/pessimistic about it? Should research on it continue or not? Should it be regulated, and if so how much and by whom?
- B. Impact on the economy:** Is AI likely to help/hurt the economy? Increase/decrease the number of jobs? Increase/decrease economic inequality? If AI creates great unemployment, what should be done? Do you support/oppose universal basic Income programs? Other plans? If retraining is needed, how difficult will it be, and who should pay for it?
- C. Specific applications:** The surveys asked about the application of AI in office work and manufacturing, autonomous vehicles, health care, education, assistance for the elderly, law enforcement and more. They asked if respondents favored/opposed such applications of AI, if they felt these applications would help/hurt the economy, create/reduce unemployment, exacerbate/reduce economic inequality, and more.
- D. Deeper beliefs:** Do you believe that AI poses a threat to the long-term survival of humanity? Is AI our greatest existential threat? Can a computer ever be considered truly alive?
- E. What arguments or factoids** would make you more or less likely to support or oppose further research on AI?
- F. Selected demographic and crosstab results:** ideology, gender, age, ethnicity, income, etc.
- G. Time Series:** How have public perceptions of AI changed over time?

III. SELECTED SURVEY RESULTS

A. WHAT GENERAL UNDERSTANDINGS, BELIEFS and OPINIONS DO PEOPLE HAVE ABOUT AI?

79% say AI is currently having a positive impact on their lives, and 77% are largely optimistic about the impact that AI will have on people’s lives and work in the future. [US5]

53% say it is important to advance the field of AI. [US1]

The [US3] survey asked very early in the survey: “Do you think we should increase or decrease our reliance on AI?” They asked a similar question about half-way through they survey, after the respondent had been exposed to a mix of topics about AI: “Do you support or oppose continuing AI research?” They asked this second question verbatim near the very end of the survey. Here are the results:

	Increase/ support	decrease/ oppose	DK
a) increase or decrease our reliance on AI?	39	39	23
b) support or oppose continuing AI research?	51	31	18
c) support or oppose continuing AI research?	50	34	17

Here are responses to other questions:

	Yes	No	DK
AI is humanity’s greatest existential threat:	50	31	19
There should be national regulations on AI:	71	14	15
There should be international regulations on AI:	67	16	17

[US3]

Only 36% support federal AI funding for university AI research and only 26% support federal tax breaks for corporate AI research. [US5]

summation: Majorities feel positively about the current and future impact of AI on their lives and society, and most of these want to see AI research continue. Nonetheless, substantial minorities are uneasy about the impact of AI, and there is strong support for national and international regulation of AI. Further, majorities do not support federal funding for university or corporate AI research. Further still, fully half agree that AI is “humanity’s greatest existential threat.” It’s not clear from the materials reviewed, however, how this last question was posed or what respondents understood the question to mean.

B. IMPACT ON THE ECONOMY, JOBS AND THE WORKFORCE

How has AI impacted your own work situation so far?

55-70% feel that their jobs/careers have been positively impacted by word processing, email, social media, software, and smartphones, [US4]

In coming years, will AI help or hurt the economy?

help	hurt	no dif	DK
28	37	9	27

[US3]

40% felt that AI would make the economy as a whole much more efficient, but 56% did not. [fts]

Will AI lead to mass unemployment and/or permanent loss of jobs?

60% say AI will lead to fewer jobs within 10 years. [UK1]

73% of all respondents, and 82% of those with blue-collar jobs, think that AI will lead to a net loss of jobs. [US5]

83% think it's likely that robots will eventually learn to do most human tasks, and 50% think that this won't be good for the average person. [US2]

75% felt that an economy using AI would produce more and better paying jobs for humans. [US4]

How worried are people about their own jobs being eliminated or taken over by AI?

24% are worried that their jobs are currently at risk. [US2]

25% are worried that their jobs will be at risk over the coming 20 years. [US2]

30% think that AI will threaten their own jobs in their lifetime; 70% do not. [US4]

23% felt that their own jobs were at risk. [US5]

28% of those with less education feel that their own jobs were at risk. [US5]

If people lost jobs because of AI, how would they respond? Could they retrain for comparable jobs?

8% believe that AI will leave everyone with abundant new free time to pursue their passions and hobbies. [US2]

64% felt that people would have a hard time finding things to do with their lives. [US4]

40% felt that humans would find their jobs more meaningful since machines would be able to do the most unappealing parts of those jobs; 59% disagreed with this. [US4]

42% agreed that people would be able to focus less on work and more on things that really matter. [US4]

57% were not confident that they could obtain the training needed to obtain a job of comparable pay. [US5]

Will AI lead to greater inequality?

70% think a futuristic robotic workforce will lead to mass inequality. [US2]

76% felt that inequality would get much worse. [US4]

What should be done if AI does create mass unemployment?

[US4]

	<u>yes</u>	<u>no</u>
Limit robots/computers to jobs that are dangerous or unhealthy for humans:	85	14
Give all Americans a Guaranteed Income that meets basic needs:	61	38
The government has an obligation to take care of those whose jobs are displaced by robots/computers, even if that mean raising taxes substantially:	50	49
Give a Universal Basic Income to help those who lose their jobs because of AI:	48	52

NB: The roughly equal national divide on UBI masks a sharper partisan divide:

	<u>support UBI</u>	<u>oppose UBI</u>
Democrat	65	35
Republican	28	72
Independent	48	52

	<u>yes</u>	<u>no</u>	<u>DK</u>
Would you be happy if you could live without having to work?	53	36	11

[US2]

summation: Respondents have a generally dismal assessment of the impact of AI on jobs and the economy. Majorities of 60-70% and higher said they believed that AI would increase unemployment and lead to greater economic inequality. At the same time, fewer respondents (20-30%) feared that AI would put their own jobs at

risk. Overall response averaged ~ 53-46 in support of a universal basic income as a response to permanent job loss generated by AI. However, this overall response conceals a stronger partisan divide: Democrats strongly supported UBI, by 65-35%, while Republicans even more strongly opposed it, by 28-72%.

C. SPECIFIC APPLICATIONS

Self-driving vehicles

28% would feel comfortable in a fully self-driving vehicle; 65% would not. [US3]

23% say they'd be comfortable using fully self-driving cars. [US5]

20% would be comfortable sharing the road with fully self-driving trucks.

46% are *not worried about* self-driving cars; 53% *are* worried. [US4]

49% are *enthusiastic* about self-driving cars; 60% are *not*.

70-90% favor various strong regulations on self-driving cars, such as restricted lanes, restricted roads, and a requirement that passengers be aboard to handle emergencies. [US4]

summation: Fewer than half of respondents, and perhaps as few as one-quarter, feel positive about self-driving cars. Strong majorities want self-driving cars to be regulated.

Eldercare

It's been suggested that in the future frail older adults could have robot caregivers to help with chores, food, medicine, notification in the case of emergency, and other tasks, and thus allow these older adults to live in their own homes.

59% thought this suggestion sounded realistic.

44% are enthusiastic about this possibility; 56% are not.

47% are worried about this possibility; 53% are not.

41% would be interested in such robot caregiving; 59 would not.

64% felt that older adults who had such robots would feel more isolated.

70% felt that young people would feel less worried about caring for aging relatives if they had such robots.

49% support the use of AI as part of household support for the elderly or disabled. [UK1]

Use of personal data

61% don't mind the use of their personal data held in large data banks, so long as the data is anonymized.

75% don't mind sharing their anonymized genetic data for non-commercial medical research. [UK2]

70% believe that the US federal government should do more to protect the privacy of consumers when their personal information is being accessed and analyzed by firms using AI. [US5]

Other applications

62% would not trust AI for flying commercial aircraft.

53% would not trust AI for surgical procedures.

49% would not trust AI for driving public buses. [UK1]

70-90% support AI for military uses, space exploration and manufacturing.

15-20% support AI for care of children and the elderly. [UK2]

65-80% felt that within 20 years we will see AI used for medical diagnosis & treatment recommendations, near-complete automation of most stores and retail businesses, and robot/drone deliveries in cities. [US4]

43% have heard about proposals that AI could review resumes, conduct interviews, and make hiring decisions. 59% think that this sounds like a realistic possibility.

57% felt that AI systems could usefully be used to screen applicants for subsequent interviews with humans. 15-30% felt that AI systems could make better final hiring decisions that humans could. [fts]

Gradations of comfort levels

There are gradations of comfort levels regarding the uses of AI. Compare these results from [US3]:

AI to be used in:	Comfortable	Uncomfortable	DK
1. cleaning a house	61	31	8
2. running a factory	38	53	9
3. medical diagnosis	27	65	8
4. choosing a romantic partner	23	68	9

In another survey respondents were asked whether for each of eight AI applications they felt that the benefits outweighed the risks (“positive”), were equal (“neutral”), or the risks outweighed the benefits (“negative”). The survey used the phrase “machine learning” throughout rather than “artificial intelligence.” [fts]

Application	positive	neutral	negative	DK
1. Computers that can make informed investments in the stock market	18	31	41	11
2. Autonomous robots that can be used by the armed forces	22	23	48	7
3. Computers which show you website/ads based on your browsing habits	24	40	28	8
4. Driverless vehicles that can adapt to road and traffic conditions	27	22	45	6
5. Robots that can adapt to a home environment e.g. to help care for elderly	38	28	28	7
6. Computers that analyze medical records to help diagnose patients	40	23	30	6
7. Computers that can recognize speech and answer questions	54	27	13	6
8. Facial recognition computers that can help catch criminals	61	18	15	6

summation: Respondents seem more accepting of AI for routine applications such as “cleaning a house” (61%) or for applications unlikely to involve them personally, frequently or directly, such as “facial recognition computers to help catch criminals” (61%). They seem less accepting of applications of potentially high personal consequence, such as “making a medical diagnosis” (27%), “driving a car” (24%), “flying an airplane” (23%), or “choosing a romantic partner” (23%).

D. DEEPER BELIEFS AND CONCERNS

	yes	no	DK	
Can a computer ever be considered truly alive:	19	79		[US1]
Does AI pose a threat to the long term survival of humanity:	36			[UK1]
Is AI humanity’s greatest existential threat:		50		[US4]
AI is humanity’s greatest existential threat:	50	31	19	[US3]

summation: Given the sensationalistic media coverage of AI both as a utopian panacea and as robot Armageddon, and given its novelty, technical nature and only recent intense media coverage, I’m not sure what these responses to these questions mean. Extended focus group discussions might be more revealing.

E. ARGUMENTS AND FACTOIDS

A section of the **Morning Consult** poll [US3] asks respondents to say whether or not a certain argument or factoid would make them more or less likely to support AI research. There are 12 questions; in the first 6 the factoid is worded to presumably motivate greater support for AI research and in the second 6 to motivate lesser support.

Such survey results are used when designing marketing and political campaigns. I presume that most respondents are aware of this or figure it out *in situ*, but I have no idea how that awareness might influence their responses. Each question asks the respondent to evaluate on the spot how a strange, unfamiliar factoid might strengthen or weaken their opinion regarding a strange, unfamiliar technology.

Here are the factoids and respondent responses. “**More**” = hearing the factoid makes them more inclined to support AI research; “**Less**” = less inclined; “**Null**” = no change in support for AI research; “**DK**” = don’t know.

Factoid	More	Less	Null	DK
1. AI can be used to overcome human limitations in space and the depths of the ocean.	61	20	6	14
2. AI can make daily life a lot easier.	52	24	9	15
3. AI can replace human beings in many labor intensive tasks.	40	41	6	13
4. Machines can perform the same task repeatedly without stopping and require no sleep	56	23	8	13
5. AI can simulate brain function and help with diagnosis/treatment of neurological problems	51	25	6	17
6. Robotic surgery may be safer because it reduces the risk for human error.	47	31	6	15
7. Maintenance and repair of AI machines is expensive.	31	45	9	15
8. Intelligence is a gift of nature and it might not be ethically correct to replicate it on machines	31	38	14	17
9. Machines are not as efficient as humans in altering their responses	29	47	8	15
10. Robots can cause mass unemployment.	31	51	6	12
11. Machines can become smart enough to control humans.	22	57	5	15
12. Abilities of humans may diminish if reliance on machines means they no longer need to use their full mental capacity.	30	51	6	14

Without knowing the purpose of the survey a real analysis is not possible, but some observations are in order:

- * Five of the six factoids intended to motivate support for further AI research do so, although not especially strongly so. The share of “null” and “DK” responses is 20-24%, which makes sense given the unfamiliar topic.
- * The one factoid that that fails to motivate greater support for further AI research, and in fact motivates greater *opposition*, is #3, which says that AI can replace humans in labor intensive tasks. This response is consistent with the high level of worry seen in all the surveys about AI causing mass unemployment.
- * The next “weakest” motivating factoid is #6, which connects AI with robotic surgery. This too is consistent with other survey responses, which suggest that people are wary of AI/robotics getting too close to their bodies, even when billed as “safer” than humans doing the same thing.
- * The 6 factoids (6-12) intended to motivate *less* support for AI research all do so, but note the high rates at which they also reinforce the *opposite* position. Consider e.g. item #10, “Robots cause mass unemployment.” Some 51% took that factoid as good reason not to support further AI research. But what about the 31% who took that as good reason to *support* further AI research? It’s possible that they figured that the research would be directed at finding means whereby AI does not in fact cause mass unemployment. If this is so, then the question is ambiguous and the results less useful. A second possible explanation might be called the transhumanist motivation. In Item #11, for example, the factoid that “Machines can become smart enough to control human beings” motivates 22% to *support* further AI research. Again, some may reason that this research would be intended to find ways to *prevent* such control from happening. Is it plausible that other respondents might sincerely *look approvingly* on a scenario in which AI controls humans, and support more AI research towards that end? I don’t know, but would think not.
- * Item #8 is something of an anomaly. It’s the only item that seeks to motivate less support for AI research by invoking an ethical stance regarding a transcendental value. By contrast, all the other items appeal to utilitarian

self-interest. This ethical/transcendental stance succeeds in motivating lesser support for AI, but by a lesser degree than the utilitarian factoids do. There is a noticeable spike in the number of respondents who say that the ethical/transcendental argument *doesn't affect* their opinion regarding further AI research either way. Why might this be? I have no good idea. Perhaps ethical/transcendental language is even more foreign to many respondents than is the language of artificial intelligence and simulated brains. Or it could be the opposite: some respondents might take ethics and transcendental values so seriously that they have difficulties with the wording of the injunction. They might believe that intelligence is not “a gift of nature” but rather a gift from God. Or the phrase “ethically correct” may put off some deeply situationally ethical respondents. As it turns out, the crosstabs show no large differences among the responses of the general public on item 8 and those of respondents who identify with any of six categories of religious belief. So the questions regarding the meaning of item #8 remain open.

summation: Given the multiple uncertainties concerning this survey I'll pass on doing a summation.

F. SELECTED DEMOGRAPHIC AND CROSSTAB RESULTS

AGE

Would it be a good thing or a bad thing if robots began to perform much of the work humans currently do?	AGE				
	<30	30-44	45-64	65+	
A good thing	37	40	37	37	[US2]
A bad thing	51	49	51	51	

	18-29	30-44	45-54	55-64	65+	
AI will help the economy:	40	34	21	20	22	[US3]
AI will hurt the economy:	27	35	44	42	33	

% in each age group saying they would be **very comfortable** having an AI:

	18-29	30-44	45-54	55-64	65+	
drive a car:	20	18	7	7	5	[US3]
pick a romantic partner	18	15	6	6	2	
fly an airplane	15	14	5	5	4	
make a medical diagnosis	14	14	5	4	4	
Strong support for AI research:	20	22	11	11	13	

	18-35	36-50	
AI will have a very/mostly positive impact on work and life in the US:	79	75	[US5]
AI will widen the gap between rich and poor in the U.S.:	69	60	
Increased use of AI will eliminate more jobs in the US than it creates:	78	75	

GENDER

	Men	Women	
Feel we should increase our reliance on AI:	52	26	[US3]
Feel that AI is safe:	54	29	"
Feel that AI will help the economy:	38	18	"
Support continuing AI research:	62	39	"
Approve of robots doing much of the work that humans do now:	48	26	[US2]
Feel optimistic about AI:	28	17	[UK1]

A portion of the difference between men's and women's responses as shown can be attributed to the fact that larger shares of women than men regularly answered “Don't know.”

How worried are you that in 20 years a robot will be able to do your job? [US2]

Worried	30	20
Not Worried	65	74

For a long list of secondary questions, e.g., “Who should pay for AI research? Should AI be regulated? If so, by what level of government?” responses of men and women did not differ appreciably.

RACE/ETHNICITY

	Should we increase or decrease reliance on AI?			Is AI safe or unsafe?			Will AI help or hurt the economy?				[US3]
	<i>increase</i>	<i>decrease</i>	<i>DK</i>	<i>safe</i>	<i>unsafe</i>	<i>DK</i>	<i>help</i>	<i>hurt</i>	<i>no diff</i>	<i>DK</i>	
White	37	40	23	40	39	31	27	36	9	28	
African American	37	39	24	39	37	24	27	32	11	29	
Hispanic	55	28	17	54	29	17	42	34	8	15	
Others	60	23	17	59	27	16	43	34	4	18	

“Other” presumably includes Asians, Pacific Islanders, South Asians and Native Americans.

The responses to all three queries (and to a fourth not shown, asking if we should continue AI research), show a consistent pattern:

- * Significant shares of all ethnicities appear broadly supportive of, broadly wary of, and uncertain regarding, AI.
- * There’s a rough, decreasing order of support for AI: Others => Hispanics => African Americans => Whites.
- * Others and Hispanics often give similarly supportive responses, and African American and Whites often give similarly wary responses. The comfort “gap” is highest between Hispanics and African-Americans.

This pattern is reproduced in other question in the survey, e.g. **Should we continue AI research?**

Here are results when respondents are asked **how comfortable or uncomfortable they would be with the use of AI** for particular purposes. I show here only the results for “comfortable”

	<i>drive a car?</i>	<i>chose a romantic partner?</i>	<i>fly an airplane?</i>	<i>perform a medical diagnosis?</i>
White	26	20	21	24
African American	37	34	28	34
Hispanic	41	43	36	39
Other	38	33	32	39

Here we see that:

- * Whites are again the least comfortable with AI, and African Americans again the next least comfortable.
- * Hispanics and Others have switched positions as the most and second-most AI-comfortable, although their differences are mostly small.
- * African Americans are now consistently more comfortable with these uses than are Whites. The comfort “gap” is now greatest between Whites and the rest.

Major differences among the four ethnic group categories mostly disappears in this final set of responses:

Agree or disagree: Artificial intelligence is humanities greatest existential threat?

	<i>agree</i>	<i>disagree</i>	<i>DK</i>
White	20	31	19
African American	28	28	17
Hispanic	28	24	18
Other	24	31	17

Good or bad if robots do currently human work? Worried that robots could take your job in 20 years? [US2]

	<u>good</u>	<u>bad</u>	<u>DK</u>	<u>worried</u>	<u>not worried</u>	<u>DK</u>
White	37	50	13	26	70	4
African American	30	53	17	34	58	8
Hispanic	29	56	16	22	75	3
Other	43	44	4	8	78	15

summation: Whites and African Americans appear to be less comfortable with AI and AI applications than Hispanics and Others (presumably including Asians, Pacific Islanders, South Asians and Native Americans.) Again, it's difficult to know how meaningful these results are. Multiple regression might show that some variable other than ethnicity was the driving factor. Even if statistically valid, these results give no hint of *why* opinion on AI should vary by ethnicity. Again, focus group discussion could be helpful.

EDUCATION / INCOME

Results from US2, US3 and US5 mostly did not show particularly significant or consistent differences in the responses received from those in different education or income categories:

	<u>< B.A.</u>	<u>B.A. or ></u>
AI will have a positive impact on life and work in the future:	74	82
AI will eliminate more jobs than it creates:	74	72

Would it be good or bad if robots do much of the work now done by humans? [US2]

	<u>< \$50K</u>	<u>\$50-100K</u>	<u>\$100K+</u>
good	34	41	36
bad	51	50	46
not sure	14	9	18

Would you be comfortable or uncomfortable with an AI driving a car? [US3]

	<u>< \$50K</u>	<u>\$50-100K</u>	<u>\$100K+</u>
comfortable	28	28	34
uncomfortable	64	67	60
DK	8	4	5

The same general pattern held for other tasks involving AI, such as flying an airplane, choosing a romantic partner or making a medical diagnosis: across all three income categories, 20-35% were comfortable, 60-70% of respondents were uncomfortable, and 5-10% were DK. [US3]

The single question that elicited noticeably different responses from the 3 income categories involved the impact of AI on the economy: Do you think increased use of AI will help or hurt the economy? [US3]

	<u>< \$50K</u>	<u>\$50-100K</u>	<u>\$100K+</u>
help	25	31	41
hurt	36	38	29
no difference	9	8	7
DK	30	24	24

RELIGION

For the most part there appear to be few major differences in the responses based on religious orientation. Three typical sets of responses are shown here: [US3]

	We should increase our reliance on AI:		AI is humanity’s greatest existential threat:		Would you feel comfortable having AI choose a romantic partner:	
	<u>yes</u>	<u>no</u>	<u>yes</u>	<u>no</u>	<u>yes</u>	<u>no</u>
Protestant	39	40	51	34	19	77
Roman Catholic	46	34	54	29	20	61
Atheist/Agnostic/none	38	32	43	34	19	66
Other	35	46	47	33	24	69
Jewish	40	32	32	36	30	62
Evangelical	40	40	58	27	27	67
Non-evangelical Catholic	41	38	49	32	22	71
All Christian	41	39	54	30	24	69
<u>All Non-Christian</u>	<u>37</u>	<u>38</u>	<u>45</u>	<u>34</u>	<u>21</u>	<u>68</u>
All Adults	39	39	50	31	23	68

IDEOLOGY / PARTISANSHIP

The [US3] surveys asked 14 questions in which the level of support for or comfort with AI was directly queried. In *all* these, respondents identified as *liberals* expressed substantially more support for or comfort with AI than did *moderates* or *conservatives*. Responses from moderates and conservatives typically showed *significantly* less support for or comfort with AI than did those from liberals and were typically within 2-3 points of one another. Here are two instances: [US3]

1. Do you support or oppose artificial intelligence?

<u>ideology</u>	<u>support</u>	<u>oppose</u>	<u>DK</u>
Liberal	63	25	11
Moderate	47	35	17
Conservative	50	35	15
Tea Party Supporter	63	28	10

2. Would you be comfortable or uncomfortable having an AI fly an airplane:

<u>ideology</u>	<u>comfortable</u>	<u>uncomfortable</u>	<u>DK</u>
Liberal	34	61	5
Moderate	21	73	7
Conservative	18	78	4
Tea Party Supporter	33	63	4

In all 14 sets of responses those from “Tea Party Supporters” track very closely those of *liberals*. Speculation on this unexpected result is deferred pending examination of survey methodology.

The responses to an additional and different sort of question *did not* seem to be strongly motivated by ideology:

3. Do you agree or disagree: AI is humanity's greatest existential threat.

<i>ideology</i>	<i>agree</i>	<i>disagree</i>	<i>DK</i>
Liberal	55	35	12
Moderate	51	30	18
Conservative	51	33	16
Tea Party Supporter	58	32	10

Responses by political party

1. Would it be a good thing or a bad thing if robots began to perform much of the work humans currently do?

<i>party</i>	<i>good</i>	<i>bad</i>	<i>DK</i>	[US2 – July 2016]
Republican	33	50	17	
Democrat	35	56	8	
Independent	39	47	15	

2. Do you think we should increase or decrease our reliance on AI?

<i>party</i>	<i>increase</i>	<i>decrease</i>	<i>DK</i>	[US3 – April 2017]
Republican	43	37	20	
Democrat	42	37	20	
Independent	31	40	29	

3.	Good/Bad for the economy	OK w AI driving a car? Yes/No	Should we continue AI research? Yes/No	Is AI humanity's greatest existential danger? Yes/No
Republican	31/38	29/65	52/34	59/27
Democrat	29/34	33/62	52/34	47/35
Independent	25/35	24/64	45/32	43/32

Cross-tabulations:

If we breakout the results shown in box 2 above to show gender preference by party, we see that the aggregate results conceal a deeper divide:

4. Do you think we should increase or decrease our reliance on AI?

<i>party</i>	<i>increase</i>	<i>decrease</i>	<i>DK</i>	[US3]
Republican	(43)	(37)	(20)	
<i>men</i>	58	28	14	
<i>women</i>	26	47	27	
Democrat	(42)	(37)	(20)	
<i>men</i>	55	32	13	
<i>women</i>	31	39	20	
Independent	(31)	(40)	(29)	
<i>men</i>	43	33	23	
<i>women</i>	26	46	34	

5. If we further break out those men and women who believe we should *strongly increase* our reliance on AI from those who believe we should only somewhat increase it, the divide is stronger still:

<u>party</u>	<u>strongly</u>	increase reliance on AI	[US3]
Republican	(15)		
men	24		
women	5		
Democrat	(13)		
men	18		
women	9		
Independent	(8)		
men	12		
women	5		

summation: Very generally, and not in all instances, **younger, better educated, higher income** respondents tended to be more approving of or comfortable with AI than were older, less educated, lower income respondents. But the margins of difference were often small. In several surveys better educated respondents more frequently replied “Don’t Know.” Little variance was seen in responses from those identifying with different **religious traditions**, and likewise between religious and non-religious respondents. All roughly tracked population-wide responses. Ideological **liberals** were generally more supportive of AI than were ideological **moderates** or **conservatives**. However, little difference of opinion was seen among **Republicans, Democrats** and **Independents**.

Among the few large differences that appeared repeatedly and consistently were the different responses from **men and women**; often their responses would vary by ~ 25%. **Gender differences transcended partisan differences:** Republican, Democratic and Independent men agreed, by an average of 52%, that we *should* increase our reliance on AI, while Republican, Democratic and Independent women agreed that we *should not*, by an average of 28%.

G. TIME SERIES

Fast and Horvitz (2016) analyzed 30 years of *New York Times* news coverage of artificial intelligence to shed light on the questions: 1) how prominently has AI figured in public discussion over this period? 2) How has optimism or pessimism in news stories concerning AI varied over this period? 3) What ideas have been most commonly associated with AI over this period? And 4) How have public concerns about AI varied over this period?

They created a data file of all 3 million NYT articles published from 1/1/86 through 5/31/16, and identified those containing the phrases “artificial intelligence,” “AI” or “robot”. They used crowdsourcing to annotate this material according to protocols. They counted and otherwise analyzed the annotations to get these results:

1) They found that AI was covered at steadily *decreasing* rates between 1986 and 1995, corresponding to a period of “AI Winter.” Coverage revived in the late 1990s but then collapsed again sharply in 2005 (they don’t say why; the period of AI coverage revival continues through both the boom and crash periods of the 1998-2002 dot-com boom and crash). In 2009 coverage revived again, quite dramatically, and has continued to grow since. Fast and Horvitz say the reasons for the dramatic revival are unclear, but note that this period follows the “renaissance” of neural nets/deep learning applications, as well as the 2009 AAAI Asilomar AI conference organized by Horvitz and Selman.

2) Regarding optimism/pessimism, they find that over the full 30 years, news articles have been consistently more optimistic than pessimistic. They note, however, that the dramatic increase of coverage that began in 2009 has generated an increase in both optimistic *and* pessimistic accounts. [But see comment 2 below].

3) Regarding associated ideas, they prepare a long list of words often associated with AI, and count how often these associations occur over time. They find, for example, that AI is frequently associated with *space weapons* in 1986, *chess* in 1997, *search engines* in 2006, and *driverless vehicles* in 2016. They find that the association of AI with *science fiction* is frequent prior to 1990 but much less so after that.

4) Regarding public hopes and concerns, they prepare a long list of (positive) hopes for AI that the public might have, and a similar list of (negative) concerns, and proceed as in #3 above. Hopes that appear to have increased in recent years include *healthcare, education, and decision-making*. Concerns that appear to have increased include: *loss of control, ethical concern, and the impact on work*. They recognize here that the same outcome might be considered positive or negative by different people. They report that both hopes and concerns about the *Singularity* have increased, but on balance hopeful assessments have increased more; the same applies to *cyborgs*. Some hopes, including *entertainment and transportation*, don't appear to have increased or decreased.

Comments:

1) Fast and Horvitz use the content and tone of NYT stories as proxy for the broader “public perception” of AI, but this seems too large a leap. Fast and Horvitz acknowledge the problem, and perform a validity test: they replicate a portion of their main NYT results using texts from a different media source: posts made by Reddit users. I don't know enough to evaluate this. I do think the original exercise can be taken as an analysis of the content and tone of NYT stories, full stop, and is still interesting and useful.

2) Fast and Horvitz refer to an *attitude ratings scale* used by the crowdsource workers to rate a given text as optimistic or pessimistic, but they don't show the scale language itself. As we saw regarding the Morning Consult poll immediately above, optimism and pessimism can mean almost diametrically different things to different people. An outcome in which scientists successfully create a Superintelligent AI that takes control of the world and gives us all hyper-IQ brain implants might be considered a very optimistic scenario by some, but I suspect would be considered a dystopic, horrific outcome by most.

3) As in #2 above, it's difficult to evaluate these results without seeing the actual text used to guide the crowdsourcers in annotating each text.

III. CONCLUDING COMMENTS

[In preparation; see Section IX, p F.3-11, for compiled summary comments.]

ATTACHMENT S. NOTES ON FUNDERS AND FUNDING OF AI ADVOCACY

[in preparation]

- S.1. Funding sources associated with different sorts of AI organizations
- S.2. The Effective Altruism movement
 - S.2.1. Introduction
 - S.2.2. Representative Effective Altruism leaders, funders, infrastructure and critique
 - S.2.3. Critique of Effective Altruism
 - S.2.4. Other AI+ Funders, with no apparent connection with the EA movement.
 - S.2.5. Case Study: Good Ventures, GiveWell and The Open Philanthropy Project
- S.3. Mainstream Funders: The Ethics and Governance of Artificial Intelligence Fund

S.1. Funding sources associated with different sorts of AI organizations

1. Old-line AI; most university-based research organizations and professional societies. (eg, AAAI)
Mostly focused on academic/professional work, less focused on PR or advocacy. Small budgets.
FUNDING: Institutional support, membership dues, small established family foundations.
2. The core AI+ advocacy organizations (FHI, CSER, FLI, LCFI, CHAI, BERI, MIRI)
Appears to be quasi-academic research, but with great attention to network building, promotion, PR, political strategy and macrostrategy. Generally but not always small-to-medium annual budgets, i.e., \$10 million or less.
FUNDING: Institutional funds, newer/techie small family foundations, grants from wealthy AI+ ideologues and from the network of Effective Altruists.
3. OpenAI
This is the singular big play organization established in 2015 by Elon Musk and Sam Altman. Its goal is to develop and release open source Artificial General Intelligence. It's ambition, size, and impracticability suggest that it is at least partly a strategic move, e.g., to normalize the AGI project and to put those who argue against open source AGI on the defensive. It has hired an elite set of high-profile AI researchers, so it might very well generate some useful products, even if it never comes close to AGI itself.
FUNDING: \$1 billion in pledged funds from Musk, Altman and (presumably) others.
4. The Partnership for Artificial Intelligence to Benefit People and Society (PAI)
This is the intended centrist, mainstream, established collective voice of all things AI. It was established in 2016 by Google, IBM, Facebook, Apple, and Microsoft, but invited many AI organizations in as partners, including the 7 core AI+ organizations. It also invited in a significant number of social and economic justice NGOs, ethics institutes and other civil society groups.
FUNDING: NA; presumably several major donors plus some sort of partnership assessment.
5. The 3 progressive AI advocacy organizations:
 - a) Data & Society: Probably more left/progressive, at least on justice/inclusion concerns.
 - b) AI NOW – liberal/progressive, but appears to collaborate easily with the major tech firms and others.
 - c) UPTURN: smaller, left/progressive, seemingly more on-the-ground activist than the other two.**FUNDING:** the mainstream liberal foundations: Ford, Knight, Kellogg, MacArthur, Open Society, Sloan, Robert Wood Johnson, Gates, Microsoft, MIT Media Lab. AI NOW has also received funds from the **Ethics and Governance of Artificial Intelligence Fund**, which regrants to particular projects from a \$27 million fund created by Knight, Hewlett, the Omidyar Network, Reid Hoffman and Jim Pallotta.

S.2. The Effective Altruism Movement

S.2.1. Introduction

“Effective Altruism,” or “EA,” is a growing social and philanthropic movement. It has a base among young science and technology entrepreneurs and professionals in Silicon Valley and at top universities in the US and the UK. It encourages members to aspire towards having “the greatest positive impact possible” on human and planetary well-being. Members are invited to pledge 10% or more of their annual income to “effective” charitable programs. The homepage of the founding EA organization states that:

“[The Centre for Effective Altruism](#) helps to grow and maintain the effective altruism movement. Our vision is an optimal world. Our mission is to create a global community of people who have made helping others a core part of their lives, and who use evidence and scientific reasoning to figure out how to do so as effectively as possible.”

The mission statement from the Foundational Research Institute captures the language and tone of many of the effective altruism groups:

“[The Foundational Research Institute](#) conducts research on how to best reduce the suffering of sentient beings in the near and far future. We publish essays and academic articles, and advise individuals and policymakers. Our focus is on exploring effective, robust and cooperative strategies to avoid risks of dystopian futures. Our scope ranges from foundational questions about ethics, consciousness and game theory to policy implications for global cooperation or AI safety.”

The movement has philosophical roots in the writing of US philosopher Peter Singer. The first effective altruism organization was established in 2011 by Oxford University philosophy faculty members William MacAskill and Toby Ord, who were 24 and 32, respectively, at the time. Since 2013 an annual “EA Global” conference has served as a focus of Effective Altruism discussion.

There are scores of independent EA groups world-wide. Many are based on university campuses, and others have been set up by wealthy individuals. As of 2017 the Center for Effective Altruism counted ~ 3,600 members as having made the 10% or more pledge, and claim that this translates into about \$1.4 billion in charitable donations, given the expected earnings and expected life spans of these members.

In 2015 EA established a career-counseling service for young people interested in having the greatest impact they could on human and planetary well-being. Students were counseled to enter management consulting, venture capital investment and other highly remunerative careers, on the grounds that the more they earn, the farther their 10% charitable contribution pledge would go towards having the greatest impact possible. Students were discouraged from entering low-impact careers such as family medicine and teaching.

The EA movement has an increasingly close relationship with the AI+ advocacy movement. The Center for Effective Altruism shares office facilities at the Oxford Martin School with Nick Bostrom’s Future of Humanity Institute. Many of the staff at these two organizations have spent time in one capacity or another with both organizations (e.g., see B.3 below). It appears that at some point over the past 5-10 years the process of “rationally and scientifically” evaluating how to have the biggest positive impact on human and planetary well-being led the core EA leadership to conclude that one of the highest priorities would be to address global existential risk, and in particular the risk that a powerful AI might either accidentally or maliciously exterminate all human and other life forever. Further, they seem to have decided that the development of *beneficial* AI would be the single *most beneficial* thing that might *ever happen in human history*. As a result, EA funding and focus, guided by rational benefit/cost assessments, has begun moving heavily to the core AI+ advocacy groups.

What might be called the *transhumanist Imperative* appears to be built into the formal structure of EA’s evaluation and decision-making process. Over the course of extended lectures and seminars, participants are taught to base

their charitable giving on objective, quantitative, rationalistic and scientific methods and criteria. This quickly prioritizes preventing illness and death, and after that to enhancing the healthier and longer lives we lead as a consequence. The transhumanist argument that technology, including AI and biotechnology, will increasingly allow more people to live longer and more fully, perhaps awesomely so, seems reasonable, and quickly leads EAs to embrace the greatest possible promise of all: the end of disease, ageing, poverty and war, the enhancement of all cognitive, sensory and behavioral abilities beyond anything we can now imagine, and so on. At the last EA Global conference, held in 2017 in San Francisco, a large portion of the workshops addressed different aspects of existential risk, AI, and the technologically transformed human future.

A largely unexpressed but obvious foundational grounding of EA is the strong rejection of any religious or otherwise “subjective,” “emotional” or “sentimental” values as a guide to philanthropic giving. This appears to include values of “justice” that motivate much of both religious and secular charitable activity. EA initiatives are thus characterized by voluminous, almost obsessive, writing on how best to quantitatively, rationally and scientifically determine both the priority need for and success of charitable and other assistance.

In addition to support for AI+ advocacy groups, EA often gives special attention to the relief of *animal suffering*. This follows directly from Peter Singler’s characterization of higher animals as sentient beings.

As far as I can tell the great bulk of EA giving commendably goes to poor communities throughout the world, and the emphasis placed on effectiveness can certainly be applauded. There is every reason to believe that EA participants are sincere and committed. I haven’t seen a professional evaluation of the EA movement, and don’t have the background to offer a personal one. I do believe that the adoption by many in the EA movement of the belief that the human community can be transformed by artificial intelligence into a Post-Human entity that will usher in the Singularity and abolish all suffering forever, is an indictment of the secular/scientistic/analytic mentality, and potentially dangerous.

S.2.2. Representative Effective Altruism leaders, funders, infrastructure and critique

A quick click-through of EA websites gives a sense of the culture, tenor, focus and membership of the EA movement.

1. Philosophy professor [William MacAskill](#) co-founded the [Center for Effective Altruism](#) at Oxford in 2008. He is the author of *Doing Good Better: Effective Altruism and a Radical New Way to Make a Difference* (2015). He participated in both the 2015 and 2017 AI strategy conferences in Puerto Rico and Asilomar.
2. Oxford philosophy professor [Toby Ord](#) co-founded CEA with McAskill. His webpage states: “His current research is on avoiding the threat of human extinction and thus safeguarding a positive future for humanity, which he considers to be among the most pressing and neglected issues we face.” Ord likewise participated in both the 2015 and 2017 AI conferences.
3. This [Resources](#) page from the CEA website illustrates some of the EA priorities, including that of support for AI+ advocacy.
4. [Giving What We Can](#) – This initiative encourages participants to pledge 10% of their annual income to benefit humanity. It was established in 2009 by MacAskill as a CEA spin-off.
6. [Good Ventures](#) – This is the \$8 billion philanthropic initiative established by Facebook co-founder Dustin Moskovitz and his spouse Cari Tuna. It has made significant contributions to AI+ advocacy organizations. See Section B.3 below for details.
10. [GiveWell](#) - Describes itself as a “charity evaluator.” It uses rational and “evidence based” criteria to evaluate charitable needs and solutions, and advises funders on giving opportunities. GiveWell worked closely with

Moskovitz and Tuna in establishing Good Ventures, and in June 2017 GiveWell and GoodVentures spun off a separate organization, the Open Philosophy Project (see following).

5. [Open Philanthropy Project](#) – According to its website, the Open Philanthropy Project does not typically give grants, but it “advises” many grantors, including Good Ventures, on effective giving opportunities. See Section B.3 for details.

7. [Give Directly](#) – Established in 2012 by students at Harvard and MIT. It makes monthly unconditional cash transfers to poor people in Kenya, Uganda and Rwanda, via mobile phone accounts. It received a strong recommendation from GiveWell and has received funds from Good Ventures and Google. In 2015 it transferred ~ \$14 million to some 21,000 individuals. It is the largest direct cash transfer program in operation worldwide, and is being watched carefully by those involved in the debates over Universal Basic Income.

8. [80,000 hours](#) - This CEA spin-off identifies those careers that offer the greatest possible benefit to humanity, and provides career counselling. Funders include Open Philanthropy Project and Y Combinator. It uses a complex quantitative formula to determine today’s most urgent global issues, and the “most exciting jobs” for addressing them. As of 2018 the first and second most urgent global issues are “Risks from Artificial Intelligence” and “Promoting Effective Altruism.” Issues #7, 8 and 9, respectfully, are “Nuclear Security,” “Developing World Health” and “Climate Change.” The recommended list of exciting jobs shows 30 jobs: 14 working with AI organizations, 4 with Effective Altruism groups, 3 with groups fighting factory farming, and 9 others on a variety of issues, including nuclear weapons, global poverty, world health, etc.

9. A two-venue Global EA conference is held each year in London and San Francisco. The [2017 San Francisco Global EA conference](#) showed the growing prominence given AI+ advocacy. The 2018 San Francisco Global AI conference is set for June 8-10.

10. [Effective Altruism Funds](#) – EAF describes itself as a set of “mutual funds” for the EA community. It has separate funds covering four topic areas judged to be of high priority by effective altruists. These are 1) Global Development, 2) Animal Welfare, 3) Long-Term Future and 4) Effective Altruism Community. EAs particularly interested in one or another of these topic areas can donate to that fund; the managers of each fund will allocate the totals they have available to priority organizations addressing those topics. Over 2013-2017 the organizations receiving funds from the Long-Term Future fund were largely AI+ advocacy groups. EAF is administered by CEA, and the fund managers are employees of the Open Philanthropy Project.

11. [Effective Altruism Foundation](#) - Is a multi-purpose EA organization for the German-speaking community. Current projects include 1) *Philanthropy* - it promotes EA within the world poker community and currently has ~ 300 professional players as EA members; 2) *Community* - EAF speaks at conferences and to the press, supports 20 local EA groups, and provides coaching on career and giving decisions; 3) [Foundational Research Institute](#) – prepares academic papers, reports and blogs, currently focused on AI threats to humanity; 4) special research and advocacy on *Wild Animal Suffering* and on 5) *Sentience Policies*.

14. [Center for Applied Rationality](#) – is a Berkeley-based training center that promotes “rational thinking.” It is part of a “rationality” movement that in recent years appears to have effectively merged with the EA, AI and transhumanist movements. It offers training workshops in “applied rationality” at \$3,900/person. A 2016 *New York Times* story described the CFAR workshop experience as “[creepy, even cultish,](#)” although it did acknowledge some positive experience as well. CFAR has received support from the Open Philanthropy Project and the Future of Life Institute.

Several EA initiatives appear to have stayed removed from the move to fund AI+ advocacy. These include:

- | | |
|--|---------------------------------------|
| 15. Jasmine Social investments | 17. Mulago Foundation |
| 16. The Life You Save | 18. Charity Science |

S.2.3. Critique of Effective Altruism

1. [Review](#) of *The Most Good You Can Do* by Peter Singer and *Doing Good Better* by William MacAskill.
2. [The Elitist Philanthropy of So-Called Effective Altruism](#) – *Stanford Social Innovation Review*.
3. Here's a self-identified [Transhumanist/Immortalist/Effective Altruist](#), acknowledging doubts.
4. Critique of EA by [David Brooks](#), *New York Times*.
5. Critique of EA by [Jason Gabriel](#), *Boston Review*.
6. "Stop the Robot Apocalypse," by Amia Srinivasan, [London Review of Books, Sept. 2015](#)

S.2.4. Other IA+ Funders, with no apparent connection with the EA movement:

1. Elon Musk Foundation – The Foundation as such appears to be little more than a bank account from which Musk writes checks. Musk has contributed at least \$20 million to the Future of Life Institute and presumably a large sum to OpenAI.
2. Thiel Foundation – Peter Thiel has funded Eliezer Yudkowsky's **Machine Intelligence Research Institute** (MIRI), several of the **Singularity Summits**, Aubrey de Gray's **Methusela Project** on immortality, the hyper-libertarian **Seasteading Institute**, and allied efforts.
3. Global Challenges Foundation – Established in 2012 with \$60 million from Hungarian billionaire [László Szombatfalvy](#), The Foundation works closely with Bostrom's Future for Humanity Institute at Oxford.
4. Ethereum - This is the platform that supports the cryptocurrency *Ether*. It was developed by Vitalik Buterin, a Russian-Canadian student who received a Thiel Fellowship to drop out of college and complete the project. Buterin participated in the 2017 AI conference at Asilomar.
- 5-11. Other funders who have given to one or more of the AI+ advocacy groups include the Hauser-Raspe Foundation, Kenneth Miller Trust, Libra Foundation, Milner Foundation, Blavatnik Family Foundation, Grantham Foundation and Templeton World Charity Foundation. Most of these are small family foundations and have given mostly small grants.

S.2.5. A Case Study of Effective Altruism: Good Ventures, GiveWell and The Open Philanthropy Project

Facebook co-founder Dustin Moskovitz and his wife Cari Tuna have a net worth of \$14 billion, and have pledged to be effective altruists and to give away much of their wealth. In 2011 they established [Good Ventures](#) as their charitable vehicle. They worked closely with [GiveWell](#) for advice and assistance on giving opportunities. In June 2017 Good Ventures and GiveWell spun off a separate organization, the [Open Philanthropy Project](#). The OPP website says that it does not typically make grants, but rather advises grantors, including Good Ventures, about effective giving opportunities. As of 2017 Open Philanthropy had a staff of 24. After a lengthy evaluation process, the Open Philanthropy Project decided on these long-term priorities:

1. US POLICY:

- a. Criminal Justice Reform; b. Farm Animal Welfare; c. Macroeconomic Stabilization Policy; d. Immigration Policy; e. Land Use Reform

2. GLOBAL CATASTROPHIC RISKS

- a. Biosecurity and Pandemic Preparedness
- b. Potential Risks from Advanced Artificial Intelligence

3. SCIENTIFIC RESEARCH [Focus areas still to be decided]

4. GLOBAL HEALTH & DEVELOPMENT [Focus areas still to be decided]

Grants made by Good Ventures in 2016-2017 in category 2b (Potential Risks from Advanced AI) were made to:

6 of the 8 core AI+ advocacy groups:

Machine Intelligence Research Institute	\$ 4,250,000
Berkeley Existential Risk Institute	404,000
OPEN AI	30,000,000
Future of Humanity Institute	1,995,000
Center for Human-Compatible AI	5,555,550
Future of Life Institute	1,186,000

Other organizations:

Berkeley AI Safety Research	\$ 1,450,000
Electronic Frontier Foundation	199,000
Monterey Institute	2,400,000
3 small grants to individuals	82,000

It's worth noting the backgrounds of several of the Open Philanthropy Project's lead staff:

Daniel Dewey, OPP Program Officer, Potential Risks from Advanced Artificial Intelligence. Carnegie Mellon Computer Science and Philosophy BS. Has worked at **Google**, the **Future of Humanity Institute**, the **Future of Life Institute** and the **Centre for Effective Altruism**.

Nick Beckstead, OPP Program Officer, Scientific Research Funding. Rutgers Philosophy PhD, former Research Fellow at the **Future for Humanity Institute**.

Chris Sommerville, OPP Scientific Advisor. "He is identifying and evaluating potential philanthropic opportunities in scientific research." Former professor at UC Berkeley and Stanford University. Member, US National Academy of Sciences and the Royal Society of London. **PhD in genetics**.

Heather Youngs, OPP Scientific Advisor. She is identifying and evaluating potential philanthropic opportunities in "transformative basic and applied research that have important societal impact." Ph.D. in biochemistry and molecular biology from **Oregon Health and Science University [primate cloning and other controversial efforts]**.

Luke Muehlhauser, OPP Research Analyst; former Executive Director, **Machine Intelligence Research Institute**.

S.3. MAINSTREAM FUNDERS: THE ETHICS AND GOVERNANCE OF ARTIFICIAL INTELLIGENCE (EGAI) FUND

THE EGAI Fund:

* supports work around the world to advance ethical artificial intelligence in the public interest.

* was launched in **January 2017**, with an initial investment of **\$27 million**.

* Founding contributors:

John S. and James L. Knight Foundation

Omidyar Network

Hewlett Foundation

Reid Hoffman (founder of Linked In)

Jim Pallotta (founder of Raptor Group)

* The Miami Foundation is fiscal sponsor.

It's first nine grants (2017) went to:

Leverhulme Centre for the Future of Intelligence (Cambridge, UK); Data & Society Research Institute (NYC) AI Now (NYC); Berkman Klein Center for Internet & Society at Harvard University; MIT Media Lab Digital Asia Hub (Hong Kong); ITS Rio (Rio de Janeiro, Brazil); Access Now (Brussels, Belgium); FAT ML (Global).

NB: The great bulk of recent controversy focused on AI has arisen in the United States and the United Kingdom. Attention has also been raised in other English-speaking countries, and in Scandinavia and Germany. I know anecdotally of controversies in other parts of the world but haven't done a formal search for these, although I would like to do so soon. Four of the projects funded by the EGAI Fund, as shown above, are from outside the centers of AI controversy and merit brief description here pending further work:

1. Digital Asia Hub (Hong Kong) - Is an "independent and interdisciplinary research exploring both the opportunities and challenges related to digital technology, innovation, and society in Asia". Topics it will address include: Digital access, privacy and other rights; Governance and infrastructure; Innovation, open manufacturing and digital trade; Trending technologies (e.g. Smart Contracts, Big Data, Internet of Things); and the Impacts of spreading mobile technologies. It was established with support from the Berkman Klein Center at Harvard.

2. Institute for Technology and Society of Rio de Janeiro (ITS Rio)- Its Mission "is to ensure that Brazil and the Global South respond creatively and appropriately to the opportunities provided by technology in the digital age, and that the potential benefits are broadly shared across society." It appears to be a mainstream/liberal civil society/academic center concerned that the growing tech sector promote equality, justice, inclusion and the like.

3. Access Now (Brussels office) - Access Now is "an international non-profit, human rights, public policy, and advocacy group" set up "to defend and extend the digital rights of users at risk around the world". It has 40 staff located in 11 offices, in Berlin, Brussels, Cordoba, Delhi, London, Manila, Nairobi, New York, San Jose, Tunis, and Washington D.C. It was started by Brett Solomon and Cameran Ashraf following the contested 2009 presidential election in Iran, and played a role in disseminating video of the turmoil in Iran. Solomon remains CEO.

4. Fairness, Accountability and Transparency in Machine Learning (FAT/ML) – This is an annual conference of AI developers, academics and others focused on challenging societal questions related to their work. It appears to be solidly center/mainstream, with good global participation. None of the AI+ advocacy groups or activists appear to be involved. The 2018 FAT/ML was held Feb. 23-24 at NYU. Funding comes from EGAI, Facebook, Google, MacArthur and H2O.ai.

ATTACHMENT T. THE AI ADVOCACY MOVEMENT AND THE HUMAN TECHNO-EUGENICS MOVEMENT

Many advocates of the AI+ transformation of the human species likewise advocate our transformation via genetic and other biotechnologies. These cross-disciplinary advocates include scientists, academics, major funders, journalists and leaders in both the infotech and biotech industries.

These integrated relationships are of particular value with respect to the techno-eugenics movement. For all the attention given over the past two decades to the possibility that new techno-eugenic practices might become acceptable and widespread, a true mass movement for this has not yet crystalized. The key players remain a fairly well-delineated set of researchers, academics, popularizers, government oversight bodies, biotech firms and fertility clinics. By contrast, the AI+ advocacy movement, as outlined in these notes, has a robust and highly motivated infrastructure in place. Importantly, it's well-represented within the mainstream Silicon Valley and infotech sectors as well. If, at any time, some set of developments took place that made it advantageous for either the AI+ advocacy network, or key sectors of the mainstream AI network, to move portions of their attention and resources into the techno-eugenics space, they could do so relatively swiftly and smoothly. The scenarios have been written, discussed, disseminated and rehearsed. We could see the emergence of a full-blown techno-eugenics movement, with facts on the ground, within a period of perhaps 2-3 years.¹⁹

The bulleted items that follow show the many ways in which the AI and other infotech advocates discussed in these notes have also been active in or otherwise associated with the advocacy and promotion of human genetic modification and related biotechnologies. These items are snapshots of a far larger network of relationships; a full analysis is yet to be done.

- The controversial personal gene-testing company [23&me](#) was co-founded in 2006 by Anne Wojcicki, then-wife of Google co-founder Sergey Brin. In 2007 **Google** invested \$3,900,000 in 23&Me; as of 2018 the firm is valued at \$1.5 billion. Anne Wojcicki continues as CEO.
- In 2013 **Google** established a “super-secretive anti-aging start-up” called [California Life Company](#), or Calico, on the initiative of **Google cofounder and CEO Larry Page**, and staked with \$1.5 billion from Alphabet/Google and the drug firm AbbVie. In 2017 Calico had 100 employees, including an elite team of anti-aging and AI researchers hired away from major US and other universities. The former head of **Google Ventures**, Bill Mavis, who was instrumental in establishing Calico, anticipates human life spans of 500 years.
- In 2016 **Facebook cofounder and CEO Mark Zuckerberg** and his wife Priscilla Chan announced a \$3 billion initiative to “[cure all diseases](#)” by the end of the century. The project is now headed by neuro-geneticist Dr. Cori Bargmann, whose immediately prior research at Rockefeller University involved “how neurons and genes impact behavior.”
- **Silicon Valley tech investor** and **Facebook board member Peter Thiel** (net worth: \$2.7 billion) has announced his intention to live forever, has given \$3.5 million to the **Methusela Foundtion** to “cure aging,” has contracted with the **Alcor Foundation** to have his body cryogenically frozen if and when he dies (pending reanimation), and has given lavishly to **Singularity University**, **Machine Intelligence Research Institute** and **DeepMind** (now owned by **Google**). His investment firm, Founders Fund, specializes in biotech, nanotech, AI and robotics.
- Amazon founder **Jeff Bezos** is a founding investor in [Unity Biotechnology](#), focused on developing anti-aging pharmaceuticals.
- In the early 2000s [Oracle CEO Larry Ellison](#) donated \$335 million to scientists studying aging and longevity.

- In 2014 a consortium of Silicon Valley tech investors established the [Palo Alto Longevity Prize](#), described on its website as “... a \$1 million life science competition dedicated to ending aging”.
- Harvard stem cell and CRISPR scientist [George Church](#) is an ambitious proponent of human genetic modification and other forms of human biological enhancement. He recently [announced](#) formation of a for-profit genome sequencing start-up, [Nebula Genomics](#), involving the controversial bitcoin and blockchain online technologies. Nebula will allow individuals to get their full genome sequenced, store it securely online, and sell it to biotech firms, in exchange for Nebula Tokens, which they can then convert to Ether and then to cash. See also [Philippidis \(2018\)](#).
- In Sept. 2001 [Stephen Hawking](#) was quoted in a German magazine as saying that humans will have to modify their DNA to stop artificially intelligent robots from “taking over the world”. He later claimed to have been [misquoted](#); he said his position is that human genetic engineering would be *too slow* to allow us to keep up with AI robots, and that we will need to develop cyborgian *brain/computer interface technologies* if we want to avoid a robot take-over.
- Past President of the UK Royal Academy [Lord Martin Rees](#) expresses concern about eugenics and DIY biotech, but [ultimately embraces the full techno-eugenic future](#), saying that humans “...will use all the techniques of genetic modification, cyborg techniques, maybe even linking or downloading themselves into machines, which, fifty years from now, will be far more powerful than they are today. The posthuman era is probably not going to start here on Earth; it will be spearheaded by these communities on Mars.”
- [Nick Bostrom](#) of Oxford’s Future of Humanity Institute is a leader in the [Transhumanist movement](#), which advocates the wholesale enhancement and reconstitution of the human species using both genetic and other biotechnologies and artificial intelligence. As noted, many of those working in the AI+ advocacy movement have transhumanist roots.
- Five of the eight key AI+ advocacy groups listed in Section II.A of this memo have as their mandate the study of “existential risk,” including risks involving AI and biotechnology. These organizations are Bostrom’s **Future of Humanity Institute**, the **Center for the Study of Existential Risk** at Cambridge, the **Future of Life Institute** in Boston, the **Leverhulme Centre for the Future of Intelligence** at Oxford, and the **Berkeley Existential Risk Institute**. Papers, reports and other output of these centers typically acknowledge the risks posed by misuse of biotechnology, e.g., as bioweapons used by terrorists, and call for appropriate regulation, but refrain from proposing that dangerous genetic and related biotechnologies be foregone. Rather, they treat the development and widespread use of genetic modification, human and otherwise, as a given. Bostrom and some others go further and add as an existential risk the danger that human genetic modification is *not developed rapidly enough*, or that it may even be *prohibited*. The reasoning is that humanity is facing so many threats – from asteroids, overpopulation, climate change, hunger, pandemics, nuclear war and more – that we are, on balance, likely to go extinct *unless* we can genetically engineer ourselves and our ecosystems to allow us to survive in the face of such multifold natural and human-caused calamities.
- The established bioethics profession that has done so much over the past three decades to normalize and prepare the way for human genetic modification is now bringing its services to bear regarding artificial intelligence. In 2016 **The Hastings Center**, the foundational US bioethics organization, launched an initiative on [“Control and Responsible Innovation in the Development of Autonomous Machines,”](#) involving three expert convenings and a series of reports. The initiative was funded by the Future of Life Institute, which advocates the transformation of the human species through artificial intelligence and germline genetic modification.

- It's worth noting that [eugenics is in Silicon Valley's DNA](#), so to speak. [David Star Jordan](#), a biologist and outspoken eugenicist, served from 1891-1913 as Stanford University's first president, advocating forcefully for mass sterilization. In 1928 he played a key role in launching the Human Betterment Foundation to promote eugenics education and policy. See Johnsson (2016). Stanford psychology professor [Lewis Terman](#) developed the Stanford-Binet IQ test in 1916, and throughout his career sought to establish links between genetics and exceptional cognitive skill. His son, **Fred Terman**, a professor of electrical engineering at Stanford, served after WW2 as dean of the School of Engineering and then as University provost. He mentored many students who went on to become Silicon Valley notables, including David Packard and William Hewlett. He shared his father's eugenic beliefs, and did much to foster a culture at Stanford and in Silicon Valley that put an exceptionally high premium on analytic skills, entrepreneurial drive and commercial success. Most notorious of Terman's colleagues was [William F. Shockley](#), who in 1956 had received the Nobel prize for his work developing the transistor. He was professor of electrical engineering at Stanford from 1963-1975, but spent much of his time as a faculty member and afterwards publicly seeking to demonstrate the genetic inferiority of persons of non-Caucasian descent. This long heritage of sharp focus on cognitive and analytic skill, entrepreneurial drive, commercial success, social Darwinism, eugenics and racism strongly shaped the Silicon Valley culture of today.

NB: Two seemingly anomalous communications

1. Elon Musk

[In an interview](#) Musk was asked why he *wasn't* working on the important topic of human genetic engineering. He replied that there was an established consensus among geneticists against it. [this is not so, unfortunately – rh] On continued query from the journalist he said he was concerned about “the Hitler problem.” On further query – the journalist was quite unhappy with Musk's responses – Musk appeared to relent: “...in order to fundamentally solve a lot of these issues, we are going to have to reprogram our DNA. That's the only way to do it.”

I mention this only because it is in fact the case that a great number of Silicon Valley I-tech moguls have a human genetics project of some sort underway, just as they do a spaceship project. Musk's attraction to the spotlight would not be inconsistent with a high-profile human genetics initiative. But in the exchange with the journalist above he sounded distinctly uncomfortable talking about human genetics. And he is idiosyncratic. If there was any chance that Musk turned out to be willing to buck the tide and oppose some important aspect of human genetic modification, that could be significant.

2. Nick Bostrom

In *Superintelligence* (2014) Bostrom goes to some length explaining why he no longer believes that human genetic modification is a priority for transhumanists. He describes how difficult it would be to genetically modify humans to have truly enhanced brainpower, and how long it would take to do so. With humans you could only go so far with any single or small group of trait modifications, and would then need time to see what their effects on cognitive skill, etc. actually were. He contrasts human genetic modification with building an intelligent AI with silicon circuits and software, and concludes that the latter could be accomplished much more rapidly to far greater effect. And speed is important because we need to reach human-level AI, and then Superintelligence and the Singularity, within the next 20-30 years at the latest. Once we have Superintelligence we'll have the Singularity within a matter of weeks, if not days or hours. After that, the Superintelligence will be in a position to quickly figure out how to make us all be immortal, omniscient and omnipotent, immediately.

[More in preparation – RH]

ATTACHMENT U. THE SOCIO-CULTURAL-PSYCHOLOGICAL DIMENSIONS OF AI

U.1. Representative Texts

These seven mostly recent books focus critically on the impact that the internet, the web, social media, the iPhone and related wired technologies, including AI are having on our personal, family and community lives.

2011. Alone Together: Why We Expect More from Technology and Less from Each Other.	Sherry Turkle
2011. The Shallows: What the Internet Is Doing to Our Brains.	Nicholas Carr
2013. To Save Everything, Click Here. The Folly of Technological Solutionism.	Evgeny Morozov
2016. Disconnected: How To Reconnect Our Digitally Distracted Kids.	Thomas J Kersting
2017. Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked.	Adam Alter
2017. Glow Kids: How Screen Addiction Is Hijacking Our Kids - and How to Break the Trance.	Nicholas Kardara
2017. iGen: Why Today's Super-Connected Kids Are Growing Up Less Rebellious, More Tolerant, Less Happy--and Completely Unprepared for Adulthood--and What That Means for the Rest of Us.	Jean Twenge

There is a large journalistic and scholarly literature on the socio-cultural-psychological impacts of artificial intelligence, info-tech and computers in general. Much is current but some dates from the inception of info-tech in the first half of the last century and further back as well to the beginnings of the industrial revolution and before.²⁰ For now I'll defer comment on this large literature.

U.2. Socio-cultural-psychological concerns raised by the spread of information/internet/social media technologies, including AI technologies.

I'm using "I-tech" to mark the full range of information-related technologies, including all web- and Internet-based activities, all forms of social media, cell phones, tablets and other mobile devices, texting, blogging, video-blogging, apps, use of Big Data, IoT, the beginnings of the network economy - Airbnb, Uber - and everything else. I presume that over the coming years AI will play some greater role in most if not all of these and their successors.

I show possible socio-cultural-psychological concerns raised by I-tech and by AI in particular that various authors have raised. I don't assume that all concerns are necessarily warranted, although I've tried to include only credible concerns.

It's appropriate to ask if these concerns, taken as a whole, suggest there is something fundamental about the ways in which I-tech/AI/robotics might impact our socio-cultural-psychological lives that needs to be at a minimum better understood, very soon.

1. Increased, habitual or addictive use of I-tech, especially among pre-teens, teens and young adults, causing:
 - lasting impairment of attention, concentration, memory, learning and related cognitive abilities
 - inability to easily converse or otherwise interact with peers and others
 - increased frequency and intensity of depression, anxiety, mood swings and related disorders
 - increased feelings of inadequacy and low self-regard resulting from constant online comparisons with others.
 - less time spent physically socializing and recreating with others, and more time spent interacting via I-tech.
 - loss of a sense of a true self. Can the constant curating of one's online imaginary, presented self, along with other "like"-maximizing curation, prevent a true self from ever really developing?
2. Degradation of societal norms and conventions regarding civility, respect and maturity.
3. Continued and increasing risk of harm from online stalking, identity theft, invasion of privacy and other criminal activity.

4. Family life is being negatively impacted by the spread of I-tech. Parents as well as children are spending more time in front of screens and less time with each other. Although I-tech also provides benefits, the net impact on family life is most likely negative.
5. Degradation of the overall culture as the proliferation of exploitative, coarse, mocking, inappropriate, trivializing and extreme I-tech content increases at the expense of quality content.
6. Degradation of the democratic political process, as a consequence of increased polarization caused by a) the prevalence of filter bubble features across all I-tech domains, and b) the wider exposure that I-tech affords political extremes.
7. Serious interference with the democratic political process as bad actors, whether individuals, groups or nations, directly seek to disrupt political decision-making, including elections, through fake and misleading news, candidate communications, and other forms of mass I-tech manipulation.
8. Loss of a general, secure *sense of reality* as AI enables creation of altered or fully fabricated images, videos, news, events and all other sorts of I-tech communication, that are indistinguishable from real ones.
9. Individual and mass risks to health and life as AI contributes to the development and deployment of DIY genetic modification of microbes, plants and animals; and through 3d-printing of weapons and other lethal devices.
10. Despite recent pledges to reinforce communalist norms on-line (e.g., the January 2017 Zuckerberg “Communitarian Manifesto”), I-tech inherently reinforces individualism, narcissism and separation, along with a relativistic moral perspective.
11. Potentially massive socio-cultural-psychological disruption if:
 - a) AI turns out to strongly reduce the need for large sectors of the workforce.
 - b) Global communications, marketing, commerce and other essential societal functions continue to be increasingly controlled by a handful of mega-scale I-tech/AI firms and individuals.
 - c) AI and I-tech continue to reinforce prejudicial and discriminatory attitudes and behaviors, across all social domains.
12. Although I-tech and AI can be expected to contribute to an increasingly efficient world energy grid, historically such efficiency improvements have led to greater, not less, total energy use (The “Jevons paradox”).
13. The growing intensity and realism of violent I-games, and the greater time spent with them during childhood, teen years and young adulthood, could have greater longer-term effects on personality, affect and mental and emotional resources, even if a one-to-one correlation of violent imagery with violent behavior is not seen.
14. “Care robots” developed to assist the elderly and patients in hospitals and other sites might seem to be a practical and benign expedient in resource-challenged situations, but their use could lead to a gradual devaluation of the care receivers, the role of human care-providers, and the caring relationship itself.
15. In her book *Alone Together* (2011), MIT professor Sherry Turkle is concerned about the *readiness and willingness* with which many people turn away from human relationships in favor of on-line and robotic ones. What does this say about the ways in which our society socializes us, or predisposes us, regarding relationships in the first place? And should we continue to encourage this?
16. Turkle is also concerned about very young children and robots. She notes that these children form attachments with reasonably anthropomorphic robots that they *don’t* form even with their favorite inanimate dolls or stuffed bears; and when the robots break or don’t respond appropriately, the child “goes ballistic.” Turkle is sure that “build more reliable and smarter robots” is *not* the right response. What is?

17. Many leaders and participants in the AI+ advocacy movement are strong supporters of genetic enhancement of the human species, e.g., for longer life spans or greater cognitive abilities, and see this and AI as part of a single mission to “make the world more awesome.” With economic inequality soaring, and with popular disillusionment with democratic institutions growing, the last think we need is to reintroduce a high-tech form of eugenics onto the world stage.
18. The vision of the future increasing being promoted by AI/I-tech leaders is one in which the only realistic choice for humanity is either that we create a Superintelligence that will not only end poverty, war and climate change, but will make us all immortal and omniscient and allow us to fly to the stars; or that we regress to the early medieval dark ages, and most likely decline further to extinction. This is a false and dangerous dichotomy. To the extent that it spreads, it will impede the real work we need to do to learn how to live together in peace on this planet.
19. The World Health Organization (WHO) 2018 International Classification of Diseases (ICD-11) has added [Gaming Disorder](#) as a recognized disease, and the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) for 2013 added Internet Gaming Addiction as a “Condition for Further Study.”
20. Countries are installing video surveillance cameras, connected to facial recognition AI, to track people and monitor their behavior. China (2016 population 1.3 billion) plans to install 626 million surveillance cameras by 2020, or ~ 1 camera for every 2 persons. In some countries such intense surveillance is not experienced as inappropriate, but in others it would be considered extremely so, and resisted.
21. In his book *The Shallows* (2010) Nicholas Carr writes, “What’s disturbing about [Google’s founders] is not their boyish desire to create an amazingly cool machine that will be able to outthink its creators, but the pinched conception of the human mind that gives rise to such a desire.” [p 176]. To expand on that: If the pinched, degraded and simply incorrect understanding of the human mind as held by so many AI+ advocates and others comes to be adopted by very large numbers of people, real harm will have been done, to individuals and to society as a whole, and our work to build a just, flourishing world will have become more difficult.
22. In his book *Irresistible* (2017), dealing with addictive technologies, Adam Alter noted the many tech leaders who strongly limit the use of screens and other tech devices by their own children, including:
 - * Steve Jobs (Apple) – didn’t let his children use iPad.
 - * Chris Anderson (Wired) – strict limits on the use of all technology.
 - * Evan Williams (Blogger, Twitter) – had his children read print books, prohibited use of iPad.
 - * Leslie Gold (data analysis) – “no screen time during the week” rule.Alter goes on to report tech workers, web developers and others who acknowledge the clinically addicting nature of the devices and sites they create. Many limit their own and their children’s use while others do not.
23. The European Parliament has been considering the pros and cons of [granting personhood to Robots and AIs](#): “...in a 17-2 vote, the parliament’s legal affairs committee voted to begin drafting... preliminary guidance on what it calls ‘electronic personhood’ that would ensure corresponding rights and obligations for the most sophisticated AI.” See [Markou \(2017\)](#). The move elicited considerable push-back; see [Delcker \(2018\)](#).

#

ATTACHMENT V. PUSH-BACK ON THE FEASIBILITY OF AI+ SUPERINTELLIGENCE

V.1. GENERAL COMMENT

Most scientists and knowledgeable others, including most of those working on AI, disagree with Bostrom, Musk, Hawking and their AI+ advocate colleagues about the nature, prospects and risks of AI.

Many simply reject the idea that machines can be intelligent in the sense that most people have long understood the word. On this view intelligence requires, at a minimum, *consciousness*, and consciousness is a property of animal life, not machinery.²¹

Many others, probably including most scientists and engineers working on AI, are comfortable saying that machines can be intelligent, but on inspection it's seen that they're using *intelligence* (or *intelligent thought*) in a sense closer to what most others would call *computation*. All that a computation requires is input, a set of rules for manipulating the input (an *algorithm*), and output. At no point is there any need to invoke consciousness, conscious calculation, deliberation, agency or volition for a machine to successfully compute.

There's no reason that machines using creative algorithms can't produce outputs that mimic human intelligent thought and behavior across many output domains. But there's no reason to believe that these machines are any more conscious, aware or volitional than is a pocket calculator or a doorknob.

Of course, many scientists and engineers also believe (as I do) that even mundane, mechanical AI has the potential to be enormously consequential, and can be used both beneficially and harmfully. Scientists differ among one another, often strongly, concerning the likelihood of various sorts and degrees of potential benefit and harm, and they differ concerning the ease or difficulty of promoting beneficial uses and avoiding harmful ones. There's a strong case to be made for precaution when dealing with the impacts and uncertainties of AI. But very few scientists claim to believe that an important thing to worry about today is the possibility of an AI suddenly waking up and deciding to destroy or enslave humanity.

V.2. AI SKEPTICS

Noted UC Berkeley professor of philosophy Hubert Dreyfus was among the earliest critics of Strong AI. See "[Alchemy and Artificial Intelligence](#) (1965), *What Computers Can't Do* (1972) and "What Computers Still Can't Do" (1975).

Some high-profile AI researchers have gone on the record about their doubts concerning the basic feasibility of Strong AI or AI+ Superintelligence. These include Gordon Moore (of "Moore's Law" renown), Baidu VP Andrew Ng, and physicist/author Douglas Hofstadter.

John Brockman's 2015 edition of his popular Edge.com series asked selected scientists and others for their response to the statement: *What to Think About Machines That Think*. A comfortable majority of the 148 responders were aligned with the techno-progressive or AI+ advocacy community, and responded with predictably enthusiastic anticipation. But a fair number pushed back, challenging the notion that machines can think in the first place. These AI skeptics included:

- * Yale computer scientist [David Gelerntner](#): "Computers will never think... [They] can imitate important aspects of *thinking-about*, narrowly understood, but *being* is beyond them. Therefore mindfulness is beyond them."
- * Historian [Noga Arikha](#): "...until we replicate the embodied emotional being—a feat I don't believe we can achieve—our machines will continue to serve as occasional analogies for thought, and to evolve according to our needs."
- * Software developer [Kai Kraus](#): "I concede to AI proponents all the semantic prowess of Shakespeare, the symbol juggling they do perfectly. Missing is the direct relationship with the ideas the symbols represent."

- * Cognitive scientist [Roger Schank](#): “Machines can’t think.... The idea that we need to... worry about them...or grant them civil rights is just plain silly.”
- * Tech editor [S. Abbas Raza](#): “I don’t think anything less than a fully Darwinian process of evolution can give [teleological autonomy] to any creature.”
- * USC engineering professor [Bart Kosko](#): “Machines don’t think. They turn inputs into outputs... Tomorrow’s thinking machines will look a lot like today’s – old algorithms running on faster computers.”
- * University of Calgary physics & biology professor [Stuart Kaufman](#): “...consciousness and will may be part of [the universe’s] furniture, and Turing machines cannot, as subsets of classical physics and merely syntactic, make choices where the present could have been different.”
- * University of Sydney linguistic professor [Nick Entfield](#)
- * University of Texas paleontology professor [Julia Clarke](#)
- * NYU psychology professor [Gary Marcus](#)
- * MIT robotics professor [Rodney Brooks](#)

Virtual Reality innovator [Jaron Lanier](#) (2014) argues that the whole concept of artificial intelligent is "an illusion" and a "stupendous con" perpetrated by the wealthy.

Former *Wired* editor Kevin Kelly, in “[The Myth of a Strong AI](#)” (2017) challenges a litany of dogmatic Superintelligence beliefs.

Duke University neurobiologist [Miguel Nicolelis](#) says that machines can’t replace human brains because “Our brains do not work in an algorithmic way and are not digital machines.” Regarding the AI-takeover warnings by Musk and others, Nicolelis says, “It used to be annoying to see these kinds of statements, but now it’s becoming serious. It’s leading to mass hysteria.”

Neurobiologist Bobby Azarian (2016) argues in “[The Myth of the Sentient Machine](#)” that AGI is impossible because a) digital computers simply manipulate patterns of symbols (0s and 1s), i.e. **syntax**; brains do that too, but in addition process meaning, i.e. **semantics**; b) human cognition requires consciousness, which is a biological phenomenon with orders of magnitude more parts, interactions, feedbacks, analogue operations, etc. that can’t be reduced to digital representation; c) simulation is not duplication. You can use a computer to *simulate* photosynthesis; the outputs are patterns of 0s and 1s. And there are in fact machines that *duplicate* photosynthesis: they use light energy to turn water and CO2 into O2 and organic molecules. The second gives you output you can breathe and eat; the first does not.

Popular Science author Erik Sofge decries the [exaggerated claims](#) regarding the Singularity, AI and robotics.

[Carlos Moedas](#), [European Commissioner for Research, Innovation and Science](#), says that the "media are too full of alarmist, hysterical doomsday scenarios.”

[Rodney Brooks](#), director of MIT’s Computer Science and Artificial Intelligence Lab, noted that none of the noted AI alarmists, including Elon Musk, have backgrounds in AI.

Atlantic magazine has run a series of insightful articles on AI. See:

- * James Hamblin’s “[But What Does the End of Humanity Mean For Me?](#)” (2014)
- * Erik Larson’s “[Questioning the Hype on Artificial intelligence](#)” (2015)
- * Jim Bogart’s “[Artificial Intelligence has become meaningless](#)” (2017)

See also Peter Kassan’s “[AI Gone Awry: The Futile Quest for Artificial Intelligence](#)” (2006), and NYU professor of psychology Gary Marcus in *The New Yorker* “[Hyping Artificial Intelligence, Once Again](#)” (2013).

V.3. A CONCEPTUAL / SEMANTIC BLUNDER?

Stanford University AI/computer instructor Jeffrey Kaplan (2016) wonders whether AI is in fact a real science or if it “is simply the Lady Gaga of computer science – performing numbers swaddled in gaudy, anthropomorphic costumes, capturing the popular imagination and lots of financial support, a carnny sideshow prone to occasional hucksterism and hubris, a parlor trick.” (p 34).

Kaplan concedes that the practice of AI does have scientific elements, but then throws cold water on its mystique. AI systems “are simply carrying out logical deterministic sequences of action, no matter how complex, changing their internal configurations from one state to another.” He concludes that intelligence is a property of living creatures: “When it comes to intelligence, as far as machines are concerned, there’s nobody home.”

Kaplan speculates that if *artificial intelligence* had been given a different and more accurately descriptive name, everything we now call AI might be considered “just another field of computer science,” like, say, operations research or software engineering. Alternative names that he and others have suggested include ***symbolic processing, augmented intelligence, statistical automation, analytic computing, anthropic computing, autonomic computing, deep computing*** and ***cognitive computing***. Somewhat sarcastic suggestions include ***glorified pattern-fitting*** and ***glorified curve fitting***.

Given the dramatic coverage that AI has received over the past half-decade, an across-the-board name change might seem unlikely. However, as noted, AI is not a coherent single discipline but rather a grab bag of somewhat disparate tools, techniques and procedures. Institutions that specialize in one or another of these might choose to use the more appropriate nomenclature and at least begin to promulgate more useful naming. A number of major firms and university departments, for example, eschew *artificial intelligence* in favor of *machine learning*.

V.4. FEAR OF AN AI WINTER?

“AI winter” refers most commonly to periods over the past half century or so during which research, development and funding of artificial intelligence slowed dramatically. The “winters” were usually immediately preceded by periods of accelerating AI excitement.

Two major AI winters are those of 1974-1980 and 1987-1993.

In a typical AI winter cycle, new theoretical approaches and techniques lead to some interesting new capabilities, which in turn spark speculation of further development, including commercial development. Venture capitalists and other investors are eager to get in early on the potentially next big thing. If continued research looks promising, at some point DARPA becomes interested and large new funding streams become available. By this point the press is talking about the spectacular, transformational future soon to be realized through new artificial intelligence. After a period of exaggerated claims, with only moderate and thus disappointing results to show, investors and DARPA begin to exit the field and winter ensues.

Some authors warn that we may be facing a new AI winter soon. [Kinsella \(2017\)](#) suggests that “virtual assistants, deep learning, machine learning and cognitive computing are at their ‘peak of inflated expectations.’” Other analysts believe that a new AI winter is unlikely, given advances in computer power, the growth of the Internet of Things (IoT), the profusion of data streams and archives, independent major AI R&D by the major I-tech firms, and continued development of collaborative arrangements between universities and the private sector. On this view, the increased scale of the AI sector will generate a greater number of useful new discoveries, techniques, innovations and products, which together should promote and sustain both supply and demand. See [eMarketer \(2017\)](#) for more.

[Ambasna \(2018\)](#) and [Chollet \(2017\)](#) differentiate between the prospects for *mainstream AI* and *General AI*. They say that the former, which includes everything now underway or planned for the near term, is robust enough, for

all the reasons just listed above, to avoid a new AI Winter. But they say that General AI (GAI or Strong AI), which is envisioned as being the functional equivalent of a human mind, and which has been at the center of the concern in recent years about superintelligent rogue robots and the like, is set to undergo its own *GAI Winter*, as the futility of research towards GAI becomes evident. On the other hand, such a GAI Winter could be delayed if even a single multi-billionaire funder takes on General AI as a part of their legacy ambition.

ATTACHMENT W. PROVISIONAL CONCLUDING COMMENTS: DISCUSSION

Given the provisional concluding comments on pp F.3-13,14, what might some of the elements be of a serious advocacy response and alternative to the trajectory of AI presently underway, grounded in concern for economic justice, ecological integrity and technological responsibility? Here are some quick thoughts; more discussion is needed to address the question properly.

Three Responses

1. A **minimal response** might entail a long list of best practices, transparency reforms, social accountability measures and such, as sketched in Attachments C.2, C.3 and C.4. These would smooth many of the rough edges that AI is likely to present and could moderate the pace of AI development but wouldn't necessarily greatly change its developmental arc.

2. A **moderately strong response** might involve the sorts of measures now being mooted to deal with the social disutility of Big Tech in general, including divestiture or break-up mandates, regulation as public utilities, strong prohibitions on business practices involving collection, use and sale of data and the like. This could be accompanied by the civil establishment of significant *cordons sanitaires* or *unmediated spaces* that allow valued social institutions and perceptual/cognitive experience to flourish free of contamination or manipulation by AI and, presumably, other on-line/digital/screen-interfaced or, most generally, mediating technology.

3. A **strong response** would likely need to be part of a more general new understanding of the desired roles of science, technology, industrialism and free markets in human life and society, along with requisite cultural and institutional change.

On the Three Responses

4. The minimal response might be successfully achieved through existing or comparatively easily realizable means and institutions. Perceptive academics, journalists, government bodies, philanthropists and others have begun to engage the new challenges raised by AI. The formation of the several new progressive AI advocacy groups, grounded in social justice and civil and human rights, is an encouraging development. They will need to become bigger, and there will need to be more of them, and they will need to be joined by allies from a broad range of civil society constituencies. But the topography of this response is familiar and can support successful efforts.

5. The moderately strong response is perhaps the most immediately interesting option. It would require that everything sketched in the minimal response be pursued and brought to pass. For the rest, it would likely require a *new social movement* or something close to it. Such a movement would need to be prepared to challenge corporate power and the logic of market outcomes in a major way. It would need to go beyond the necessary legislative and regulatory reforms and involve changes that individuals and families make in the way we choose to use technology. Importantly, it would involve real changes in the way we recreate and spend time with one another. It would need to fully engage in local, national and international arenas. It would need to appeal to people across the full mix of demographic and other backgrounds. It would involve new ways of connecting and coming together in person in contrast to the ersatz, illusory connecting that so much online tech has touted. If it's real and big enough it would likely give rise to a variety of expressions with all the healthy tension that that can entail. It would need to foster simple joy and delight in a way that was once common despite all hardships and has now become scarce.

6. What I'm calling the strong response is a highly speculative notion and any useful comments would require further prior discussion. That needs to happen. The working paper outline touches on many topics that a strong response would likely need to address.

ADDENDA

These are preliminary notes on topics that might be later incorporated into the working paper.

ADDENDUM A. AI+ ADVOCATES' MACRO-STRATEGY AND SHORT-TERM STRATEGY

A.1. AI+ Macro-Strategy

The A+ advocates want to achieve personal immortality, omniscience and omnipotence, and have what they believe is a credible strategy for doing so. A greatly simplified summary follows. Much of this is from Bostrom's *Superintelligence* (2014) and from talks given at the 2015 Puerto Rico and 2017 Asilomar "Beneficial AI" convenings. I've kept some of the reasoning and tone used by the AI+ advocates, as the macro-strategy is primarily a rhetorical narrative.

After much study, the AI+ advocates concluded that strategies involving genetic enhancement, neural emulation, computer-brain interface or pharmaceuticals would for various reasons either be impractical or take too long.

They now figure that the strategy-of-choice should be grounded in **Artificial Intelligence**, i.e., using computer hardware to run software capable of doing everything that a human brain can do, and more. They figure that computers should be powerful enough to run a human brain program within ~50 years. However, it's possible that with the right policies and support human-level AI could be achieved within as few as 20 years.

Human-level AI can run much faster on a computer than in a wet brain, so once human-level AI on a computer has been achieved, it will be only a short period until that intelligence has engineered itself into **Superintelligence**. Most AI+ advocates estimate that it will take a few years to reach Superintelligence, but some believe that it would happen within months, weeks or even days.

Once Superintelligence is achieved, the **Singularity** will take place, perhaps within hours or minutes. The Superintelligence will expand in power without limit, perhaps explosively, elevating the entire planet and everything on it into a new and enhanced dimension of reality. This new reality will necessarily include everything and anything that anyone could ever possibly desire: health, wealth, happiness, sustainability, immortality, omniscience, omnipotence, and starships with which to colonize the Galaxy and beyond.

Humanity will now be Post-Human. Human-level AI will thus have been *humanity's last invention*.

But for all this to come to pass, and ideally within the next 20-30 years, we need to avoid several dangers.

One is that the Superintelligence doesn't share our values, or is unresponsive to our desire to be included in the Singularity, and either ignores, enslaves or kills us. To avoid this, we need to figure out how to program the early AI so that it *is aligned with human values*. Once we do that, we can more or less relax, confident that *whatever* happens, the AI, and thus the Super-AI that follows it, will act in ways that *are aligned with human values*.

We don't know how to do this at present, so it is imperative that a major research initiative get underway ASAP to figure it out. Scholars at such top universities as Oxford, Cambridge and MIT have already begun working on it.

Another danger is that as we get closer to achieving human-level AI, and begin talking in more detail about the path to Superintelligence and the Singularity, people might react uneasily. In the worst scenario they might revolt and unknowingly destroy the last chance that humanity has of reaching health, wealth, happiness, sustainability, immortality, omniscient and omnipotence. This would literally be the worst event for all of humanity in all of human history. Not only would the well-being and happiness of those living today be tragically cut short, but the well-being and happiness of *everyone who would otherwise have ever lived, i.e., our many, many immortal descendants, would be blocked from ever happening at all!*

We have to avoid this, no matter what. The best and preferred means is through popular education. We need to begin educating people *now* so that they understand what is happening as it unfolds. Once everyone understands that we are working so that *everyone* can have health, wealth, happiness, sustainability, immortality, omniscience, omnipotence and anything else they'd ever want, forever, who would possibly object?

Our only remaining significant danger is simply one of time. The mean estimate of ~ 50 years before we achieve human-level AI, Superintelligence and the Singularity is simply too long for many people alive today to be comfortable with. How can we achieve our goals by, say, Ray Kurzweil's target date of ~ 2040, about 20-25 years from now?

Again, mass popular education is the way. Once people understand how close we are to the final goal, there will be tremendous, spontaneous global demand that we *accelerate* our efforts. People in countries world-wide will demand that their governments commit whatever funds are necessary to bring this one-time most awesome moment in human history to completion; as noted, it will be the *last* moment in human history, because after the Singularity we will all be *Post-Human!*

Our immediate challenge, therefore, is to make sure that we proceed as fast as possible, but still without making any tragic mistakes. That's why such top universities as Oxford, Cambridge and MIT are establishing academic programs, including degree programs, focusing on artificial intelligence *safety*. As AI becomes the dominant technology world-wide, there will be a strong demand for thousands, perhaps tens of thousands, of AI safety experts. They will work for firms, governments and NGOs to advise them on the steps we need to take to realize the promise of AI and avoid any tragic mistakes.

Of course, we have to be realistic, and we have to plan for all scenarios. It's unlikely but possible that we could, for one reason or another, miss our goal of achieving the Singularity by 2040 or so and that the Singularity doesn't happen until, say, 2100 or even later. Again, many of those living today will be greatly disappointed. There is, however, a back-up plan, one with a very high likelihood of success.

Today, for only \$50,000, you can arrange to have your brain cryogenically deanimated and stored indefinitely in liquid nitrogen. Once the Singularity happens (say, in 100 or 200 years *at most*), the Superintelligence will figure out within seconds how to reanimate your brain and how to clone your body and quickly bring it to youthful maturity for you. Alternatively, you might simply decide to upload the neural code that defines you from your wetware brain onto a computer platform, one housed in a solar-powered vehicle about the size of a walnut that you can use to travel throughout the Galaxies, whenever and where ever you wish, forever.

Comments

The danger posed by the AI+ advocates is not that their vision of a superintelligent, singularitarian, post-human world might come to pass (it won't) but that the ideological/mythic force and coherence of their vision comes to appeal to a significant share of a significant sector of the world's population and begins to shape events.

I don't think the hard-core vision as recounted above will have great appeal. But I can imagine toned-down versions that, given the right conditions of global crisis, would.

The default focal sector would be the world's STEM-trained elites. A significant share of this sector already shares a fatalistic attitude towards technological developments in general ("if it can be done it will be done"). And technology is their baby. This sector is chomping at the bit to save humanity from climate change (through geo-engineering), from disease and suffering (through human genetic modification), from resource shortages (through nanotechnology and synthetic biology), and from inefficient governmental, commercial, financial, educational, transportation, industrial, defense and all other systems through the pervasive application of AI.

There are other possible focal sectors as well. It's not difficult to imagine a techno-utopian ideology being grafted onto any of many possible nationalist-populisms.

The hyper-technological world vision is an attempt to find a way around the real, inevitable changes that all people will need to make over the coming century and beyond if we are to live together in peace and justice on a finite planet. Well before it's finally accepted that we're not going to upload our minds onto nanochips and colonize the extra-solar rocky planet frontier, there will be ample opportunity for aggrieved havoc and worse. The rise of techno-totalitarian political sensibilities, accompanied by new racist biotech-eugenic ideologies, is only one of a number of completely plausible scenarios under the circumstances towards which we are now converging.

A.2. AI+ Short-Term Research and Cadre-Building Strategy

Here are the three key slides that Bostrom displayed during his January 2015 presentation in San Juan, Puerto Rico. The title of the presentation was *Superintelligence Explosion – The Road Ahead*.

Slide I. What needs to be done

1. Make theoretical progress happen on the control problem.
2. Build collaboration between AI developers and AI safety community to strengthen mutual understanding and trust
3. Create hireable safety experts
4. Ensure sufficient risk awareness so that serious AGI projects want to hire safety experts
5. Ensure the field is sufficiently well ordered so that merit can be ascertained
6. Promote sense of ethical responsibility & look for opportunities for positive-sum trades

Slide II. The Common Good Principle

1. Superintelligence should be developed only for the benefit of all of humanity and in the service of widely shared ethical ideals.
2. And: is consistent with commercial rewards for work along the way

Slide III. Research infrastructure needs (0---2 yrs)

1. FHI (Oxford University)
2. MIRI
3. CSER (Cambridge University)
4. FLI (individual researchers, events, etc.)
5. e.g. Berkeley?
6. e.g. MIT/Harvard?
7. Industry (e.g. DeepMind, Vicarious)
8. CFAR (talent scouting)

Brief comments:

The three slides display a 2-year political organizing agenda. During 2015 and 2016, key leadership among the AI+ community, supported by fund such as Musk, Tillman and Moskowitz worked to make it happen. They appear to have largely succeeded.

In his opening remarks at the January 2017 Asilomar conference, Jaan Tillman was beaming, as he inventoried the new research projects, academic institutes, media coverage and more that had gotten underway because of their efforts.

Here are brief comments of some of the items listed in Bostrum's three slides

What needs to be done

1. The "control problem" is the danger that an AI might escape human control and act autonomously and against the interests of humans.
2. The "AI developers" are the big firms that are doing the real AI work – Google, Apple, Facebook, etc – and the big university programs at Stanford, MIT, Carnegie Mellon, etc. The "AI safety community" are the *Transhumanists and Superintelligence advocates*, who are using the meme of "Threat of AI" to secure a perch in the media and other institutions for themselves.
3. The "hirable safety experts" are MA and PhD graduates of the Transhumanist/Superintelligence institutes at Cambridge, Oxford, etc. This niche is seen as their foot in the door of respectability and real influence.
4. Same goal: respectable, paying jobs for the apparatchiks.
5. Not sure what the key motivation is.
6. Sounds obvious. Not sure what the backdrop is.

The Common Good Principle:

1. Obvious. Nothing in this statement is inconsistent with the Superintelligence/signularitairn/post-Human vision
2. Part of the need to build credibility with Google, Apple, Facebook, etc.

Research infrastructure needs (0--2 yrs):

This list references the need for organizational development at 6 of the 8 AI+ advocates organizations identified at the beginning of this memo, along with mention of the need for continuing growth of AI research in the private sector. CFAR is the **Center for Applied Rationality**, another allied non-profit group and described on p. 64.

What is called the AI "safety community" are the AI+ Superintelligence groups, and what is called the "developers" are the mainstream big tech firms and university AI researchers. Many of the mainstream long dismissed the AI+ advocates as little more than wannabe's and groupies. They mostly still do, but they realize that the use of the "Filler AI Robots" meme by the AI+ community is able to generate enormous attention, and considerable funding. What Bostrom is proposing is that the mainstream start hiring "safety experts", i.e., PhDs and Mas that the dozen or so AI+ groups are generating at Cambridge, Oxford, MIT, Berkeley and elsewhere. The role of these safety experts is, at bottom, a public relations and political organizing role. Under the cover of "ethical reflections," stakeholder involvement" and the like, they will be working to promote an extreme version of what the human future should be. Even if the Singularity never happens (which it won't), the intents of many in the AI+ community, as well as many, perhaps most, of those active in the mainstream AI movement, for a hyper-techno-centric world, could come to pass. There are additional levels of nuance, and differences of strategy and ultimate objectives, between and among all the players, that are yet to be documented.

Note finally that all 8 of the research institutions that Bostrom identified as priority research infrastructure were well represented at the 2017 Asilomar meeting and have received significant funding over the preceding 24 months.

The Common Good principle is, again, part of the sleight-of-hand. There's nothing in that statement that is consistent with Superintelligence triggering the Singularity so that everyone can be happy, healthy, wealthy and immortal. The important move that the AI+ people are engaged in is seeing to it that people from their ranks occupy the entire set of commanding heights whenever there is any discussion of "The Human Future," whether at the universities, media empires, Davos, United Nations, and elsewhere.

ADDENDUM B. ADDITIONAL TOPICS FOR POSSIBLE INCLUSION IN THESE ATTACHMENTS

The numbered items are possible topics or elements of topics that might be included in the final version of this Attachment F.3 on artificial intelligence, or possibly in other Attachments or sections of the final working paper.

1. Review of the full set of plenary presentations and debate at the 2015 Puerto Rico convening and the 2017 Asilomar convening on “Beneficial AI.” See the website of the [Future of Life Institute](#) for all videos and pdfs.
2. Commentary on four videos from the Asilomar 2017 conference would document and expand upon many of the points made in these notes:
 - A. Friday morning [Welcoming Keynote](#) by Max Tegmark
 - B. Friday afternoon – [Presentation by Jeffrey Sachs](#)
 - C. Saturday afternoon Panel – Musk, Kurzweil, Bostrom, et al.:
[“If we succeed in building human level AI, then what are likely outcomes?”](#)
 - D. Sunday morning – [AI and the Law](#) [incl. comment by ACLU executive director Anthony Romero]
3. More material should be presented on the plans of the market dominant I-tech firms regarding research and development of AI, use of AI as part of internal operations and AI products for commercial sale. I haven’t touched on – because I don’t have much knowledge about – the larger corporate/financial environment within which each of the major firms is operating, what their strategies are, and how that might bear on our particular interests in the AI debate.
4. A note should be included on the unheralded role of Oxford and Cambridge Universities in promoting the political and philosophical framing of techno-eugenics, transhumanism and now artificial Intelligence. This directly derives from their historic role in the development and promotion of materialist reductionism, utilitarianism, analytic philosophy, classical liberalism, scientific racism and imperialism. In particular, the story of the Oxford Martin School needs to be told. At its inception it housed, among other centers, the Institute for the Future of Humanity and The Uehiro Center for Practical Ethics, both well-endowed and headed by leading transhumanists/techno-eugenics advocates Nick Bostrom and Julian Savulescu.
5. It might be useful to explore the question of clinical sociopathy, autism, introversion and related disorders associated with those engaged in AI research. This would need to be handled responsibly and with careful documentation. A knowledgeable doctoral student at a major U.S. engineering school estimates that 50% of the students there are clinically introverted.
- 6a. It might also be useful to explore the megalomania and hubris that characterizes most of the AI+ community and much of the mainstream AI community. In his 16 February 2017 manifesto “Building Global Community,” Facebook CEO Mark Zuckerberg seemed to be using the phrase “our community” to denote and conflate both the network of Facebook users and the entire human population.
- 6b. Here are representative quotes from the two “Beneficial AI” convenings (2015, 2017) regarding the work of the AI and AI+ communities:
 - The AI researchers are doing “the most important thing ever in human history.”
 - “... you are not in the business of creating artificial people... you are in the business of rewriting the laws of physics.”
 - “We are on the edge of the biggest invention ever; if we can solve intelligence, we can solve anything and everything.”
 - “...sponsoring [this conference] might have been the best way humanity has spent money, ever.”
 - “Very soon, we are going to have everything free”.
 - The goal of the Future of Life Institute is “to make the future as awesome as possible...”
 - “Everything we have is the product of intelligence... Access to significantly greater intelligence would be a step change in civilization.”

6c. Related: The utter cluelessness of many of the out front Kurzweil/Bostrom/transhumanist enthusiasts, and many of the less bizarre mainstream scientists as well (e.g., Russell, Tallinn, Tegmark), about the texture and grain of human life and human society. They talk about the challenges they face in trying to figure out what the most important global human values are, so that they can build these into the objective function and algorithms of their AIs, thus assuring “values alignment” between the AIs and humanity. If and when the robots take us over at least they’ll do so in a way that “aligns with human values.”

6d. Despite this criticism, most of the conference participants appear to sincerely want to do the right thing. The problem is that they also appear to have accepted the inevitability of a superintelligent/singularitarian/posthumanist future as a fact of nature. With that belief as a starting point, seemingly bizarre speculations of the sort freely shared, at Asilomar and in their many books, can be seen as rational and reasonable. I am concerned that the combination of megalomania, neediness and naivete as profound at that shown at the Asilomar conferences can, under the wrong circumstances, come to very bad ends, especially if coupled with access to real power.

7. It might be useful to include an attachment or appendix section illustrating Hebbian learning and neural networks.

8. The fact that so many lay people appear not only willing but eager to believe, and to *even say they welcome*, scenarios of AI or robotic domination or replacement is of concern. At a minimum it’s an expression of a feeling of disempowerment. Further on, it’s a statement of a lack of confidence in the prospects offered by a globalized, industrial, technologically-dominated world, even as it affirms and celebrates the creation of that world.

9. AI has received little attention from the wide range of civil society, non-governmental and non-commercial organizations, in the U.S. and elsewhere. It remains to be seen what sort of response will be forthcoming from, say, the environmental community, the human and civil rights communities, the labor movement, the international peace and justice communities, the religious community, the political parties and others.

10. From its inception until now the Silicon Valley/InfoTech community has been characterized by a strong libertarian socio-political ethos. This is discussed in detail in the working paper. In just the last ~ three years, however, this ethos has changed. Among the large, expanding, and youthful employee base, the libertarian ethos still dominates, but is being complexified by a strong concern over gender and racial justice and to a lesser but still noticeable extent economic justice. Government policies and programs to remedy, prevent or offset such inequities are now being strongly affirmed by leading Silicon Valley and AI figures. The proposal for a Universal Basic Income, in particular, has been widely embraced. At the 2017 “Beneficent AI” conference in Asilomar, Eric Brynjofsson of MIT presented this menu of what he called “Design Parameters” to address technologically-generated inequality:

- Universal Basic Income
- Earned Income Tax Credit
- Minimum Wage
- Education Investment
- Educational Transformation
- Anti-trust/Competition Policy
- Progressive Income Taxes
- Wealth Taxes
- Property Taxes
- R&D investment
- Infrastructure and Public Goods
- Privacy and Surveillance
- Federalism
- Trade policies
- Occupational licensing
- Intellectual Property Protection
- Distributed Capital (Robot) Owners

This is hardly a libertarian policy menu, to say the least. Questions have been raised about the sudden enthusiasm for such strong measures. It’s possible that this support is cynical: Silicon Valley might be assuming that none of this will ever happen at any scale – we can’t even agree on universal health care – and so there is little to lose, and

many PR points to gain, by forceful egalitarian advocacy. Alternatively, support may be sincere, at least among top Silicon Valley leadership, because they know that even ambitious programs of income transfer would still leave them filthy rich.

11. I've noted the fantastical visions of the human future propounded by AI+ advocates, and I've noted the still profoundly consequential accounts of the human future suggested by mainstream AI advocates. But I haven't given examples of the latter, other than to note the general agreement that an AI-dominant world will generate enormous wealth unequally distributed, as a consequence of which measures such as those just listed in (10) above will need to be adopted.

12. A more thoroughly thought-through account of the coming decades is offered by **Kai-Fu Lee (2017)**, chairman and CEO of Sinovation Ventures and president of its Artificial Intelligence Institute. He gives a sense of the magnitude of the policy challenge we face. He says that the extreme economic inequality that AI and other advanced technology will likely generate over the coming decades, along with the fact that the great majority of economic gains will be concentrated in the United States and China, requires that the U.S. and China stand as a **dual global coalition** that agrees to tax the wealthy firms and individuals within its borders and to distribute the proceeds to countries within the U.S. and China's respective "spheres of responsibility" to be used as income support intended to preserve global harmony. This account is likely fantastical as well. The fact that it was offered in all seriousness by mainstream leadership indicates how large the challenge we face actually is.

13. An organization that might be considered part of the larger AI+/transhumanist network is the [Foundational Questions Institute](#), aka [FQXI.org](#). Described as "exploring the foundations and boundaries of physics and cosmology," it brings many of those associated with the AI+ and transhumanism together with others in physics and cosmology to share ideas in what might be called speculative science. It has held conferences, sponsors research, creates podcasts and videos and the like. Since 2006 it has given \$15 million in research grants. The source of the funds is not disclosed on the website. FQXI was founded in 2005 by MIT cosmologist Max Tegmark.

14. **ADDITIONAL ORGANIZATIONS CHALLENGING OR CRITICALLY ADDRESSING AI, BIG TECH, TECHNOCRACY**

This is a preliminary list. More, and with descriptions, will be included in the final working paper. This list doesn't include groups associated with the AI+/superintelligence movement.

Additional Groups and Initiatives

1. [Campaign to Stop Killer Robots](#) – Human Rights Watch, European Union
2. [Center for Humane Technology](#)
3. [Freedom from Facebook](#)
4. [Common Sense Media](#)
5. [Tactical Technology Collective](#) => [Our Data Ourselves](#)
6. [Tech Workers Coalition](#)

Slow Tech / Unplugging

1. [Slow Tech Parenting](#)
2. [National Day of Unplugging](#) March 1-2, 2019 → [Sabbath Manifesto](#) → Reboot

Academic Centers, institutes and Collaborations

1. [Council on Big Data, Ethics and Society](#)
2. [The Center for Internet and Society](#)

ADDENDUM C. UPDATE ON EFFECTIVE ALTRUISM, TRANSHUMANISM, LONGTERMISM AND AI – JUNE 2023

Since July 2018 the EA movement had grown dramatically in participation, infrastructure and influence. By mid-2022 EA controlled “... philanthropic resources on the order of thirty billion dollars.”²² For much of this period EA’s most prominent donor and representative in high-level financial, tech, media and political circles was Sam Bankman-Fried (SBF). He established and led multiple cryptocurrency investment and exchange entities, including major firms Alameda Research (AR) and FTX, and was a close friend and supporter of EA co-founder Oxford philosophy professor William MacAskill. In late 2022 FTX and AR simultaneously underwent scandalous impulsive bankruptcies; SBF was arrested and indicted for fraud, conspiracy and money-laundering. Here is a timeline of selected events in the growth and development of EA, with special attention to the involvement of SBF.²³

= = =

2008 – William MacAskill begins graduate study in philosophy at Oxford. He grounds his moral thinking in the utilitarian/consequentialist philosophy of Peter Singer. He lives frugally and gives 10-30% of his stipend to charity.

2009 – MacAskill meets similarly motivated fellow Oxford philosophy grad student Toby Ord, who pledges 50% of his earnings to good causes chosen by ‘rational’ rather than ‘emotional’ assessment. The two create the non-profit **Giving What You Can** to promote this philosophy and practice.

2011 – Ord and MacAskill start **80,000 hours** to encourage students at elite universities to ‘earn to give,’ i.e., pursue high-paying careers and contribute as much as you can to rigorously evaluated good causes.

2012 – MacAskill sets up the **Centre for Effective Altruism (CEA)** to guide and support the growing EA community. He shares office space with Oxford philosophy professor Nick Bostrom’s transhumanist **Future of Humanity Institute**. MacAskill and Bostrom come to see EA and transhumanist thinking and strategy as mutually supportive.²⁴

2012 – While visiting MIT in Cambridge, MA, MacAskill meets undergrad physics student Sam Bankman-Fried (SBF), who subsequently aligns with EA and its earn-to-give ethic. MacAskill and SBF develop an ongoing friendship.

2014 – MacAskill finishes his philosophy PhD at Oxford. SBF finishes his physics BA at MIT and goes to work for global proprietary trading firm Jane Street. He gives half his salary to EA-endorsed causes and for EA infrastructure.

2015 – MacAskill is appointed assistant professor of philosophy at Oxford. At 28 he’s the youngest such appointee ever. His first book, **Doing Good Better**, outlines the EA moral and philanthropic philosophy.

2015-2016 – EA grows rapidly among libertarians, ‘rationalists,’ techies and transhumanists in the UK, the US and the rest of the Anglosphere.²⁵ It begins a shift in focus from alleviating present suffering to preventing human extinction. The utilitarian/consequentialist rationale for this shift, as formulated and propounded by Nick Bostrom, is that the timely prevention of human extinction would allow 10^{52} human beings to come into existence who would otherwise not be able to. Again following Bostrom’s lead, key EA leaders come to identify malevolent AI+ as the most likely and near-term agent of human extinction.²⁶

2017 – **Open Philanthropy (OP)** is set up in San Francisco to advise on and manage charitable donations to EA causes. Facebook billionaire Dustin Moskovitz tasks OP to spend down his fortune in alignment with EA values.

October 2017 – SBF joins CEA as Director of Development. He leaves shortly afterward but remains in regular contact with the informal and largely unacknowledged EA leadership circle.

November 2017 – SBF starts **Alameda Research (AR)**, a quantitative crypto-focused hedge fund. AR presents itself as an EA firm whose profits will go to EA causes. AR quickly succeeds with an arbitrage strategy involving crypto

price differentials in Japan and the U.S. AR then sets on an aggressive growth path using a full set of strategies, including market making, yield farming, volatility trading, venture investment and investment management.²⁷

November 2017 – April 2018 - From the start, AR staff find SBF routinely ignoring conventional norms of business ethics, lying to investors and creditors, comingling investor capital and firm capital, indulging a “staggering appetite for risk,” designing “predatory” financial products intended to “scam” investors, having “inappropriate sexual relationships with subordinates” and in general being “extremely difficult to work with.” Key staff tell SBF that unless this stops they will quit. SBF rebuffs them. More than half the staff quit, including the full management team other than SBF.²⁸

Spring 2018 – Throughout this turmoil MacAskill and the other EA leadership rally around SBF and dismiss the fracas as normal start-up frictions and ‘he said-she said.’

Late 2018 – SBF restarts AR with a core of loyal staff and moves to Hong Kong, which allows crypto trade activities that are illegal under U.S. law.

May 2019 – SBF establishes **FTX**, a user-friendly crypto asset trading platform. The timing is right – crypto is the hot new thing, and thousands of users begin trading crypto through FTX. It is set up in The Bahamas because China had criminalized crypto-related activities. Funds that customers have deposited in their FTX accounts are illegally drawn upon by SBF to support trades, deals and ventures at AR, and for personal use, e.g. political contributions.²⁹

March 2020 – Toby Ord publishes *The Precipice*, stating that extinction due to runaway AI+ is humanity’s greatest threat and is thus an EA priority. MacAskill had been skeptical but was won over by reports of Open AI’s GPT-3.

2020 – SBF donates \$10 million to EA candidate for US Congress from Oregon Carrick Flynn, in an open primary whose leading candidate is a progressive Democrat. Flynn runs on a platform of AI safety and longtermism in this significantly rural, working class and Latino district and loses badly. SBF donates \$5 million to the Joe Biden presidential campaign.

2021 – *Forbes* estimates that SBF, now 29, has net worth of \$23 billion.³⁰

2021 – SBF tells MacAskill he wants to give \$1 billion annually to EA-endorsed longtermist causes and for EA institution-building. SBF sets up the **FTX Future Fund**, hires longtermist Oxford philosopher Nick Beckstead as CEO and enlists MacAskill as an FTX adviser.

2021 – MacAskill begins discussions with Hollywood figures who suggest “setting up a pipeline from the EA community to Hollywood... and seeing if we can’t get these ideas out there into the world.”³¹

2021 – CEA buys an ancient estate in the English countryside, turns it into a luxury retreat and meeting venue for EA staff, membership and invited guests, including potential funders and political candidates.³²

Late 2021 – Discord arises within the EA community over the shift from directly alleviating current human suffering to ensuring that malevolent AI+ doesn’t destroy all humanity and thus prevent 10^{52} future humans from ever having lived. Staff and members present MacAskill with proposals for greater democracy, transparency and accountability in EA decision-making. MacAskill is unresponsive.³³

April 2022 – 80,000 hours posts a cringe-worthy three-hour podcast further promoting SBF as a visionary EA decabillionaire who has pledged 99% of his fortune to EA causes.³⁴

April 26-29 – FTX is lead co-host of the invitation-only [1st Annual Crypto Bahamas Conference](#), attended by 2000 investors, developers, entrepreneurs, journalists and others immersed in the world of finance, technology and crypto. Speakers include Bill Clinton, Tony Blair, Gisele Bündchen and Tom Brady. But the event is in fact SBF’s coming-out party to a new level of global celebrity and influence.³⁵

Summer 2022 – The Future Fund, now based in Berkeley CA, grants \$160 million to 67 EA causes, mostly for AI+ safety, biorisk and scenario forecasting. \$33 million of these grants go to organizations connected to MacAskill: \$13.9 M to Center for Effective Altruism, \$17.9 M to Longview Philanthropy, \$1.2 M to Global Priorities Institute.³⁶

July 2022 – Core EA leadership hear that a U.S. federal investigation of SBF may be underway. They discuss the risk of further association with SBF but take no action.

August 2022 – FTX is ordered by a federal bank regulatory official to stop making “false and misleading” claims concerning the extent to which depositor funds are insured by the U.S. government.

August 2022 – MacAskill’s new book *What We Owe the Future* presents longtermism as a human imperative and EA as its moral foundation. The book presents the techno-utopian transhumanist future as the good to which all humans self-evidently aspire. Elon Musk tweets that the book “... is a close match for my philosophy.”³⁷

Through early November 2022 – SBF gives at least \$39.2 million to Democratic Party mid-term campaigns, becoming the largest party donor after George Soros.³⁸

Early November 2022 – Reports circulate of irregularities in FTX’s financial structure and operations. Customers begin withdrawing funds, further irregularities are revealed, charges of fraud surface and a classic liquidity crisis ensues. SBF denies there is a problem but seeks bailouts from within the crypto community. He is unsuccessful.³⁹

November 11 – FTX, AR and associated entities declare bankruptcy; \$8-10 billion in customer funds ‘vaporize.’⁴⁰

November 11 – The entire staff of the FTX Future Fund resigns. MacAskill tweets that if the charges of fraud are true “...I am outraged, and I don’t know which emotion is stronger: my utter rage at Sam (and others?) for causing such harm to so many people, or my sadness and self-hatred for falling for this deception.”⁴¹

December 12 – SBF is arrested in the Bahamas and later extradited to the U.S. He is charged with twelve counts of wire fraud, securities fraud, conspiracy and money laundering.⁴²

Addendum C – References

Adams, Carol J., Alice Crary and Lori Gruen, eds. 2023. *The Good It Promises, the Harm It Does: Critical Essays on Effective Altruism*. Oxford: Oxford University Press.

Alter, Charlotte. 2023. “[Exclusive](#): Effective Altruist Leaders Were Repeatedly Warned About Sam Bankman-Fried Years Before FTX Collapsed.” *TIME*, 15 March.

Altraide, Dagogo. 2023. “[The FTX Disaster is Deeper than you Think](#).” (Video). *ColdFusion.com*.

Bostrom, Nick. 2003. “[Astronomical Waste](#): The Opportunity Cost of Delayed Technological Development.” *Utilitas*, Vol. 15, No. 3, pp. 308-314.

Bostrom, Nick. 2002. “Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards.” *Journal of Evolution and Technology*, Vol. 9, No. 1.

Douglas, Sylvia and Nick Fountain. 2022. “[Sam Bankman-Fried and the spectacular fall of his crypto empire, FTX](#).” (audio and transcript). *Planet Money*. National Public Radio. 16 November.

Ehrlich, Steven. 2021. “[Meet The World’s Richest 29-Year-Old](#): How Sam Bankman-Fried Made A Record Fortune In The Crypto Frenzy.” *Forbes* (cover story). 6 October.

Fleck, Anna. 2022. “[Midterm Election Mega-Donors](#).” *Statistica.com*. 28 October.

Goldstein, Matthew, Alexandra Stevenson, Maureen Farrell and David Yaffe-Bellany. 2022. “[How FTX’s Sister Firm Brought the Crypto Exchange Down](#).” *New York Times*. 18 November.

Legraien, Lea. 2023. "[£15m Oxford property purchase not funded by failed crypto platform](#), says charity." *civilsociety.co.uk*. 22 February.

Lewis-Kraus, Gideon. 2022a. "Do Better" – *The New Yorker*, 15 August.

Lewis-Kraus, Gideon. 2022b. "[Sam Bankman-Fried](#), Effective Altruism and the Question of Complicity." *The New Yorker*, 1 December.

Lowrey, Annie. 2022. "Effective Altruism Committed the Sin It Was Supposed to Correct." *Atlantic*. 17 November.

MacAskill, William. 2022. [Tweet thread](#). 11 November.

Musk, Elon. 2022. [Tweet thread](#). 1 August.

Robinson, Nathan J. 2022. "Defective Altruism." *Current Affairs*. 19 September.

Szalai, Jennifer. 2022. "[How Sam Bankman-Fried Put Effective Altruism on the Defensive](#)." *NY Times*, 9 December.

Táiwò, Olúfémí O. and Joshua Stein. 2022. "Is the effective altruism movement in trouble?" *The Guardian*. 15 November.

Torres, Émile P. 2023. "[Why Effective Altruism and "Longtermism" Are Toxic Ideologies](#)." *Current Affairs*. May.

United States Attorney's Office – Southern District of New York. 2022. *Bankman-Fried Charged in an Eight-Count Indictment with Fraud, Money Laundering, and Campaign Finance Offenses* (Press Statement). 12 December. See Attachment B below.

United States Security and Exchange Commission. 2023. *Litigation Release No. 25616*. 19 January. See Attachment A below.

Wallace, Benjamin. 2021. "[The Tech Elite's Favorite Pop Intellectual](#)." *New York*. 13 April.

Wang, Tracy. 2022. "[Wall Street Goes Crypto in the Bahamas](#)." *Coindesk.com*. 4 May.

Wiblin, Robert, and Keiran Harris. 2022. "[Sam Bankman-Fried on taking a high-risk approach to crypto and doing good](#)." *80,000hours.org*. 14 April.

Yaffe-Bellany, David. 2022. "[A Crypto Emperor's Vision](#): No Pants, His Rules." *New York Times*. 14 May.

#

ADDENDUM D. TRANSHUMANISM AS PRECURSOR TO AND MAJOR INFLUENCE ON DEVELOPMENT OF EA AND LONGTERMISM – JULY 2023

This addendum displays a timeline concerning:

- 1) the rise and decline of *transhumanism* as an explicit ideology and social movement.
- 2) the development of *effective altruism* and *longtermism* as together incorporating, building upon, refining and supplanting the ideology and program of transhumanism in a form more suited for movement-building, in particular among young tech and tech-adjacent academics and entrepreneurs.

Before 1988: Authors and researchers independently begin work that later forms the foundational core of transhumanism. Areas of interest include life extension and cryonics, human genetic enhancement and techno-eugenics, nanotechnology, artificial intelligence, cyborgs and human-machine integration, the nature of and prospects for the Singularity, and large-scale space colonization.⁴³

1988: Futurist writers Max More and Tom Morrow publish the first issue of *Extropy* magazine to address the above and related topics and their synergies.

1992: The **Extropy Institute** is established by More, Morrow and others, and the Extropy Mailing List is compiled to allow online exchange of ideas among interested colleagues.

1994: The first Extropy Institute Conference on Transhumanist Thought is held in Sunnyvale, CA. Major sessions focus on the philosophy of Extropianism, digital uploading of human minds and life extension/cryogenics.

Late 1990s-early 2000s: Scientific and technological developments encourage transhumanist enthusiasm:

- 1996 - Cloning of Dolly the Sheep by Ian Wilmut in Edinburgh UK
- 1997 - IBM computer Deep Blue defeats world chess champion Gary Kasparov
- 2002 - Elon Musk establishes SpaceX with the ultimate goal of colonizing Mars
- 2003 – Near-complete sequencing of the human genome is announced by U.S. President Bill Clinton

1997: UK grad student Nick Bostrom estimates that we will develop superhuman AI sometime before 2034.⁴⁴

1998: Bostrom and British philosopher David Pearce found the **World Transhumanist Association (WTA)** as an umbrella organization for national, local and individual transhumanist initiatives.

2000: **Seasteading Institute** is set up to establish sovereign polities in international waters exempt from national proscriptions on human experimentation, rDNA applications, drug development, patent infringement, etc.

2002: In his paper “Existential Risk: Analyzing Human Extinction Scenarios and Related Hazards,” Bostrom defines and explicates the concept of *existential risk* as it is widely understood today.^{45 46 47}

2003: In his paper “Astronomical Waste: The Opportunity Cost of Delayed Technological Development,” Bostrom explicates the rationale for longtermism as it is widely understood today, but doesn’t use that word itself.

Early 2000s: First explicit opposition to transhumanism is voiced, coming from both the progressive left and the social conservative right:

- 2000 - Human Genetics Alert (HGA) is established in London, UK
- 2000 - Center for Genetics and Society (CGS) is established in San Francisco, CA
- 2002 - President George Bush’s Council on Bioethics is appointed with noted ethicist Leon Kass as chair
- 2004 - Francis Fukuyama, in *Foreign Policy*, names transhumanism “The World’s Most Dangerous Idea”

2004: In a move to appear less bizarre and cultish, WTA changes its name to *Humanity+* and espouses a less right-libertarian and more centrist-liberal public face, without, however, significantly changing its vision or program.

2004. As part of this move, Bostrom and transhumanist sociologist James Hughes establish the **Institute for Ethics and Emerging Technologies (IEET)** to promote transhumanist perspectives within the mainstream professional bioethics community.

2005: Bostrom and colleagues establish the **Future of Humanity Institute (FHI)** at Oxford University as a base for developing and promoting transhumanist thought.

2005: Ray Kurzweil's *The Singularity is Near* is a NYT bestseller. **Singularity University** is founded in 2008 in Silicon Valley to prepare young people for careers in the growing info-nano-bio-cogno-neuro-robo industrial complex.

late 2000s – early 2010s: "...a golden era for transhumanism."⁴⁸ Selected events of that period include:

2008 - Nick Bostrom appointed Professor, Faculty of Philosophy, Oxford University.

2009 - Aubrey de Grey founds the SENS Foundation for research to find a cure for aging.

2009 - *The Transhumanist Declaration* is promulgated by the WTA.⁴⁹

2009 - Ideological 'rationalist' Eliezer Yudkowsky starts the online forum *LessWrong* in support of 'beneficial AI.'

2011 - *Time* magazine cover story promotes transhumanist vision of body hacking and near-term immortality.

2011 - Max More becomes CEO of Alcor Life Extension Foundation.

2012 - Google hires Ray Kurzweil as Chief Engineer to work on machine learning and AI.

2013 - Larry Page establishes Calico Labs for research on life extension and associated human enhancements

2014 - The Transhumanist Party is founded by [Zoltan Istvan](#); he runs for U.S. President in 2016.

2012: Philosophy PhD student William MacAskill sets up **Centre for Effective Altruism (CEA)** at Oxford, sharing offices with Bostrom's FHI.⁵⁰ At its inception EA supported mostly conventional poverty-relief and humanitarian charities which passed the high bar of EA's quantitative screening, and encouraged personally frugal lifestyles so that larger shares of personal income could go to good causes.⁵¹

2014: Bostrom's best-seller *Superintelligence: Paths, Dangers, Strategies* argues that humanity should make the achievement of AI superintelligence its major priority, as this will allow all other problems (e.g. climate, poverty, war) to be quickly and easily solved. He says that "AI Safety" must be ensured before superintelligent AIs are developed, lest one decide to enslave or exterminate us. The term 'transhuman-' doesn't appear in the book.^{52 53 54}

2013-2017: Developments in machine learning, neural networks, pattern recognition, robotics and large language models spark public debate among developers concerning risks of AI. Seemingly competing views of luminaries (Musk, Gates, Hawking et al.) are headlined. Open Letters on the dangers of AI are publicized (by *pro-AI* partisans; see next). "Safe AI" institutes are established in Berkeley CA, Cambridge MA, Boston MA, Oxford UK, elsewhere.⁵⁵

January 2015: Invitational convening in San Juan Puerto Rico of key AI, EA and related leadership and funders to develop a strategy for controlling public debate and policy concerning AI. Core proposal: say AI poses existential risks (i.e. human extinction), but that beneficial/safe/friendly AI holds the greatest opportunity ever for human flourishing and abundance. Development of beneficial/safe/friendly AI should thus be a top priority for humanity.⁵⁶

2015: MacAskill is appointed assistant professor of philosophy at Oxford.

July 31-Aug 2 2015: Participants at the EA Global Summit Conference held at the Mountain View CA Googleplex hear Musk, Bostrom, MacAskill and others declare that the development of AI holds the key to maximum human flourishing and abundance, but that it must be friendly AI lest something goes wrong and it decides to kill us all.⁵⁷

December 2015: Elon Musk and Sam Altman launch **Open AI** in San Francisco with \$1 billion in funding and the declared mission of developing Artificial General Intelligence (AGI) and releasing it free to all as open source.^{58 59}

January 2017: Second invitational convening, now at Asilomar, CA, to refine and agree upon a consensus strategy in support of beneficial/safe/friendly AI.⁶⁰

October 2017: MacAskill, fellow Oxford philosophy professor Toby Ord and colleagues agree upon the term *longtermism* to identify the new philosophic and programmatic stance that EAs are expected to espouse.^{61 62}

January 2018: Bostrom and others establish the **Global Priorities Institute** at Oxford to provide academic research supporting longtermism.⁶³

December 2018. MacAskill and others establish the **Forethought Foundation for Global Priorities Research (FF)**.⁶⁴

2019. The **Center for Security and Emerging Technologies (CSET)** is established at Georgetown University to promote longtermism within DC policy and leadership networks, and to train young longtermist cadres and place them in influential positions in Washington, the United Nations and elsewhere.⁶⁵

March 2020. Toby Ord publishes *The Precipice*, which highlights Bostrom's arguments about the centrality of AI and Superintelligence as simultaneously humanity's great risk and greatest potential benefit, and thus the proper EA priority. This is the first book-length introduction to longtermism. 'Transhuman-' doesn't appear in the book.

September 2020. Ajeya Cotra, senior researcher with EA-aligned **Open Philanthropy** in San Francisco, estimates 15% chance that "transformative AI" will be achieved by 2036 and 50% chance by 2050.^{66 67 68}

2021. Opposition to longtermism begins to surface from diverse sources, including from the AI research community and from within the EA network itself.^{69 70 71}

August 2022. *The New Yorker* reports that as of mid-2022 the EA network controlled "... philanthropic resources on the order of thirty billion dollars."^{72 73}

August 2022: MacAskill publishes *What we Owe the Future*, specifically billed as an introduction to and explication of the longtermism movement. The term 'transhuman-' appears nowhere in the book.

August 2022 – Transhumanist George Dvorsky posts "**Whatever Happened to the Transhumanist?,"** in which he notes the decline of transhumanism as an ideology, community, social movement and topic of media attention.⁷⁴

November 2022 – Open AI releases ChatGPT-4. This sparks a wave of enthusiasm, fear, reaction and calls for action to regulate AI significantly greater than the earlier 2014-2017 wave.⁷⁵

November 2022 – Multi-billionaire and major EA advocate and funder Sam Bankman-Fried's cryptocurrency empire collapses following accusations of fraud and mismanagement. He is later arrested and charged with multiple federal offenses. EA and MacAskill are widely criticized for their past embrace of SBF.⁷⁶

January 2023 – An explicitly racist 1996 on-line post by Nick Bostrom is uncovered and made public. The post is condemned by Oxford University officials and an investigation of the matter is begun. Bostrom offers a poorly expressed apology. Some EA members call upon Bostrom to resign as head of the FHI.⁷⁷

Spring 2023 – Two *Open Letters* are circulated, both signed by noted AI luminaries. One calls for a 6-month pause on further training of AI systems more powerful than GPT-4. The other calls for the risk of extinction from AI to be made a priority global concern comparable to concern about pandemics and nuclear war.^{78 79}

June 2023 – The European Parliament approves the **Artificial Intelligence Act (AIA)** to establish strong regulatory oversight of AI, including formal bans on AI applications judged to be dangerous to individuals and/or society.⁸⁰

July 2023 – *NATURE* editorial says that domination of the AI debate by those who say AI is an all-powerful technology that could lead to human extinction is simultaneously generating enormous investment funds and letting big tech avoid regulations of AI practices causing real societal harm right now. It says that this debate is being framed and led by "a homogeneous group" and that other communities are being left out.⁸¹

Additional Note for Addendum D: *The State of Transhumanism Today*

George Dvorsky (2022) queries noted transhumanist activists and writers as to the state of their movement and what accounts for its decline. Responses (here heavily rewritten for concision and clarity) include:

* We don't hear much about transhumanism anymore because it's become normalized and largely accepted as where we're going, even if not as rapidly as we had hoped. We live in a world of online immersion, virtual reality, gene therapies, ubiquitous AI, increasingly transhuman prostheses, and space tourism. Ours is a world of growing transhumanism on its way to posthumanism, even if we don't use those words.

* Transhumanism arose in the 1990s during a time of great enthusiasm about technology in general. But over the last fifteen years we've seen the ugly downsides of tech: disinformation, deep-faking, surveillance, China's 'social credit' regime, social isolation, mental health issues, subversion of the democratic process and more. And Big Tech has proven to be just as greedy and exploitative as any other powerful capitalist industry.

* People world-wide are confronting climate catastrophe, economic injustice, racism, sexism, police and military violence and more. To expect anyone, and especially members of front-line communities, to become excited about uploading their minds, having designer babies and living in space colonies is elitist privilege on steroids. Before we embark on transhumanity we need to make sure everyone is treated humanely in the first place.

* Transhumanist thinkers have refined their focus. Rather than foregrounding morphological freedom, cryonics and living as cyborgs – all of which many people find disturbing, even grotesque – they now emphasize the humanitarian imperative of protecting future generations from existential catastrophe. AI is now presented as both our greatest existential risk, and so needs to be regulated, and our greatest means for an abundant, transformative future, and so needs should be supported with that end in mind. People like that kind of narrative.

* Hollywood and the media in general have fully embraced transhumanist themes and images, but unfortunately can't resist portraying these in an ambivalent if not dystopian manner. And the day is saved not by technology, much less by transhumanist technology, but by conventional, predominantly male, forcefulness and/or violence.

* Transhumanism quickly divided into often incompatible political camps – techno-libertarians, techno-progressives, techno-authoritarians, even techno-monarchists. Transhumanists became partisans of particular technologies and pathways: life extension vs. mind uploading vs. genetic enhancement vs. space colonization and so on. They'd tout their favored tech scenario and trash the others. Now it's unlikely that we'll ever see the emergence of an explicitly transhumanist movement. Rather, we'll see transhumanist sensibilities and practice advance incrementally as one or another of these separate political and technological movements show results.

* Early transhumanist proposals such as cryonics, human cloning and nanobots to clean out your arteries now feel like "that '90s stuff," i.e., old-fashioned, almost comical fantasies, like aging baby-boomers still waiting for their personal jet-packs, atomic powered automobiles and underwater cities.

Zoltan Istvan (2023) shared these thoughts regarding current and future prospects for transhumanism:

1. The longevity project that has been central to transhumanism has been disappointing. It has not produced the results that many transhumanists expected would be realized by this time.
2. The AI project, on the other hand, has moved faster than expected, and now promises to completely reshape our livelihoods and our systems of education, business, scientific research and more. Profound changes in human civilization will begin to be evident beginning 3-5 years from now, i.e., 2026-28.
3. Superintelligent AIs coupled with human brains will quickly figure out a practicable road to longevity and then immortality. Once on this path we will move at an accelerating pace to transhumanity and then posthumanity. ⁸²

Addendum D - References

- Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford UK: Oxford University Press.
- Bostrom, Nick. 2003. "[Astronomical Waste](#): The Opportunity Cost of Delayed Technological Development." *Utilitas*, Vol. 15, No. 3, pp. 308-314.
- Bostrom, Nick. 2002. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology*, Vol. 9, No. 1.
- Bostrom, Nick. 1997. "[How Long Before Superintelligence?](#)" *International Journal of Future Studies*. Vol 2.
- BostromAnonAccount. 2023. "[Nick Bostrom should step down as Director of FHI.](#)" *Effective Altruism Forum.com*. 4 March.
- Chugg, Ben. 2020. "[Against Strong Longtermism: A Response to Greaves and MacAskill.](#)" *Medium.com*. 17 December.
- Crary, Alice. 2023. "[The Toxic Ideology of Longtermism.](#)" *Radical Philosophy*. Spring.
- Cotra, Ajeya. 2022. "[Two-year update on my personal AI timelines.](#)" *LessWrong.com*. 2 August.
- Cotra, Ajeya. 2020. "[Draft report on AI timelines.](#)" *LessWrong.com*. 18 September.
- Cremer, Carla Zoe and Luke Kemp. 2023. "[Democratizing Risk and EA with Carla Zoe Cremer and Luke Kemp.](#)" *Critiques of EA* (Podcast). 2 February.
- Cremer, Carla Zoe and Luke Kemp. 2021. [Democratizing Risk: In Search of a Methodology to Study Existential Risk](#). 28 December.
- Davis, Jacob. 2023. "[Longtermists Are Pushing a New Cold War With China.](#)" *Jacobin.com*. 25 May.
- Dvorsky, George. 2022. "[What Ever Happened to the Transhumanists?](#)" *Gizmodo.com*. 1 August.
- European Parliament News. 2023. "[MEPs ready to negotiate first-ever rules for safe and transparent AI.](#)" *European Parliament press office*. 14 June.
- Gebru, Timnit. 2022. "[Effective Altruism is pushing a dangerous brand of 'AI Safety.'](#)" *WIRED*. 30 November.
- Gebru, Timnit and Émile Torres. 2023. [Understanding TESCREAL with Dr. Timnit Gebru and Émile Torres](#). *Dave Troy Presents*. 14 June.
- Greaves, Hilary and William MacAskill. 2019. [The Case for Strong Longtermism](#). Working Paper. Oxford: Global Priorities Institute. September.
- Hao, Karen. 2020. "[We read the paper that forced Timnit Gebru out of Google. Here's what it says.](#)" *MIT Technology Review*. 4 December.
- Hogan, John. 2022. "['Longtermism' Movement Misses the Importance of War.](#)" *Scientific American*. 28 September.
- Istvan, Zoltan. 2023. "[The Current State of Transhumanism.](#)" Interview. *Singularity Radio*. March.
- Jacobson, Matt. 2022. "[The Problems with Longtermism.](#)" *Quillette.com*. 27 October.
- Karnofsky, Holden. 2021. "[Forecasting Transformative AI, Part 1: What Kind of AI?](#)" *Coldtakes.com*. 10 August.
- Karnofsky, Holden. 2016. "[Some Background on Our Views Regarding Advanced Artificial Intelligence.](#)" *OpenPhilanthropy.org*. 6 May.
- Khatchadourian, Raffi. 2015. "[The Doomsday Invention](#): will artificial intelligence bring us utopia or destruction?" *The New Yorker*. 23 November.
- Lewis-Kraus, Gideon. 2022. "Do Better." *The New Yorker*, 15 August.

Liboreiro, Jorge and Aida Alonso. 2023. "[MEPs endorse blanket ban on live facial recognition in public spaces, rejecting targeted exemptions.](#)" *Euronews.com*. 14 June.

Linton, Samara. 2023. "[Tech Elite's AI Ideologies Have Racist Foundations, Say AI Ethicists.](#)" *People of Color in Tech. (POCIT)*. 24 May.

MacAskill, William. 2022. *What We Owe the Future*. New York: Basic Books.

MacAskill, William. 2019. "[Longtermism](#)." *Forum.effectivealtruism.org*. 25 July.

Matheny, Jason G. 2007. "[Reducing the Risk of Human Extinction.](#)" *Risk Analysis*. Vol 27, Issue 5. October.

Matthews, Dylan. 2015. "[I spent a weekend at Google talking with nerds about charity. I came away ... worried.](#)" *vox.com*. 10 August.

McGoey, Linsey. 2023. "[Elite Universities Gave Us Effective Altruism, the Dumbest Idea of the Century.](#)" *Jacobin.com*. 19 January.

Miller, Asher, Rob Dietz and Jason Bradford. 2023. "[How Longtermism Became the Most Dangerous Philosophy You've Never Heard of.](#)" Crazy Town Podcast, Episode 73. *Resilience.org*. 17 May.

Nature. 2023. Editorial. "Stop talking about tomorrow's AI doomsday when AI poses risks today." 29 June.

Naughton, John. 2022. "[Longtermism: how good intentions and the rich created a dangerous creed.](#)" *The Guardian*. 4 December.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. New York: Hachette Books.

Torres, Émile P. 2023a. "[Existential Risks, Transhumanism and Longtermism.](#)" Interview. *SingularityFM*. 3 May.

Torres, Émile P. 2023b. "[Why longtermism and EA are such toxic ideologies.](#)" *Current Affairs*. 7 May.

Torres, Émile P. 2023c. "[Longtermism and Eugenics: A Primer.](#)" *Truthdig.com*. 4 February.

Torres, Émile P. 2022. "[Selling 'longtermism': How PR and marketing drive a controversial new movement.](#)" *Salon.com*. 10 September.

Torres, Émile P. 2021a. "[Against longtermism.](#)" *Aeon.com*. 19 October.

Torres, Émile P. 2021b. "[The Dangerous Ideas of "Longtermism" and "Existential Risk."](#)" *Current Affairs*. 28 July.

Troy, David. 2023. [The Wide Angle: Understanding TESCREAL — the Weird Ideologies Behind Silicon Valley's Rightward Turn.](#) *The Washington Spectator*. 1 May.

Varanasi, Lakshmi. 2023. "[OpenAI's Sam Altman is the latest tech entrepreneur making a play to extend the human lifespan.](#)" *www.businessinsider.com*. 8 March.

World Transhumanist Association. 2009. *The Transhumanist Declaration*.

Wynroe, Keith, David Atkinson and Jaime Sevilla. 2023. "[Literature review of Transformative Artificial Intelligence timelines.](#)" *EPOCHAI.org*. 7 January.

Yannick, Fritz. 2023. "[Philosophy Against the Present: The Foundations and Critique of Longtermism.](#)" *UMBAU*. 11 May.

Zaitchik, Alexander. 2022. "[The Heavy Price of Longtermism.](#)" *The New Republic*. 24 October.

de Zwart, Hans. 2022. "[Beware of 'Effective Altruism' and 'Longtermism'.](#)" *Racism and Technology Center*. 28 October.

REFERENCES - ATTACHMENT F.3

- AI Now 2017 Report*. 2017. Alex Campolo, Madelyn Sanfilipp, Mededith Whittaker, Kate Crawford.
- Alter, Adam. 2017. *Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked*. New York: Pegasus Press.
- Ambasna-Jones, Marc. 2018. "Winter is coming for AI. Fortunately, non-sci-fi definitions are actually doing worthwhile stuff." *The Register.co.uk*. 8 February.
- AAAI Presidential Panel on Long-term AI Futures. 2009. [Interim Report from the Panel Chairs](#). American Association for the Advancement of Artificial Intelligence. August.
- Armstrong, Stuart. 2014. *Smarter than Us: The Rise of Machine Intelligence*. Berkeley: Machine Intelligence Research Institute.
- Azarian, Bobby. 2016. "The Myth of Sentient Machines." *Psychology Today*. 1 June.
- Barrat, James. 2016. *Our Final Invention: artificial intelligence and the end of the human era*. New York: St. Martin's Press.
- Bogost, Ian. 2015. "[The Cathedral of Computation](#)". *The Atlantic*. 15 January.
- Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
- Bostrom, Nick. 1998. [How Long Until Superintelligence?](#) *International Journal of Future Studies*. Vol 2.
- Brockman, John. 2015. *What to Think about Machines that Think*. Edge publications.
- Bain, Marshall. 2015. *The Second Intelligent Species: how humans will become as irrelevant as cockroaches*. BYG Publishing.
- Brynjolfsson, Eric and Andrew McAfee. 2016. *Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York: W. W. Norton & Co.
- Carr, Nicholas. 2016. *Utopia is Creepy and other provocations*. New York: W. W. Norton & Co.
- Carr, Nicholas. 2014. *Glass Cage*. New York: W. W. Norton & Co.
- Carr, Nicholas. 2010. *The Shallows: What the Internet is Doing to Our Brains*. New York: W. W. Norton & Co.
- Caudill, Maureen and Charles Butler. 1992. *Understanding Neural Networks: computer explorations*. Cambridge: The MIT Press.
- Chase, Calum. 2015. *Surviving AI: The Promise and Peril of Artificial Intelligence*. Three Cs Publishing.
- Chase, Calum. 2016. *The Economic Singularity: Artificial intelligence and the death of capitalism*. Three Cs Publishing.
- Chang, Emily. 2018. *Brotopia: Breaking Up the Boys' Club of Silicon Valley*. New York: Penguin Random House.
- Chollet, Francois. 2017. Prediction: there will not be a real AI Winter... [Twitter post](#), 13 December.
- Church, George and Ed Regis. 2014. *Regenesis: How Synthetic Biology Will Reinvent Nature and Ourselves*. New York: Basic Books.
- Cohen, Noam. Noam Cohen. 2017. *The Know-It-Alls: The Rise of Silicon Valley as a Political Powerhouse and Social Wrecking Ball*. New York: The New Press.

Del Monte, Louis. 2014. *The Artificial Intelligence Revolution: Will Artificial Intelligence Serve Us or Replace Us?* Louis Del Monte.

Delcker, Janosch. 2018. "[Europe divided over robot 'personhood.'](#)" Politico.eu. 11 April.

Diamandis, Peter. 2014. *Abundance: The Future Is Better Than You Think* Paperback. New York: The Free Press.

Dietterich, Thomas G. and Eric J. Horvitz. 2015. [Rise of Concerns about AI: Reflections and Direction.](#) *Communications of the ACM*. October.

Dreyfus, Hubert. 1965. "[Alchemy and Artificial intelligence](#)" Rand Co. Free PDF.

Dreyfus, Hubert. 1972. *What Computers Can't Do*. New York: HarperCollins.

Dreyfus, Hubert. 1975. *What Computers Still Can't Do*. Cambridge: MIT Press.

eMarketer (IBM). 2017. *Artificial Intelligence: What's Now, What's New and What's Next*. May.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Executive Office of the President. 2016. *Preparing for the Future of Artificial Intelligence*. October.

Foer, Franklin. 2017. *World without Mind: the existential threat of Big Tech*. New York: Penguin Press.

Ford, Martin. 2016. *Rise of the Robots: Technology and the Threat of a Jobless Future*. New York: Basic Books.

Future of Humanity Institute. 2018. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. PDF.

Galloway, Scott. 2017. *The Four: The Hidden DNA of Amazon, Apple, Facebook, and Google*. New York: Portfolio/Penguin.

Halpern, Sue. 2016a. "They Have, Right Now, Another You." *The New York Review of Books*. 27 December.

Halpern, Sue. 2016b. "Our Driverless Future." *The New York Review of Books*. 24 November.

Halpern, Sue. 2015. "How Robots & Algorithms Are Taking Over." *The New York Review of Books*. 2 April.

Halpern, Sue. 2014. "The Creepy New Wave of the internet." *The New York Review of Books*. 20 November.

Halpern, Sue. 2011. "Mind Control. & the Internet." *The New York Review of Books*. 23 June.

Hamblin, James. 2014. "[But What Does the End of Humanity Mean For Me?](#)" *The Atlantic*. 9 May.

Hanson, Robin. 2016. *The Age of Em: Work, Love and Life When Robots Rule the World*. Oxford: Oxford University Press.

Harari, Yuval. 2017. *Homo Deus: A Brief History of Tomorrow*. New York: HarperCollins.

Hebb, Donald. 1949. [The Organization of Behavior](#). New York: Wiley.

Johnsson, Lars. 2016. "The Inconvenient Truth about David Star Jordan." *Palo Alto Online*. 19 February.

Kaplan, Jeffrey. 2016. *Artificial Intelligence: What Everyone Needs to Know*. Oxford: Oxford University Press.

Kaplan, Jeffrey. 2017. "[AR's PR Problem.](#)" *MIT Technology Review*. 25 April.

Kardara, Nicholas. 2017. *Glow Kids: How Screen Addiction Is Hijacking Our Kids - and How to Break the Trance*. New York: St. Martin's Press.

Kassan, Peter. 2006. "[AI Gone Awry: The Futile Quest for Artificial Intelligence](#)". Blog post.

Keep, Elmo. 2016. "The Strange and Conflicting World Views of Silicon Valley Billionaire Peter Theil." *Splinter News*. 22 June.

Kelly, Kevin. 2017. "[They Myth of a Strong AI](#)" *Wired*. April.

Kelly, Kevin. 2011. *What Technology Wants*. New York: Viking Press.

Kelly, Kevin. 1994. *Out of Control: The New Biology of Machines, Social Systems and the Economic World*.

Kelly, Kevin. 2017. *The Inevitable: Understanding the 12 Technological Forces That Will Shape Our Future*. New York: Viking.

Kersting, Thomas. 2016. [Disconnected: How To Reconnect Our Digitally Distracted Kids](#).

King, Bret. 2016. *Augmented: Life in the Smart Lane*. Marshall Cavendish Editions.

Kurzweil, Ray. 1999. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. New York: Penguin.

Kurzweil, Ray. 2006. *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking.

Kurzweil, Ray. 2012. *How to Create a Mind: The Secret of Human Thought Revealed*. New York: Viking.

Larson, Erik. 2015. [Questioning the Hype on Artificial intelligence](#). *The Atlantic*. 14 May.

[Lanier, Jaron](#). 2014. "Enthusiasts and Skeptics Debate Artificial Intelligence." *Vanity Fair*. November.

Lanier, Jaron. 2017. *Dawn of the New Everything: Encounters with Reality and Virtual Reality*. New York: Henry Holt & Co.

Lee, Kai-Fu. 2017. "[The Real Threat of Artificial Intelligence](#)." *The New York Times*. 24 June.

Malik, Om. 2016. "[The Hype and Hope of AI](#)". *The New Yorker*. 28 August.

Marcus, Gary. 2013. "[Hyping Artificial Intelligence, Once Again](#)." *The New Yorker*. 31 December.

Markoff, John. 2015. *Machines of Loving Grace: The Quest for Common Ground between Humans and Robots*. New York: HarperCollins.

Markou, Robert. 2017. "[Robots and AI could soon have feelings, hopes and rights ... we must prepare for the reckoning](#)." *Yahoo.news.com*. 24 February.

Martinez, Antonio. 2016. *Chaos Monkeys: Obscene Fortune and Random Failure in Silicon Valley*. New York: HarperCollins.

Morozov, Evengy. 2013. *To Save Everything, Click Here. The Folly of Technological Solutionism*. New York: PublicAffairs.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

Nicolelis, Miguel (2017) cited in Solon, Oivia (2017). "Elon Musk says humans must become cyborgs to stay relevant. Is he right?" *The Guardian*. 15 February.

O'Neill, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Penguin Random House.

Philippiadis, Alex. 2018. "Startup Cofounded by George Church Marries Low-Cost Sequencing with Blockchain." *Genetic Engineering and Biotechnology News*. 8 February.

Rees, Marin. 2003. *Our Final Century: Will the Human Race Survive the Twenty-first Century?* New York: Basic Books.

Russell, Stuart and Peter Norvig. 2011 (3rd ed). *Artificial Intelligence: A Modern Approach*. Essex, UK: Pearson.

Somers, James. 2017. "Is AI Riding a One-Trick Pony?" *MIT Technology Review*. 29 September.

Stone, Peter, and Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee Saxenian, Julie Shah, Milind Tambe, and Astro Teller. 2016. "Artificial Intelligence and Life in 2030." One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University, Stanford, CA. September. Doc: <http://ai100.stanford.edu/2016-report>.

Tapin, Jonathan. 2017. *Move Fast and Break Things: How Facebook, Google, and Amazon Cornered Culture and Undermined Democracy*. New York: Little, Brown & Co.

Tegmark, Max. 2017. *Life 3.0: Being human in the age of artificial intelligence*. New York: Alfred P. Knopf.

Turing, Alan M. 1950. "Computing Machinery and Intelligence". *Mind* 49: 433-460.

Turkle, Sherry. 2011. *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.

Twenge, Jean. 2017. *iGen: Why Today's Super-Connected Kids Are Growing Up Less Rebellious, More Tolerant, Less Happy--and Completely Unprepared for Adulthood--and What That Means for the Rest of Us*. New York: Simon & Schuster.

Vaidhyanathan, Siva. 2018. *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy*. Oxford: Oxford University Press.

Wachter-Boettcher, Sara. 2017. *Technically Wrong: Sexist Apps, Biased Algorithms, and other Threats of Toxic Tech*. New York: W. W. Norton & Co.

Wolfe, Alexandra. 2017. *Valley of the Gods: A Silicon Valley Story*. New York: Simon & Schuster.

#

F.3 DISCUSSION NOTES [in progress]

¹ In these notes I use “artificial intelligence” and “AI” to include robotics and associated technologies.

² Test versions of autonomous cars began driving regularly on open roads in 2013. Until now (May 2018) there have been 4 fatalities, 1 in China and 3 in the U.S. Three were driver fatalities and 1 was a pedestrian fatality. Three involved Tesla vehicles and one an Uber vehicle. There is a wide range of opinions as to how long it will be before the balance of risk and benefit allows widespread general use of autonomous vehicles. Some predict several years and others predict several to many decades. It’s likely that autonomous vehicle use will be incremental. It will be allowed on restricted terrain in controlled and monitored situations, e.g. vacation resorts and military bases, before being more generally allowed. See Halpern (2016b).

³ This background memo was prepared in May 2018. It’s based almost entirely on web/journal/library research. I had relatively little prior familiarity with the recent controversy over AI. I haven’t yet spoken with any knowledgeable people working in the field of AI. I hope to do so as part of the preparation of the full working paper. I expect that those conversations will identify errors of omission, commission and interpretation in this memo.

⁴ The current memo focuses largely on developments in the United States and the United Kingdom, where the AI+ network has been most active. A fully useful background memo would include developments in all countries, and I hope to survey those prior to completion of the final working paper.

⁵ I use “AI+” (read “AI plus”) to distinguish these ideologically-motivated superintelligence/posthumanist/singularitarian advocates from the mainstream corporate and academic AI advocates and the progressive AI advocates. The “+” notation follows the naming convention of the major transhumanist organization, “Humanity +” or “H+”. The leading AI+ advocates are further motivated by an expectation of achieving personal immortality, and believe that humanity will, possibly very soon, be replaced by or transformed into some higher artificial superintelligent technological entity.

⁶ Many would argue that *if* AGI could be developed, it *shouldn’t* be, that is, that its development should be banned. But Musk has announced that he is in fact developing AGI, and there has been no protest. Further, he’s announced that he will make the AGI code publicly available, i.e. open source. This might be likened to developing a technique for constructing hydrogen bombs in a suburban garage and sending starter kits to everyone in the world over age 14. The absence of protest may be due to the reasonable assessment that the objective is unattainable. Alternatively, the lack of protest may be due to the fact that vanishingly few people fully understand what any of this is about, and those that do either support Musk’s initiative or feel isolated and helpless when trying to think of what they can do to stop it. OpenAI is intended to produce shock and awe in those who might be in any position to challenge in some effective manner the forward development of AI by the AI+ advocates.

⁷ [Further analysis of motivation, interests and strategy is in preparation.]

⁸ It’s well known that differences in wording, methodology, sample size and selection, and other factors, can generate very different responses to seemingly similar survey questions. The review in Attachment R, and the summary items, should be taken as provisional pending more careful review.

⁹ Truly accidental harm is less likely than is harm inflicted as a result of human intention. Most advanced technologies are intricate. They need to be designed with multiple inter-related features to accomplish their intended purpose. It’s unlikely that an AI-empowered municipal sanitation vehicle (for example) would *accidentally* charge off into the desert and start mining uranium with which to construct a nuclear weapon. And inherently risky devices, such as an AI-empowered nuclear missile submarine, have multiple layers of safety features

engineered into them. Truly catastrophic risk comes from devices *intentionally engineered* to do catastrophic harm, and that include engineered features intended to *facilitate* activation and to thwart counter-measures.

¹⁰ An analogy might be the practice of smoking tobacco. This was done for centuries with little knowledge of its harmful effects. Even after its potential risk was suspected and then confirmed it took concerted effort before people began to abandon the practice.

¹¹ An “erosion of the sense of human autonomy and dignity that allows human flourishing” is of consequence because it would open the door to, among other harms, a world shaped and dominated by AI technologies and those who own and control them. Although it would not be a world of AI+ superintelligent/singularitarian/posthumanism, it could still be an AI-dominated transhumanist world.

¹² Two levels of restriction that we will need to become comfortable with imposing are:

a) **limits on research:** It’s frequently said that we needn’t and shouldn’t impose limits on research, as the generation of new knowledge is an unambiguous good; rather, the proper focus of any limits, proscriptions or bans should be on *applications*, which, if wrongly or poorly deployed, could cause real harm. But this argument is blinkered. The research enterprise brings into being a constituency of researchers, funders, potential direct and ancillary beneficiaries and of course potential commercial beneficiaries. The often lengthy and expensive process of research requires multiple parties to become passionately committed to the potential beneficial applications that the research might allow and to play down the risks. By the time a research initiative has produced the sought procedure, datum, molecule or other result, its application has almost always been long thought-through and planned for. We will need to learn to draw lines much further upstream if we want to ensure against strongly unwanted outcomes.

b) **constitutional amendments or equivalent:** technologies that are judged to be undesirable should be regulated or proscribed at the appropriate level. Some should come under administrative regulation and be subject to fairly easy modification. Others should be prohibited by law enacted by national legislatures. The most potentially dangerous technologies will likely need to be proscribed via constitutional amendment. [Further discussion in process.]

¹³ There is of course a wide-ranging debate over whether a machine can *ever* become conscious. I believe the strongest case is that consciousness is an embodied, evolved feature of animal life. See Attachment V of this memo on the push-back against the notion of AI+ superintelligence.

¹⁴ These definitions are taken and modified from Kaplan (2016), eMarketeer (2017), Russell and Norvig (2011) and Wikipedia and other websites.

¹⁵ He argues strongly that superintelligence, and AI and AGI as well, are *substrate independent*, that is, they could operate on “a digital computer, an ensemble of networked computers, cultured cortical tissue or what have you.” Others differ strongly, and there are a wide range of opinions about this. Substrate independence is taken as axiomatic among AI+ advocates, for obvious reasons. But there is absolutely no evidence, and there is absolutely no reason to believe, that substrate independence is the case. The notion is an example of what has been called “Potemkin science” or “cargo cult science,” i.e. arguments dressed up to look and sound like science but without the chain of arguments supported by evidence that leads from initial statements of fact to reasonable conclusions.

¹⁶ [IBM eMarketing](#) (2017).

¹⁷ Canadian psychologist Donald Hebb (1949) first proposed a form of neural network learning. At U.C. Berkeley in the mid-1990s I learned to construct early textbook-level computational neural networks. The text was Caudill and Butler (1994), *Understanding Neural Networks: computer explorations*.

¹⁸ Rejected surveys included those used for promotional advertising, as well as those with no description of methodology and/or poorly worded questions.

¹⁹ Consider two scenarios, an outside or rogue scenario and an inside or establishment scenario. [in preparation].

²⁰ I can remember when in the early 1960s U.C. Berkeley students first revised and repurposed, as a placard-and-button slogan, the cautionary instruction routinely printed on the then-ubiquitous IBM punch-cards: *I am a human being. Do not fold, spindle or mutilate.*

²¹AI+ advocates have three responses. Many, notably *eliminativists* like Dennet and the Churchlands, say that consciousness is unnecessary for intelligence because it's obvious that consciousness doesn't exist in the first place, and humans are intelligent. Many others argue that the question of consciousness is *irrelevant*. It's impossible to know if anyone or anything other than oneself is conscious. All a computer needs to be able to do is to simulate human intelligence to the point that it will fool a human in a Turing test; that's all we can ever expect so far as indicators of consciousness are concerned, and all that we need. Others say that when a computer can fully simulate human intelligence that's evidence that it *is in fact* conscious. This is the argument from *substrate independence* (see Endnote 14, above). It's frequently invoked by AI+ advocates but as noted there is no reason at all to suppose that it's true, there's no test that can prove it to be so, and an exercise applying Occam's razor would suggest that it is not so.

Addendum C - Discussion Notes and Citations

²² Lewis-Kraus (2022a, p48)

²³ Selected critical commentary on EA prepared prior to mid-2018 is shown in ATTACHMENT S.2.3. Additional, more recent critiques of EA are referenced in this timeline. Other useful recent critiques include Adams et al. (2023), Torres (2023), Szalaf (2022), Lowrey (2022), Robinson (2022) and Táíwò and Stein (2022).

²⁴ It's important to reiterate that the strong and now dominating EA emphasis on avoiding existential risk and the possibility of human extinction, especially through the action of malevolent AI+, is not the result of EA opposition to AI+ itself in the manner that, say, environmentalists oppose climate change or biodiversity loss. Rather, EAs look forward to a planet, and a universe, totally transformed by the hyper-technologies of AI+, human genetic modification, space colonization and the rest. They believe this is the path to unimaginably greater good for unimaginably greater numbers of people. They call for controls on AI+ not because they want to prevent it from being developed and deployed, but because they don't want catastrophic accidents or intentionally harmful use to motivate reactions and measures that would shut their enterprise down for good.

²⁵ The present-day 'rationalist' movement is largely a movement of atheist scientific libertarian technocrats. EA developed, in part, as a response to the criticism that the rationalist mindset lacks the compassion and charity exhibited by religion and by both classical and progressive liberalism. In effect, the EAs are saying that not only is their compassion and charity more effective than what religion and liberalism offer, EAs are more fervently compassionate and charitable in the first place. For more on contemporary 'rationalism' see Wallace (2021). One qualifier: if popular protest or anything else appears to be putting their grand universal project – effectively, transhumanism + space colonization – at risk, EAs are able to quickly drop their libertarian vibe and remake themselves as benevolent techno-authoritarians.

²⁶ For examples of Bostrom's utilitarian consequentialist calculus see:

- * Bostrom (2003), in which he argues that we lose at least one hundred trillion (10^{14}) potential human lives for every second that we delay expanding into space.
- * Bostrom (2002), in which he derives the value of 10^{52} potential human lives un-lived if humanity goes extinct before having developed friendly AI+.

²⁷ See U.S. Security and Exchange Commission (2023); Excerpt and link are shown in ATTACHMENT B below.

²⁸ See Lewis-Kraus (2022b) and Alter (2023).

²⁹ See Altraide (2023) and Goldstein et al. (2022) for details on the set-up and operation of SBF’s scams. See too the [Comments](#) on the Goldstein et al. piece that call out the financial press for their kid-gloves coverage of SBF.

³⁰ Ehrlich (2021)

³¹ Lewis-Kraus (2022a, p56)

³² Legraien, Lea (2023)

³³ Lewis-Kraus (2022a, p56-58)

³⁴ Wiblin and Harris (2022)

³⁵ After noting the rarefied money-and-celebrity quotient of the event, the digital currency news site *Coindesk* reported (Wang, 2022),

“But perhaps the biggest celebrity of all was FTX founder Sam Bankman-Fried, who was met with cheers and applause during his several on-stage appearances. If anything, the event was a testament to his crypto empire’s growing might and an extended opportunity to show off his unconventional star power.”

For an only slightly less star-struck account of Crypto Bahamas and SBF’s ascending clout , see Yaffe-Belany (2022).

³⁶ Alter (2023)

³⁷ Musk (2022) tweet thread

³⁸ Fleck (2022)

³⁹ For more on the rise and collapse of FTX and SBF see Douglas and Fountain (2022).

⁴⁰ See Alter (2023) and Lewis-Kraus (2022b)

⁴¹ MacAskill (2022) tweet thread

⁴² See both U.S. Attorney’s Office – Southern District of New York (2022) and U.S. Security and Exchange Commission (2023) for the full set of criminal charges and civil complaints against SBF. Excerpts and links are shown in Attachments A and B below.

Addendum D - Discussion Notes and Citations

⁴³ For more fully detailed timelines of the development of transhumanism see a) [Timeline of Transhumanism](#); and b) [A Timeline of Transhumanism](#).

⁴⁴ Bostrom (1997).

⁴⁵ Bostrom (2002) defines existential risk as one “...where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential.” Most people might agree that large asteroid impacts, “spasm” nuclear warfare, genetically modified super-pathogens and climate change driven by strong positive feedback loops could pose existential risks. Bostrom and other longtermists believe that the greatest existential risk is posed by powerful AIs that remain “unaligned with human values” and then slip from human control. Some celebrity AI scientists say they share this belief but the great majority of AI scientists most likely do not. See Working Paper [ATTACHMENT F3, pp 73-76](#).

⁴⁶ Bostrom’s definition of existential risk was carefully crafted. If humanity develops truly intelligent computers capable of defending and sustaining themselves under the most adverse conditions, then a catastrophic event that destroyed all human life forever but did not destroy those computers would not have been an *existential* catastrophe. Further: those who argue that we should not aspire to uploading our minds onto nanoscale

spacecraft and then colonizing the galaxies are placing humanity in a state of existential risk, because if they prevail they will have drastically curtailed its potential.

⁴⁷ Longtermist critic Émile Torres (2023a) turns Bostrom's analysis of existential risk on its head. Torres notes that longtermists insist on the priority of developing strong AIs, albeit friendly ones, and then using them to become omniscient, live forever, colonize the galaxies, etc. But, says Torres, an ideological commitment to this ludicrous scenario is more likely to leave us in ruins, or worse, than it is to succeed. Thus, he says, "One of the existential risks to the future is longtermism itself."

⁴⁸ Dvorsky (2022).

⁴⁹ [The Transhumanist Declaration](#) (2009).

⁵⁰ Presumably Bostrom and MacAskill had met earlier and at some point realized an alignment of interests and strategies. I haven't found a record of this but I expect that one exists. Both the transhumanists and the effective altruists see themselves at the pivot of human history and they document their goings-on meticulously.

⁵¹ See my account of the early years of EA in Working Paper [ATTACHMENT F3, pp 62-66](#).

⁵² See Bostrom (2014). He defines superintelligence as "... any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest." He considers five paths over which superintelligence might be realized: 1) Artificial Intelligence (i.e. machine intelligence, probably digital); 2) Whole brain emulation (synapse-by-synapse replication of a brain *in silico*); 3) Biological cognition (through genetic enhancement and drugs); 4) Brain-computer interfaces; and 5) societal networks and organizations. He says that each has advantages and disadvantages, and that all might have some contribution to make, but concludes that AI is the likely most expeditious path and should be prioritized.

⁵³ Bostrom main case is, first, that the best possible world is one in which human minds merge with superintelligent AIs and together quickly realize omniscience, omnipotence and immortality, and then colonize the galaxies; and second, that in the pivotal early stages (now) it's possible that this best possible world could be denied us if either a) powerful AIs slip from our control and destroy us or b) the fearful masses who don't want to merge with AIs destroy *them*. To avoid these unhappy fates the priority now is to a) develop "safe" AI that won't slip from our control (or if it does, at least won't want to destroy us); and b) keep the pitchforks from mobilizing.

⁵⁴ Bostrom doesn't appear to use the word "transhuman-" in his recent writings but neither does he deny the motivating philosophy or the label. [His website](#) displays his early transhumanist writing, and his November 2015 *New Yorker* profile story called him "...arguably the leading transhumanist philosopher today."

⁵⁵ For a detailed timeline over this 2013-2017 wave of AI attention see Working Paper [ATTACHMENT F3, pp 6-7](#).

⁵⁶ For the Agenda, list of participants and full set of pdfs of presentations given at the Puerto Rico gathering see Working Paper [ATTACHMENT F3, pp 27-28](#).

⁵⁷ See Matthews (2015). He says the plenary on existential risk and AI was "...the most hyped event at EA Global."

⁵⁸ Artificial General Intelligence (AGI) is AI capable of doing everything a human can do, including writing improved computer programs. Enthusiasts like Bostrom argue that the first AGI would last only a few days, at most, before it developed an improved AGI, which would make greater improvements even more rapidly, and so on. Further, Bostrom says the very first AGI would quickly disable all other AGIs under development and declare itself 'The Singleton,' and would go on to control the Universe. Some identify this scenario with The Singularity.

⁵⁹ Many, perhaps most, hands-on AI scientists dismiss AGI as a fantasy. The Big Data + Pattern Recognition approach being used by OpenAI and other firms to supposedly develop AGI is simply not how the human mind works and is orders of magnitude too crude to even simulate creative higher-level thought. That's not to deny its potential for innovative and powerful capabilities that could be used for either great good or great harm.

⁶⁰ For the Agenda, list of participants, and full set of pdfs and videos of the presentations see Working Paper [ATTACHMENT F3, pp 33-40](#). The Asilomar conference center is famous in science and technology circles as the site of the 1975 conference at which leading molecular biologists, many associated with the nascent biotech industry, gathered to agree on strategy for defusing public and governmental anxiety over proposed research and applications involving the use of recombinant DNA.

⁶¹ See MacAskill (2019). He says that in October 2017 Toby Ord was completing the manuscript for *Precipice* and realized he didn't have a good identifier for the growing numbers of people who were becoming concerned about existential risk reduction; after some reflection MacAskill proposed *longtermism* and it was adopted.

⁶² MacAskill says he was initially skeptical about existential risk as a proper priority ethical concern. He credits work by Open Philanthropy senior researcher Ajeya Cotra on the high near-term probability of transformative AI being realized as importantly helping "bring him around" to a longtermist worldview. See Lewis-Kraus (2022).

⁶³ The [Global Priorities Institute](#) (GPI) says its vision is "A world in which global priorities are set by using evidence and reason to determine what will do the most good," and that its mission is "To conduct and promote world-class, foundational academic research on how most effectively to do good." Its research agenda has two major foci: 1) the "longtermist paradigm" and 2) "cause prioritization." It has 40+ staff, affiliates and scholars, most with backgrounds in philosophy or economics. One of GPI's earliest working papers was *The Case for Strong Longtermism* (2019) by GPI Director Hilary Greaves and MacAskill ([updated 2021](#)).

⁶⁴ The [Forethought Foundation](#) (FF) promotes studies on "... how to use our scarce resources to improve the world by as much as possible." Its three research areas are Longtermism, Mitigating Catastrophic Risk, and Affecting the Very Long Run. It has seven staff and apparently hosts 35-40 fellows annually for one-year research projects. It "works closely with" the Global Priorities Institute.

⁶⁵ See Davis (2023). The [Center for Security and Emerging Technology](#) (CSET) was founded by [Jason Matheny](#), former Research Director of Bostrom's Future of Humanity Institute at Oxford and now President and CEO of RAND Corporation. Matheny is hugely embedded in DC and global technology and foreign policy networks. He played a key role in shaping the Biden Administration's decision to impose unprecedentedly strong controls on semiconductor exports to China. His early paper *Reducing the Risk of Human Extinction* (2007) is a restatement of Bostrom's proto-longtermist arguments made in 2002-2003 and draws heavily on utilitarian/consequentialist (Peter Singer, Derek Parfit) and transhumanist (Eliezer Yudkowsky, Martin Rees) sources.

⁶⁶ Cotra (2020) follows the definition of *transformative AI* (TAI) presented by [Open Philanthropy](#) Co-CEO Holden Karnofsky. It is "AI powerful enough to bring us into a new, qualitatively different future." He cites the industrial and agricultural revolutions as examples of developments that generated qualitatively different futures. He notes that an AI can be transformative but still fall short of superintelligence or artificial general intelligence. See Karnofsky (2021, 2016).

⁶⁷ Cotra updated her TAI estimates in 2022. As of that year she foresaw, very roughly:

- 15% probability of TAI by 2030
- 50% probability by 2040
- 60% probability by 2050

⁶⁸ Wynroe et al. (2023) reviews ten alternative forecasts for the arrival of TAI. Each alternative relies upon differing methodologies and assumptions. The year at which TAI can be expected with 50-50 probability ranges from 2039 to later than 3000. Various averaging and weighting moves show TAI expected between 2045-2089.

⁶⁹ For criticism of and opposition to longtermism from a wide range of perspectives – social conservative, liberal, eco-activist, democratic socialist, anti-racist and post-modern left – see Linton (2023), McGoey (2023), Miller et al. (2023), Yannick (2023), Jacobson (2022), Hogan (2022), Naughton (2022), de Zwart (2022), Zaitchik (2022), and Chugg (2020). Especially useful is the extended critique by apostate transhumanist/longtermist Émile P. Torres; see e.g. his papers and interviews 2023a, 2023b, 2023c, 2022, 2021a, 2021b and his website [xriskology.com](#).

⁷⁰ The key critique from within EA was the December 2021 paper *Democratizing Risk* by Carla Zoe Cremer and Luke Kemp. At the time they were researchers at, respectively, Bostrom’s Future of Humanity Institute and the Center for Existential Risk at Cambridge. Their paper charged that the field of *existential risk studies* (ERS), as developed by Bostrom, Ord, MacAskill et al., and for which longtermism had become the nearly universal ethical and programmatic stance, with failures of definition, methodology, integrity, diversity, values and constituency pluralism, governance, democratic decision-making and more. The paper traced many of these failings to the nearly hermetic grounding of ERS in what Cremer and Kemp called the *techno-utopian approach* (TUA). See Cremer and Kemp (2021) as well as the podcast Cremer and Kemp (2023).

⁷¹ A noted critic of longtermism, EA and their ideological companions is computer scientist [Timnit Gebru](#), who until December 2020 served as technical co-lead of Google’s Ethical AI Team. At that time a dispute over a paper she co-authored that focused on dangers of large language models led to her contentious departure from Google. See Hao (2020). Gebru is now founding executive director of the [Distributed AI Research Institute](#) (DAIRI), a community-centered initiative focused on bringing diverse perspectives and deliberative methodologies to the production of needed, wanted and beneficial AI. She and colleague Émile Torres coined the acronym **TESCREAL**, standing for *transhumanism, extropianism, singularitarianism, cosmism, rationalism, effective altruism* and *longtermism*, to identify the overlapping ideologies spreading among hardcore techno-triumphalists in the U.S., U.K. and elsewhere. See Gebru and Torres (2023), Troy (2023) and Gebru (2022).

⁷² See Lewis-Kraus (2022). For these timeline notes I haven’t reported the extraordinary sums of investment and philanthropic capital that has been mobilized via the networks associated with EA, longtermism and their many allied and interlocking networks. I haven’t found a comprehensive compendium and analysis of the sources and recipients of these funds.

⁷³ The EA/longtermist infrastructure now includes dozens, if not scores, of institutes, centers, NGOs and philanthropies world-wide. I haven’t seen a comprehensive compilation of these and their interlocking leadership, staff, program and funders. One important hub is [Effective Ventures Foundation UK](#), the non-profit legal entity of which the following organizations are structured as “projects,” and which serves as a pass-through for funding, provides administrative, personnel and legal support, and more:

- | | |
|--|---|
| 1. Center for Effective Altruism (CEA) | 6. Center for the Governance of AI (CGAI) |
| 2. 80,000 Hours | 7. Longview Philanthropy |
| 3. Forethought Foundation for Global Priorities Research | 8. asterisk |
| 4. EA Funds | 9. Non-Trivial |
| 5. Giving What We Can | 10. BlueDot Impact |

⁷⁴ See Dvorsky (2022) and Addendum A (this notes).

⁷⁵ I had expected that the strong reaction to the release of ChatGPT-4 would generate stories in the press saying that the call by EA for priority attention to AI safety had now been vindicated. I haven’t as yet seen signs of this.

⁷⁶ See my timeline tracing the involvement of EA and SBF: ATTACHMENT F, S.2.6. Update - June 2023.

⁷⁷ McGoey (2023), BostromAnonAccount (2013).

⁷⁸ Both letters received considerable [press coverage](#) but, perhaps as expected or even intended, their calls were not immediately heeded. The first letter (22 March) was prepared by the [Future of Life Institute](#) in Cambridge MA, and as of June ’23 had 33,000 signatories. The second letter (30 May) was prepared by the [Center for AI Safety](#) in Berkeley, CA, and was signed by ~ 350 leaders in AI science/tech/policy and related fields. Both organizations are EA-embedded and funded by EA philanthropies. Such open letters primarily function to cast the AI-tech leadership as caring stewards of the human future, and thereby to forestall the emergence of any voices or initiatives actually representing affected communities and, especially, the human community as a whole.

⁷⁹ The Open Letter prepared by the Future of Life Institute (FLI) referenced a [series of policy recommendations](#) for consideration during the called-for 6-month pause. These policy recommendations are:

1. Mandate robust third-party auditing and certification.
2. Regulate access to computational power.

-
3. Establish capable AI agencies at the national level.
 4. Establish liability for AI-caused harms.
 5. Introduce measures to prevent and track AI model leaks.
 6. Expand technical AI safety research funding.
 7. Develop standards for identifying and managing AI-generated content and recommendations

Note that none of the proposed measures are explicit prohibitions on the uses to which AI is put. This stands in contrast to the measures adopted by the European Parliament in its Artificial Intelligence Act (see DN 38 following). Given the reasonable assumption that the proposed FLI policies would be developed and then administered by personnel from or aligned with the AI tech community, they would in fact benefit that community by rationalizing the current chaotic industry environment.

⁸⁰ The [Artificial Intelligence Act](#) (AIA) assigns AI applications one of three levels of risk: *limited*, *high* or *unacceptable*. Those judged to pose *unacceptable* risks are explicitly banned. These banned applications are:

1. Biometric categorization systems using sensitive characteristics (e.g. gender, race, ethnicity, citizenship status, religion, and political orientation)
2. Real-time remote biometric identification systems in publicly accessible spaces
3. Post-time remote biometric identification systems, with the only exception of law enforcement for the prosecution of serious crimes and only after judicial authorization
4. Predictive policing systems based on profiling, location or past criminal behavior
5. Emotion recognition systems in law enforcement, border management, the workplace, and education
6. Untargeted scraping of images obtained from the Internet in order to create facial recognition databases

See Liboreiro and Alonso (2023) and European Parliament News (2023). I haven't yet seen commentary from major stakeholders, analysts and others on perceived pros and cons of the AIA.

⁸¹ *Nature* magazine, editorial, 29 June 2023.

⁸² See the growing commitment to *immortalism* among elite tech billionaires in Varanasi (2023).