AI Strategy, Policy, and Governance

...

Allan Dafoe
Center for the Governance of AI
Future of Humanity Institute
University of Oxford

What is the Governance of AI?

Descriptive definition: The processes by which decisions are made and implemented. This includes norms, policies, institutions, and laws.


"
Whoever leads in AI will rule the world
Vladimir Putin
"
RT

What is the Governance of AI?

Descriptive definition: The processes by which decisions are made and implemented. This includes norms, policies, institutions, and laws.

Normative definition: A **good** set of such of such processes. Good governance usually means that it is effective, legitimate, inclusive, adaptive.

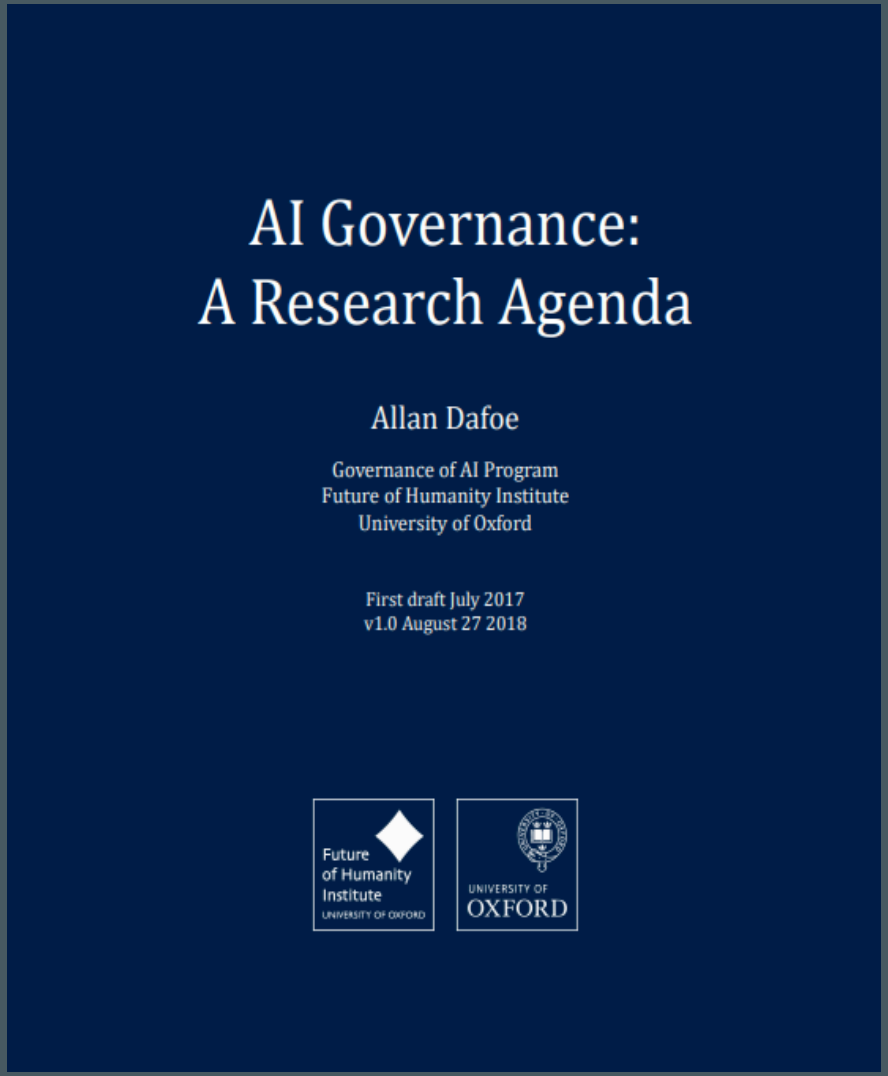
Governance of AI Will Not Be Easy

AI is a General Purpose Technology.

GPTs fundamentally transform economic, social, military processes, often in ways that are hard to govern.

Governance Properties of AI

- Diffuse harms and benefits
- High uncertainty
- Fast moving, dynamic problem
- Irreversible achievements
- Unclear responsibility
- Dual-use, broadly available
- Highly technical
- Competitive incentives



Technical landscape: capabilities, mapping, forecasting, safety

Politics: international geopolitics, domestic and mass politics, IPE, international security

Ideal Governance: values, principles, appealing positive visions, institutional design, norm building

Policy: translation of long-term goals into concrete near-term policy actions

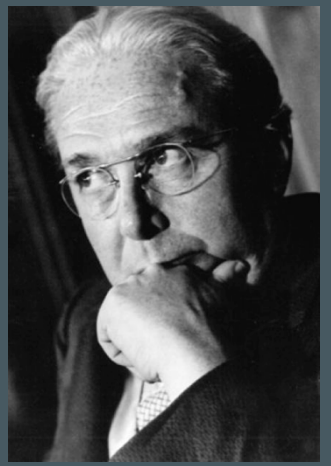
Scientific Conservatism and Policy Conservatism

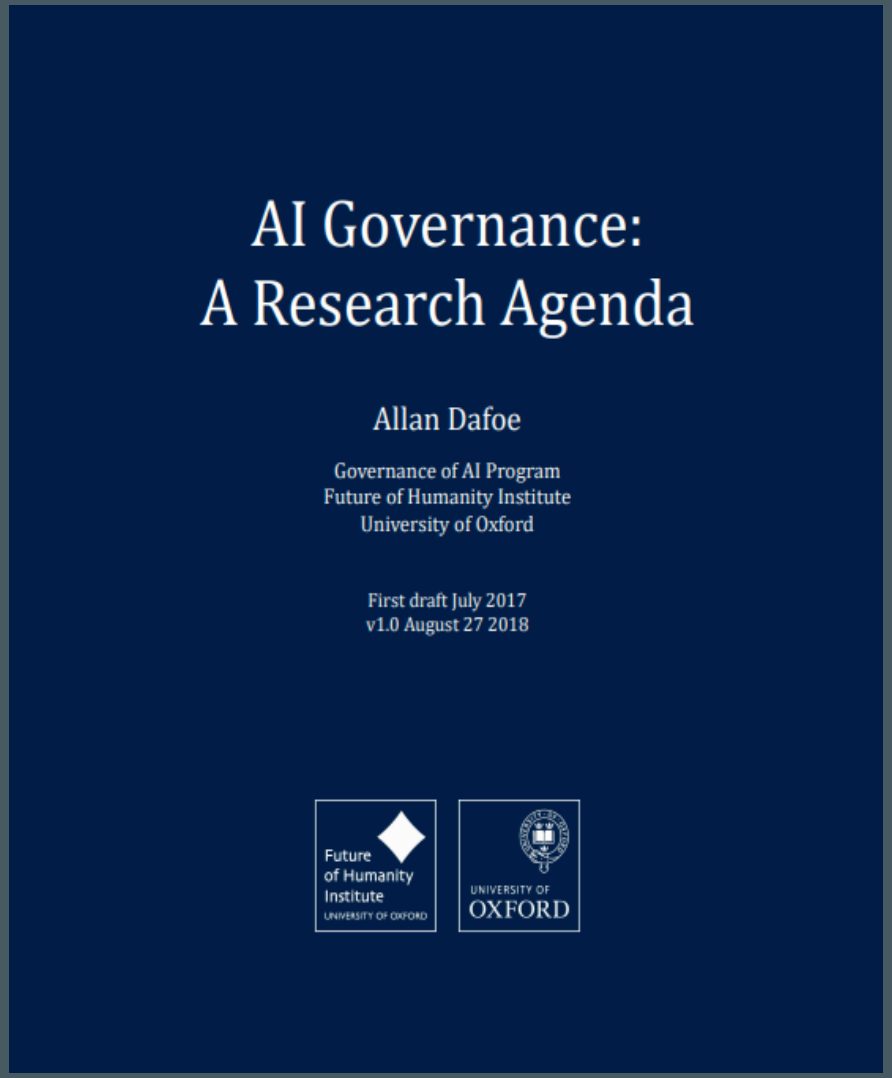
From the very beginning [1939] the line was drawn [...]

Fermi thought that the conservative thing was to play down [his 10%] possibility that [a nuclear chain reaction] may happen,

[Szilard] thought the conservative thing was to assume
that it would happen and take all the necessary precautions.

-Leo Szilard (quoted in 1978)





Technical landscape: capabilities, mapping, forecasting, safety

Politics: international geopolitics, domestic and mass politics, IPE, international security

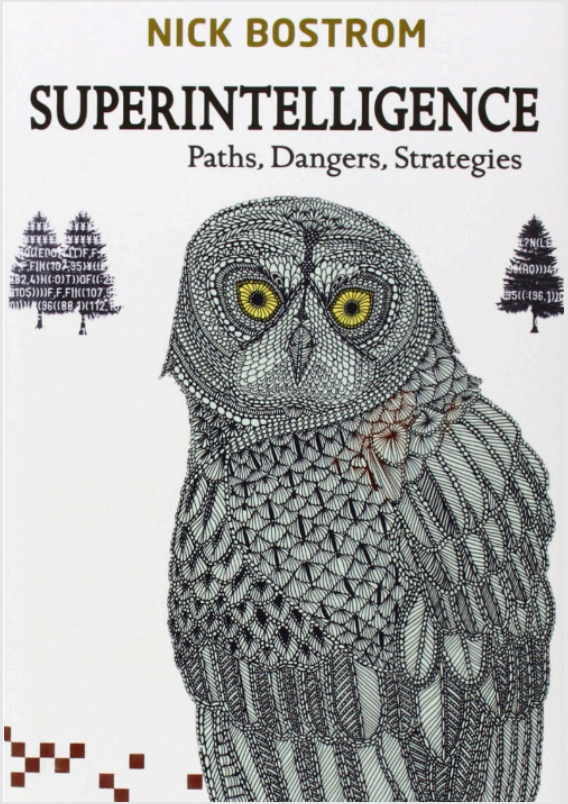
Ideal Governance: values, principles, appealing positive visions, institutional design, norm building

Policy: translation of long-term goals into concrete near-term policy actions

Technical Landscape

- Rapid and broad progress?
- Kinds, capabilities, and properties?
- Strategic properties of technology?
- Measuring inputs, capabilities, performance.
- Modeling AI progress
- Forecasting and indicators
- AI safety

Technical Landscape: Mapping



Political Implications of Crypto

[Ben Garfinkel]

Recent Developments in Cryptography and Possible Long-Run Consequences

Ben Garfinkel*

Abstract

Historically, progress in the field of cryptography has been enormously consequential. Over the past century, for instance, cryptographic discoveries have played a key role in a world war and made it possible to use the internet for business and private communication. In the interest of exploring the impact the field may have in the future, I consider a suite of more recent developments. My primary focus is on blockchain-based technologies (such as cryptocurrencies and smart contracts) and on techniques for computing on confidential data (such as homomorphic encryption and secure multiparty computation). I provide an introduction to these technologies that assumes no previous knowledge of cryptography. Then, I consider eight speculative predictions about the long-term consequences these emerging technologies could have. These predictions include the views that a growing number of information channels used to conduct surveillance may "go dark," that it may become easier to verify compliance with agreements without intrusive monitoring, that the roles of a number of centralized institutions ranging from banks to voting authorities may shrink, and that new transnational institutions known as "decentralized autonomous organizations" may emerge. Finally, I close by discussing some challenges that could limit the significance of emerging cryptographic technologies. On the basis of these challenges, it is premature to predict that any of them will approach the transformativeness of previous technologies. However, this remains a rapidly-developing area well worth following.[1]

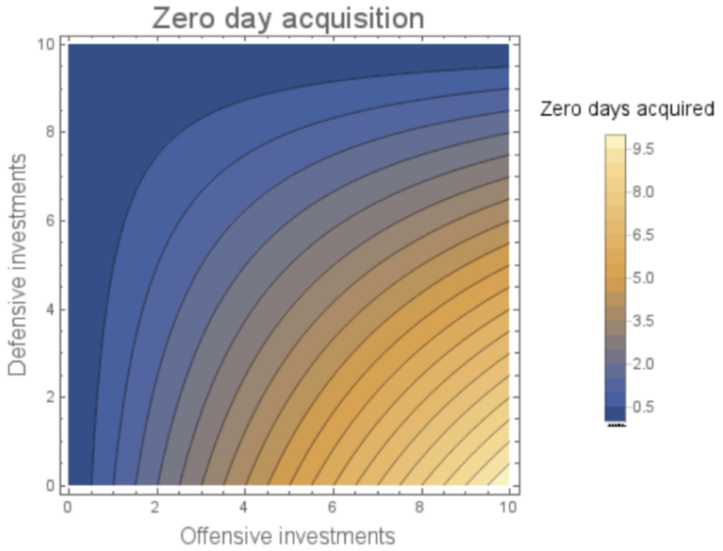
Strategic properties of artificial intelligen

[Ben Garfinkel, Allan Dafoe]

How Does the Offense-Defense Balance Scale? *

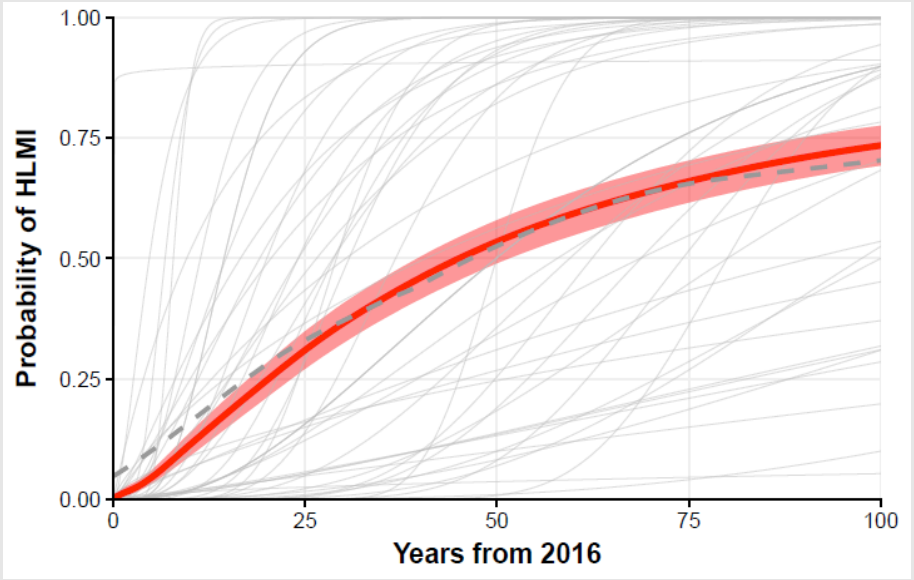
Ben Garfinkel, Allan Dafoe†

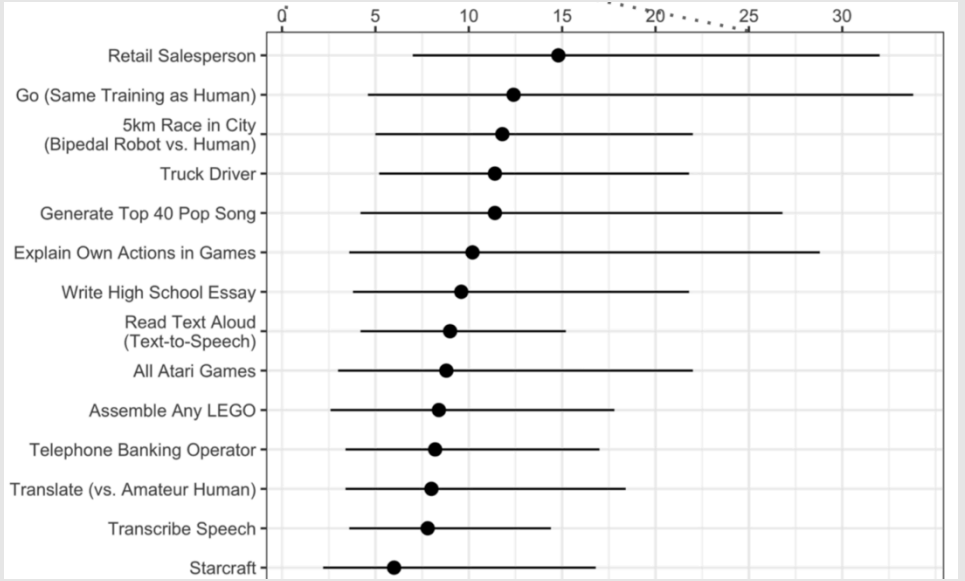
May 15, 2018

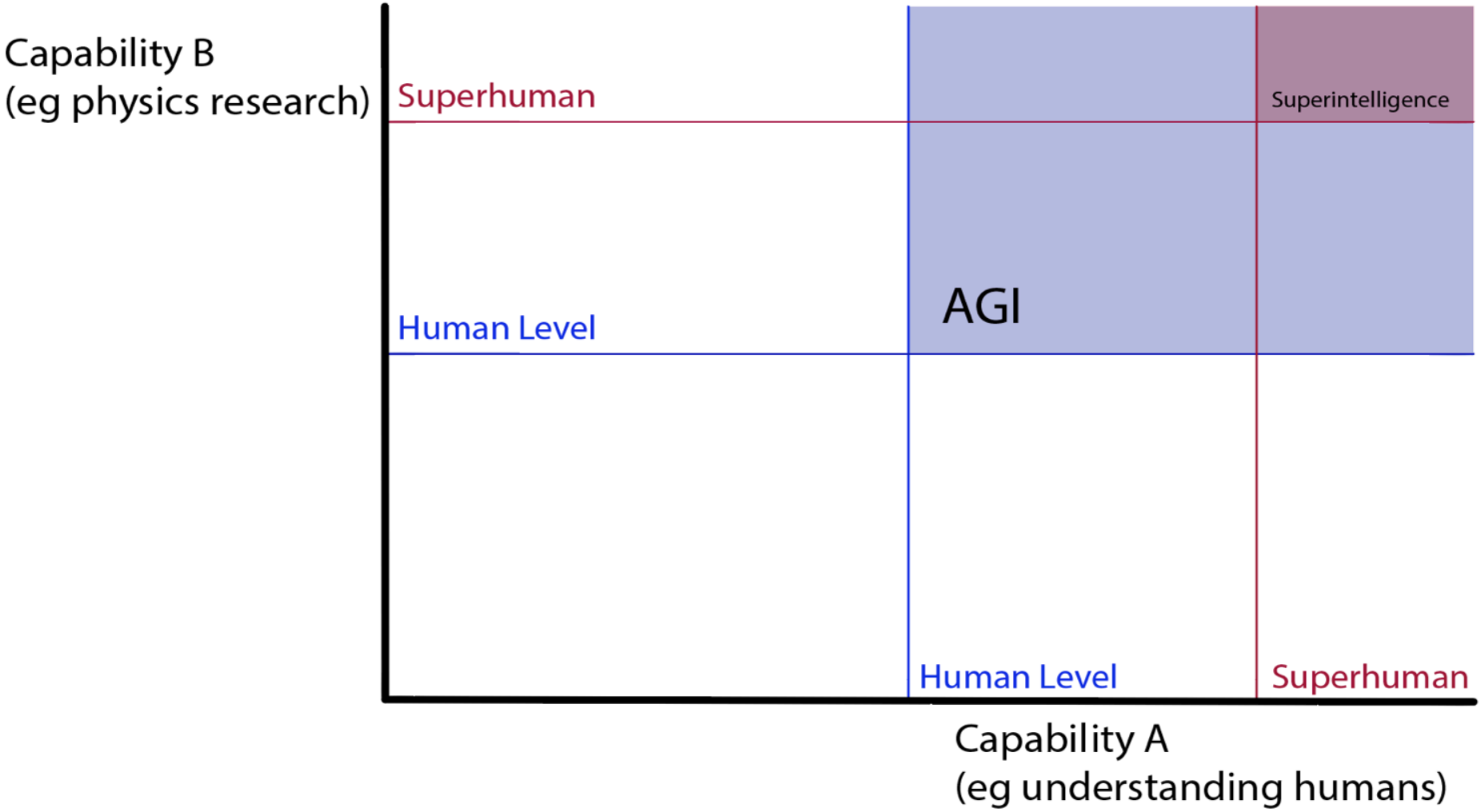


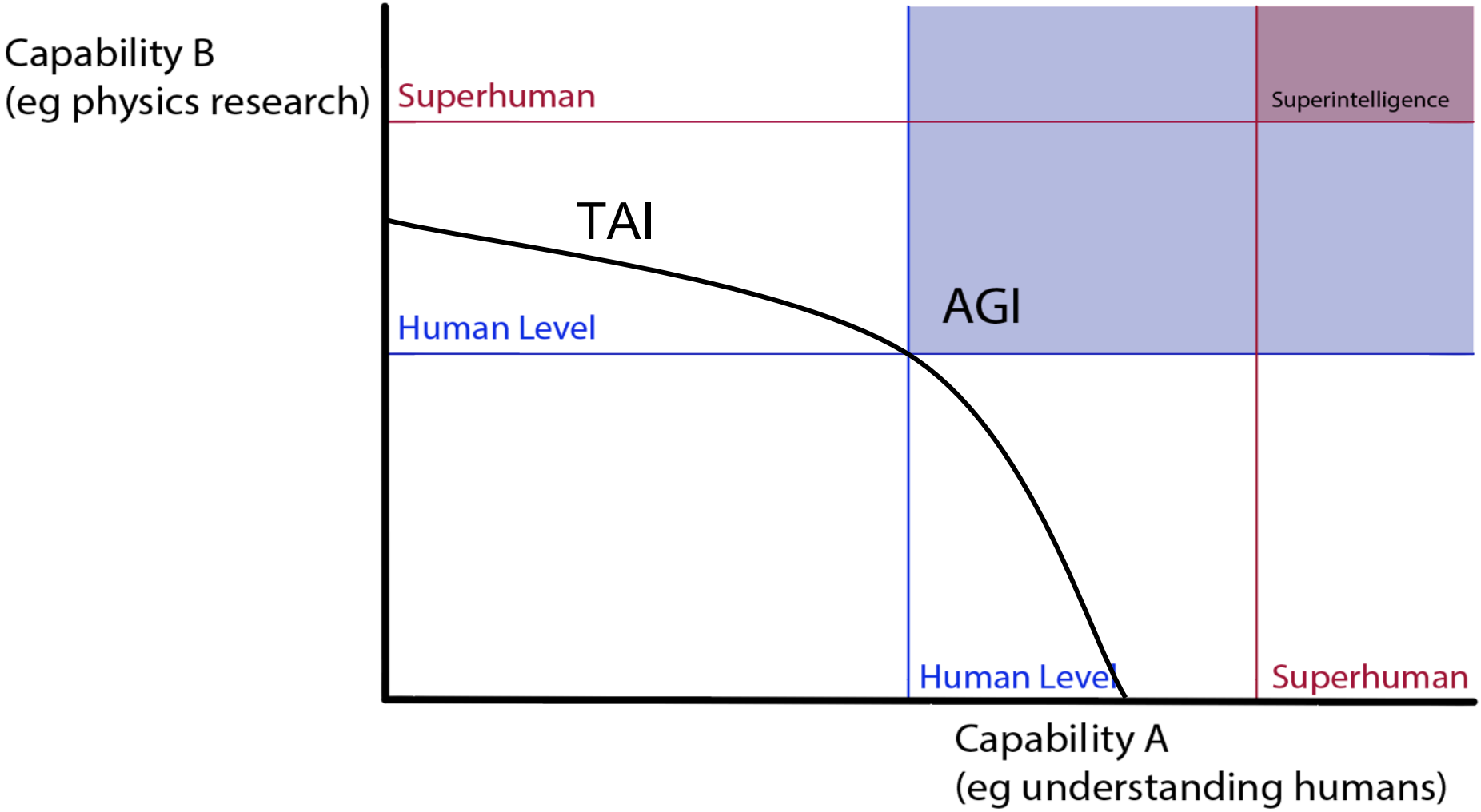
Expert surveys

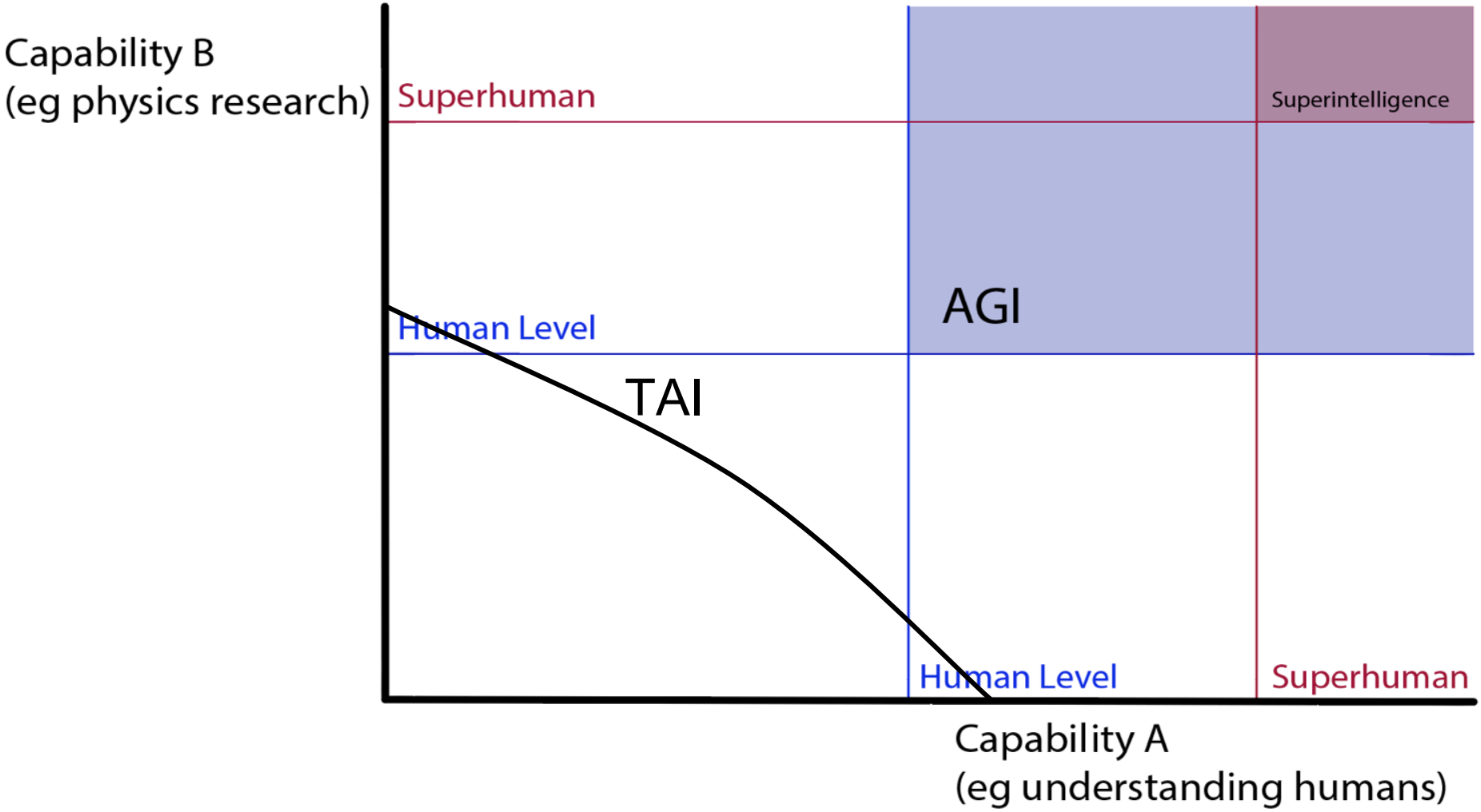
atja Grace, John Salvatier, Baobao Zhang, Allan Dafoe, Owain Eva

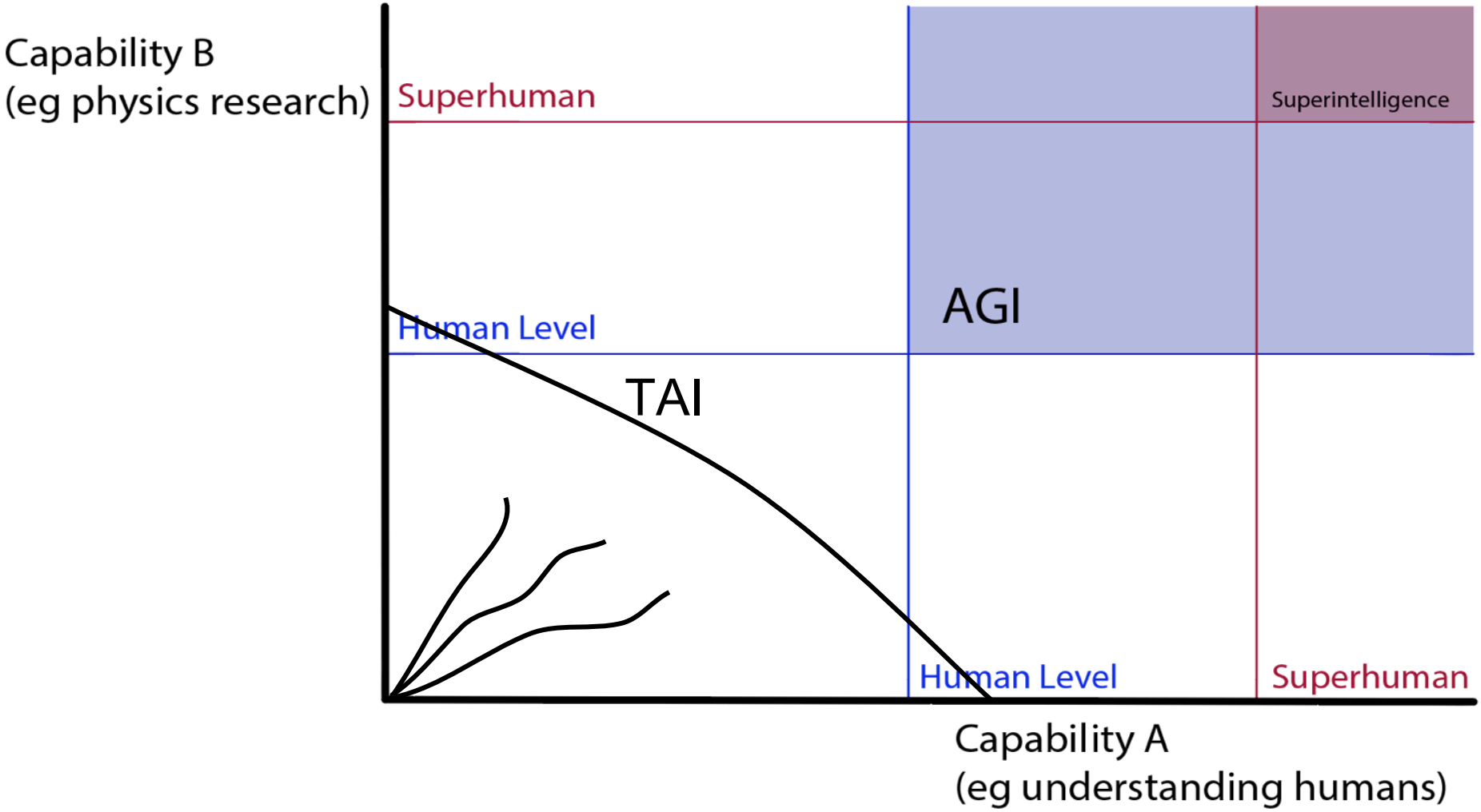


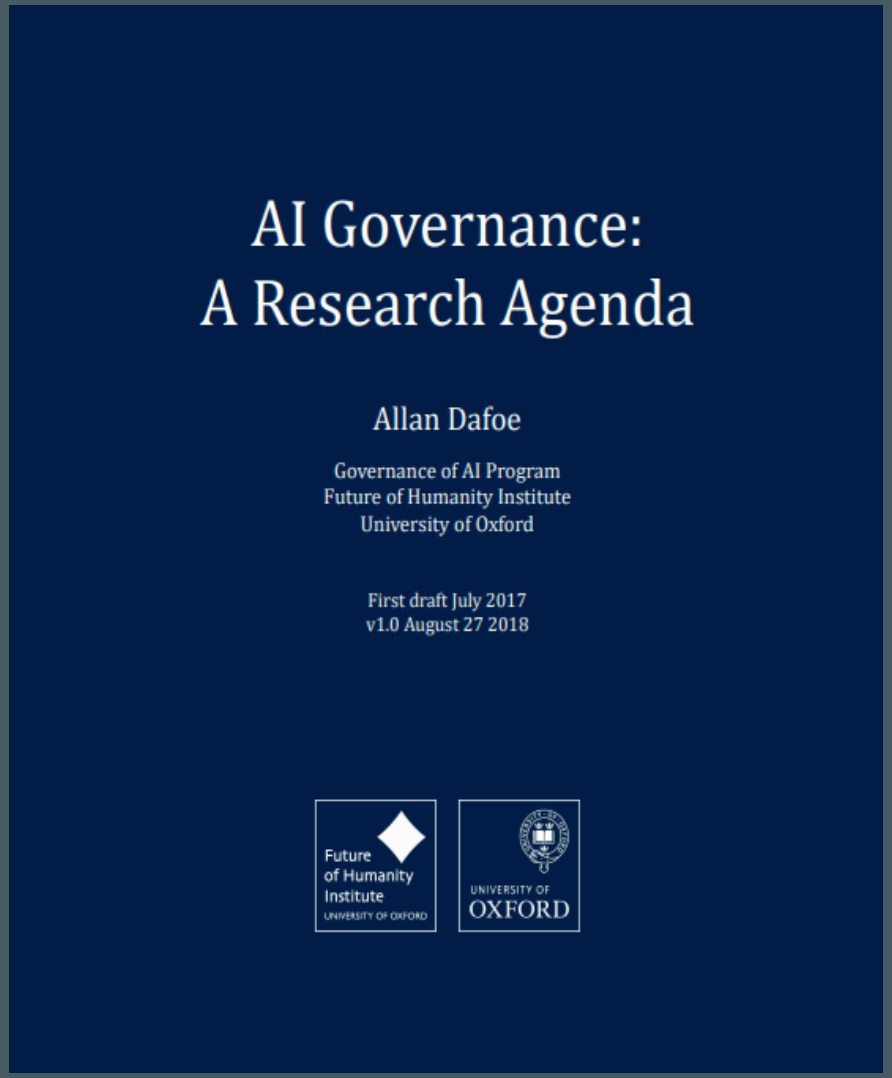



Capability B
(eg physics research)
Superhuman
Superintelligence
AGI
Human Level
Human Level
Superhuman
Capability A
(eg understanding humans)


Capability B
(eg physics research)
Superhuman
Superintelligence
TAI
AGI
Human Level
Human Level
Superhuman
Capability A
(eg understanding humans)


Capability B
(eg physics research)
Superhuman
Superintelligence
AGI
Human Level
TAI
Human Level
Superhuman
Capability A
(eg understanding humans)


Capability B
(eg physics research)
Superhuman
Superintelligence
AGI
Human Level
TAI
Human Level
Superhuman
Capability A
(eg understanding humans)



Technical landscape: capabilities, mapping, forecasting, safety

Politics: international geopolitics, domestic and mass politics, IPE, international security

Ideal Governance: values, principles, appealing positive visions, institutional design, norm building

Policy: translation of long-term goals into concrete near-term policy actions

Political Challenges from (Near-Term) AI

Politics of Algorithms

1. Privacy
2. Fairness
3. Transparency; Interpretability; Auditability
4. Accountability
5. Robustness
6. Safety
7. Security
8. Alignment
9. Innovation

Domestic Politics

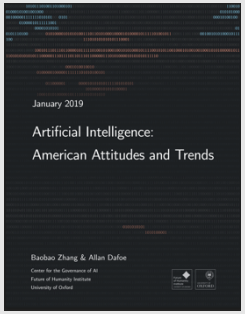
10. Labor displacement and inequality
11. Surveillance and control
12. Influence
13. Fearful backlash; clumsy policy

International Political Economy

10. Natural global oligopolies
11. Tax law
12. Competition policy (antitrust)

International Security

10. LAWs and cyber
11. Power shifts
12. Strategic stability
13. Militarization

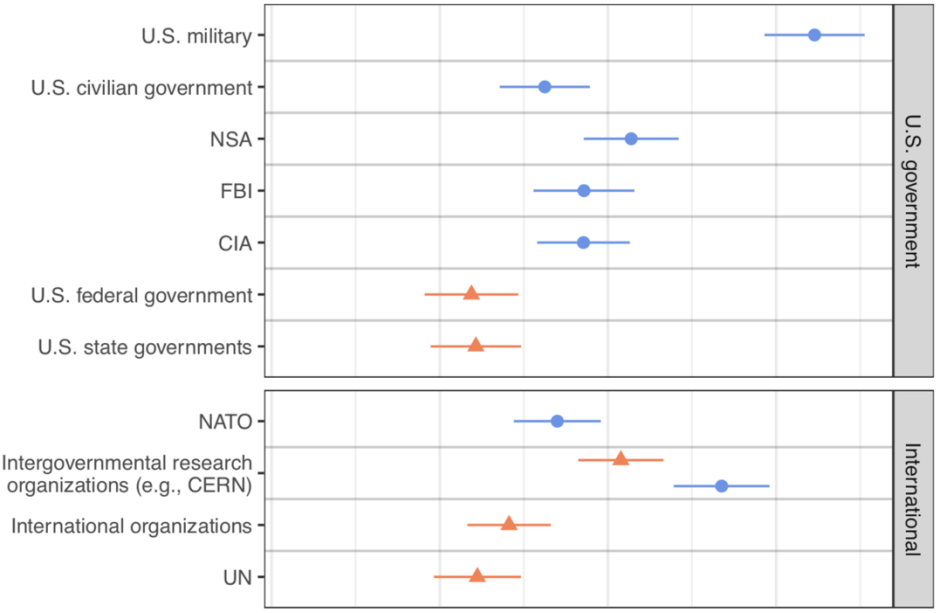


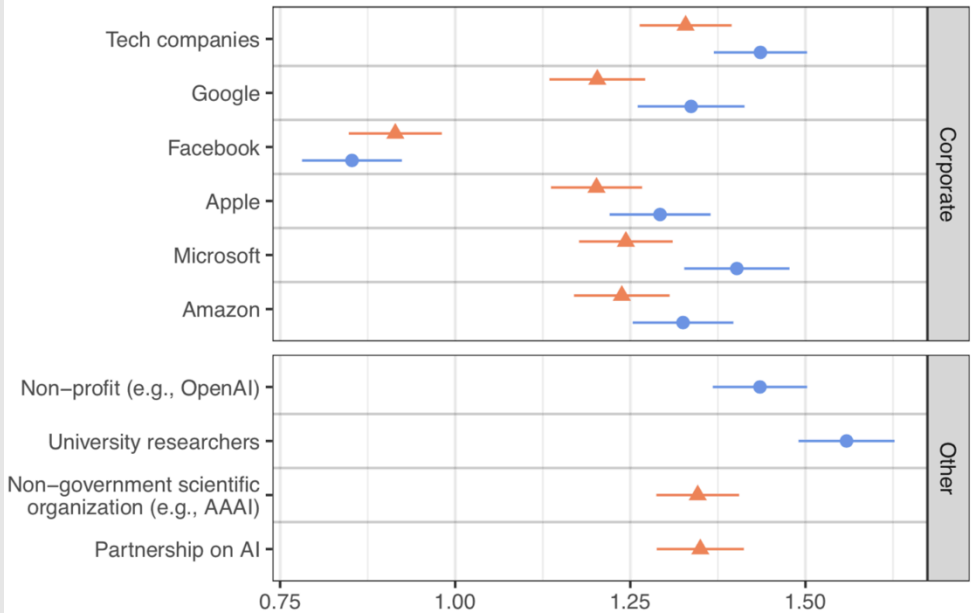
Public opinion surveys

[Baobao Zhang, Allan Dafoe]

Trust in actors to develop/manage AI in the interest of the public.

0=No Confidence. 1=Not too much confidence. 2=A fair amount of confidence. 3=A great deal of confidence



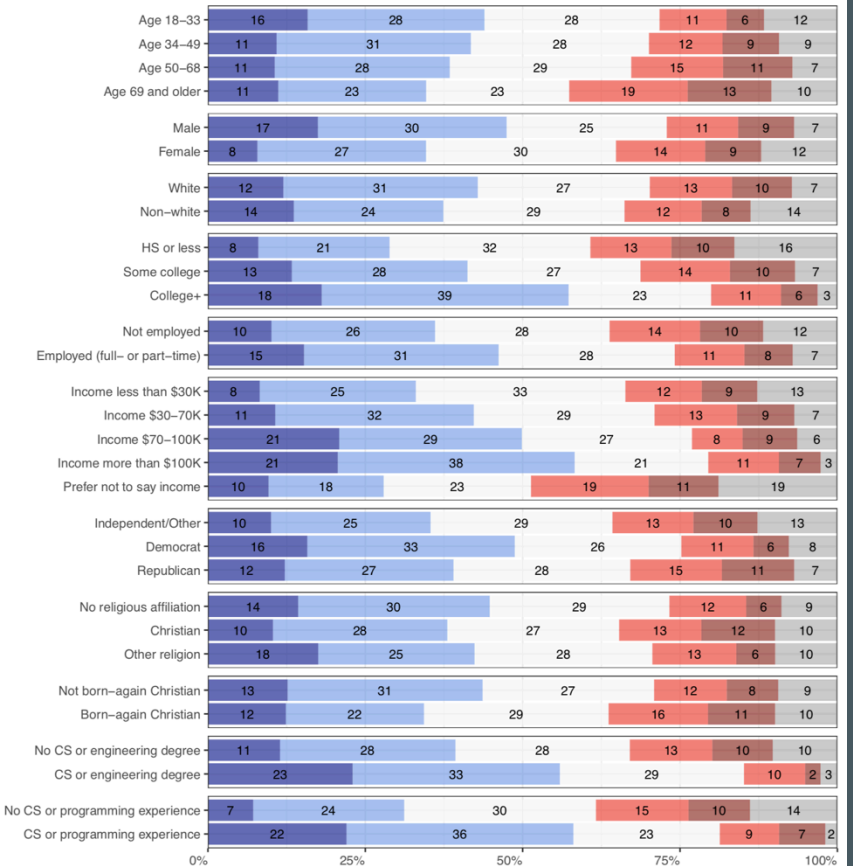


Support for developing AI

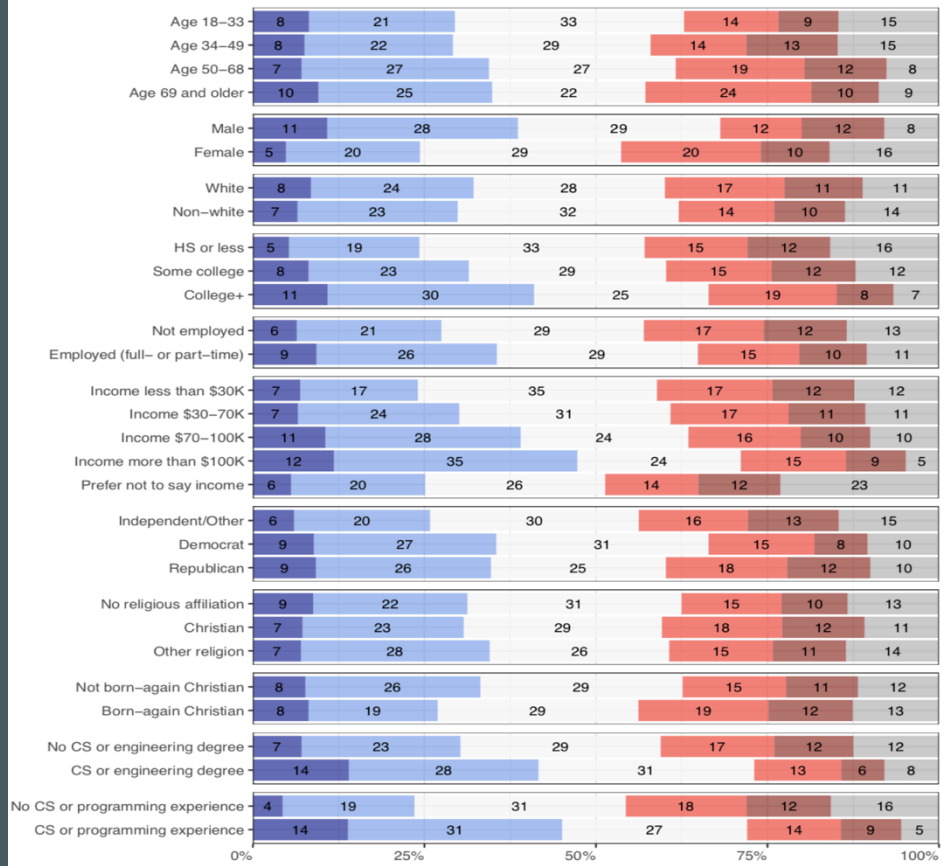
Support for developing HLMI

Profile of concern: female, less educated, poor, without CS experience

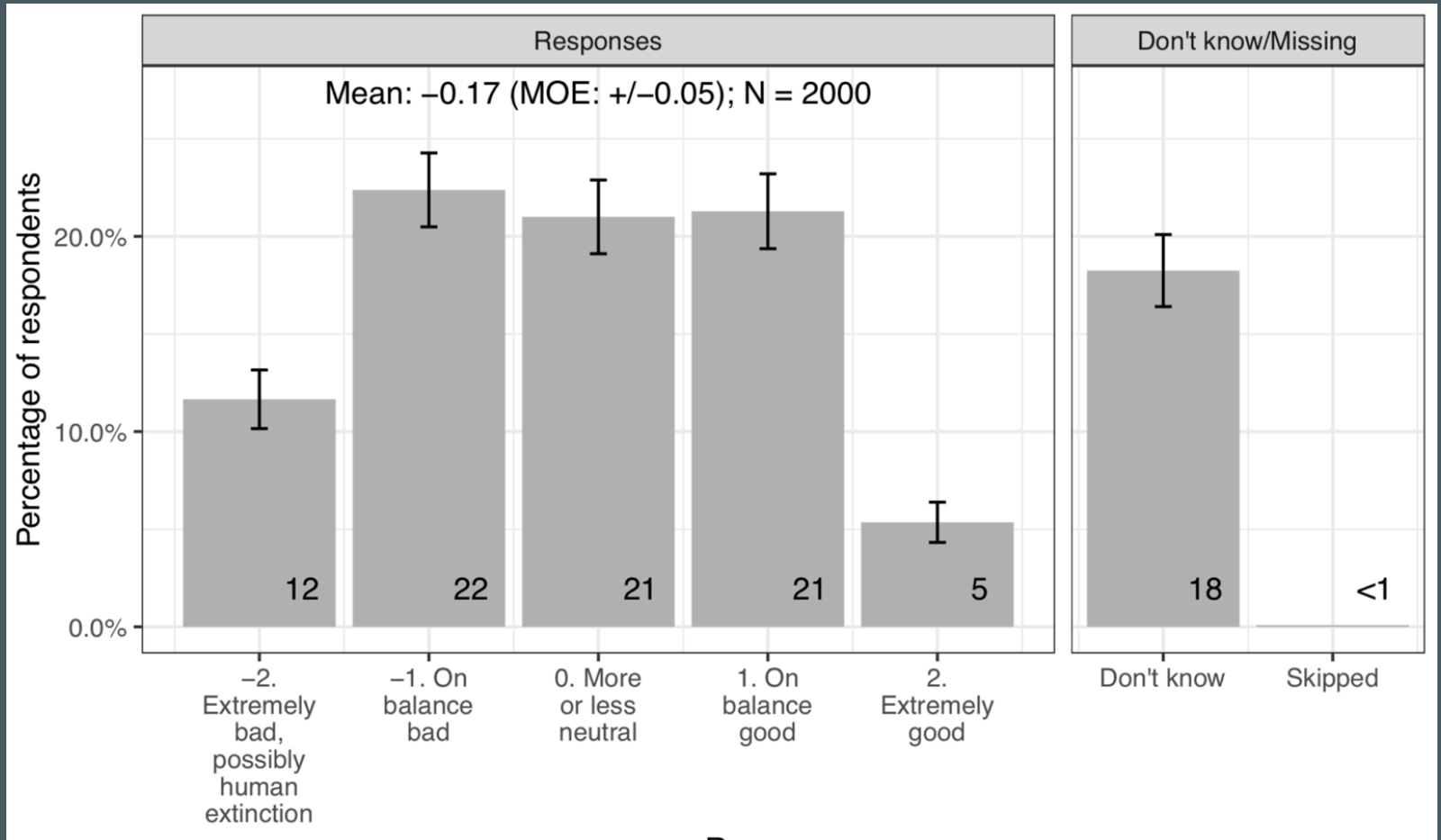
Support for developing AI



Support for developing HLMI



Expected impact of high-level



Political Challenges from (Near-Term) AI

Politics of Algorithms

1. Privacy
2. Fairness
3. Transparency; Interpretability; Auditability
4. Accountability
5. Robustness
6. Safety
7. Security
8. Alignment
9. Innovation

Domestic Politics

10. Labor displacement and inequality
11. Surveillance and control
12. Influence
13. Fearful backlash; clumsy policy

International Political Economy

10. Natural global oligopolies
11. Tax law
12. Competition policy (antitrust)

International Security

10. LAWs and cyber
11. Power shifts
12. Strategic stability
13. Militarization

Malicious Use of AI

[Brundage et al.]

The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation

February 2018

Miles Brundage[1] Shahar Avin[2] Jack Clark[3] Helen Toner[4] Peter Eckersley[5] Ben Garfinkel[6] Allan Dafoe[7] Paul Scharre[8] Thomas Zeitzoff[9] Bobby Filar[10] Hyrum Anderson[11] Heather Roff[12] Gregory C. Allen[13] Jacob Steinhardt[14] Carrick Flynn[15] Seán Ó hÉigeartaigh[16] Simon Beard[17] Haydn Belfield[18] Sebastian Farquhar[19] Clare Lyle[20] Rebecca Crootof[21] Owain Evans[22] Michael Page[23] Joanna Bryson [24] Roman Yampolskiy[25] Dario Amodei[26]

The Vulnerable World Hypothesis

[Nick Bostrom]

The Vulnerable World Hypothesis[1]

(2018) Nick Bostrom
Future of Humanity Institute
University of Oxford

[*Working Paper*, v. 3.15]
[www.nickbostrom.com]

ABSTRACT

Scientific and technological progress might change people's capabilities or incentives in ways that would destabilize civilization. For example, advances in DIY biohacking tools might make it easy for anybody with basic training in biology to kill millions; novel military technologies could trigger arms races in which whoever strikes first has a decisive advantage; or some economically advantageous process may be invented that produces disastrous negative global externalities that are hard to regulate. This paper introduces the concept of a *vulnerable world*: roughly, one in which there is some level of technological development at which civilization almost certainly gets devastated by default, i.e. unless it has exited the "semi-anarchic default condition". Several counterfactual historical and speculative future vulnerabilities are analyzed and arranged into a typology. A general ability to stabilize a vulnerable world would require greatly amplified capacities for preventive policing and global governance. The vulnerable world hypothesis thus offers a new perspective from which to evaluate the risk-benefit balance of developments towards ubiquitous surveillance or a unipolar world order.

Political Challenges from (Near-Term) AI

Politics of Algorithms

1. Privacy
2. Fairness
3. Transparency; Interpretability; Auditability
4. Accountability
5. Robustness
6. Safety
7. Security
8. Alignment
9. Innovation

Domestic Politics

10. Labor displacement and inequality
11. Surveillance and control
12. Influence
13. Fearful backlash; clumsy policy

International Political Economy

10. Natural global oligopolies
11. Tax law
12. Competition policy (antitrust)

International Security

10. LAWs and cyber
11. Power shifts
12. Strategic stability
13. Militarization

Political Challenges from (Near-Term) AI

Politics of Algorithms

1. Privacy
2. Fairness
3. Transparency; Interpretability; Auditability
4. Accountability
5. Robustness
6. Safety
7. Security
8. Alignment
9. Innovation

Domestic Politics

10. Labor displacement and inequality
11. Surveillance and control
12. Influence
13. Fearful backlash; clumsy policy

International Political Economy

10. Natural global oligopolies
11. Tax law
12. Competition policy (antitrust)

International Security

10. LAWs and cyber
11. Power shifts
12. Strategic stability
13. Militarization

Many of these exacerbated by competition, esp great power security competition

Structural Risks from Artificial Intelligenc

[Remco Zwetsloot, Allan Dafoe]

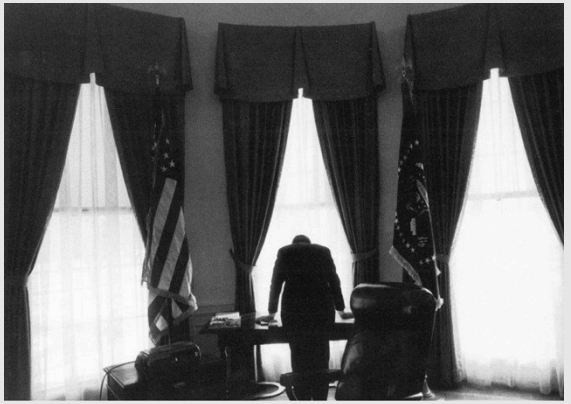
Accidents

Misuse

Structural Risks from Artificial Intelligenc

[Remco Zwetsloot, Allan Dafoe]

Accidents



Misuse

Structural Sources of Risk:

1. Diffuse harms and benefits
2. High uncertainty
3. Fast moving, dynamic problem
4. Irreversible achievements
5. Unclear responsibility
6. Dual-use, broadly available
7. Highly technical
8. Competitive incentives

Levers of Influence

[Sophie-Charlotte Fischer, Jade Leung, Cullen O'Keefe, Allan Dafoe]

Research questions:

What levers of influence does the U.S. government have over AI companies?

What levers of influence do AI companies have over the U.S. government?

How are they likely to be used in various scenarios of AI development?

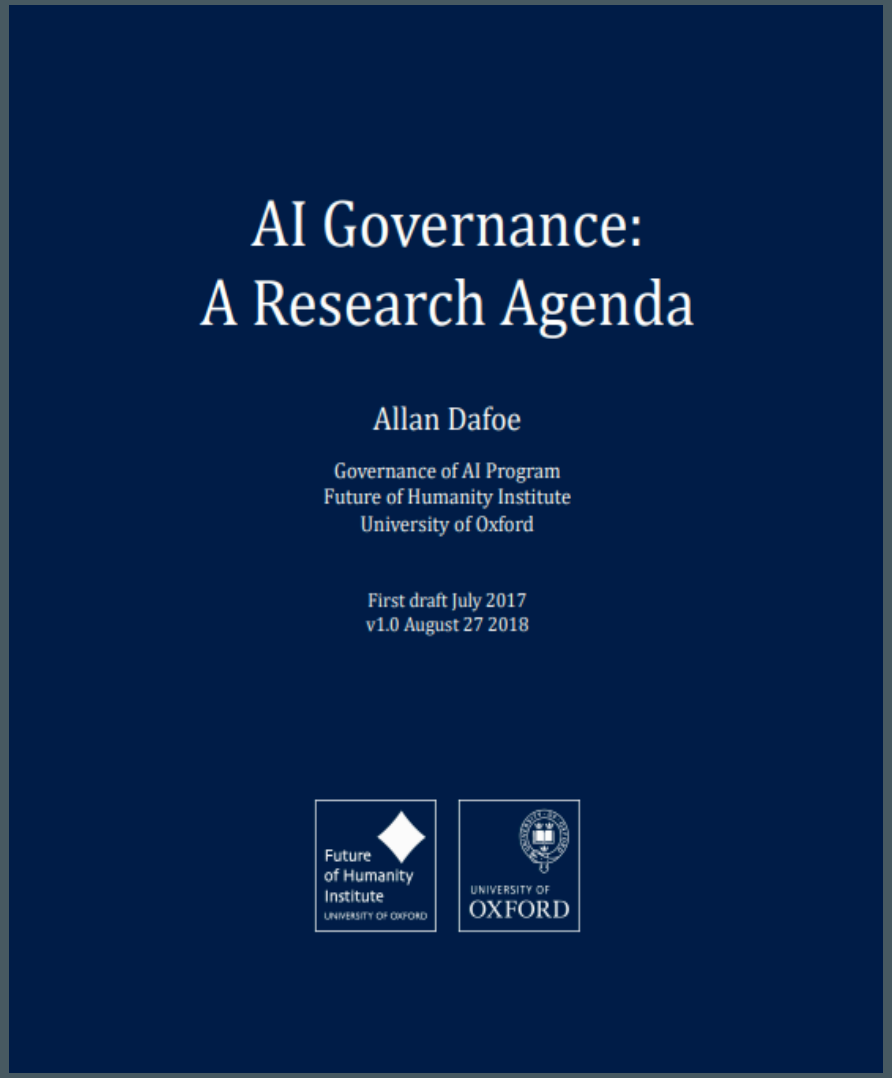
How do these levers compare to those used in other countries?

International control of powerful technolog

[Allan Dafoe, Waqar Zaidi]

Lessons

1. Scientists can be politically powerful.
2. Scientists can play crucial role enabling cooperation.
3. Radical proposals are possible.
4. Confusion. Range of perspectives. Misunderstanding.
5. Messy politics. Muddling through.
6. Ugly decisions made under “necessity”.
7. Realism. Cynicism.
8. Public sphere is crucial.
9. Terrible epistemics, especially given secrecy.
10. Secrecy and fear yields domestic power.
11. Cooperation hinges on trust.



Technical landscape: capabilities, mapping, forecasting, safety

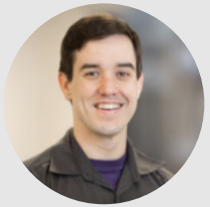
Politics: international geopolitics, domestic and mass politics, IPE, international security

Ideal Governance: values, principles, appealing positive visions, institutional design, norm building

Policy: translation of long-term goals into concrete near-term policy actions

Policy Desiderata

[Nick Bostrom, Allan Dafoe, Carrick Flynn]



Policy Desiderata for Superintelligent AI: A Vector Field Approach[1]

(2018) version 4.3 (first version: 2016)

Nick Bostrom[†], Allan Dafoe[†], Carrick Flynn[†]

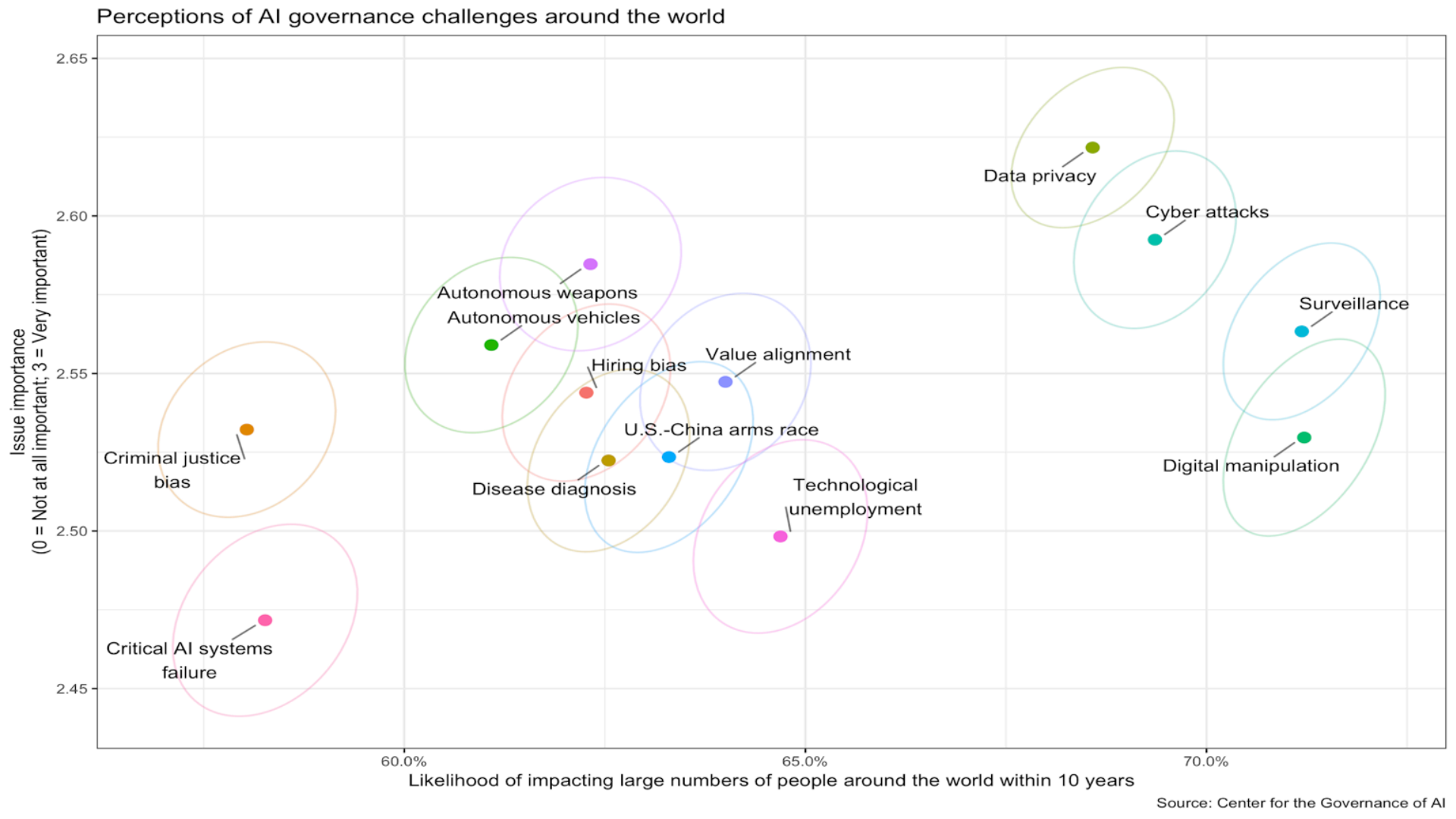
[forthcoming in Liao, S.M. (ed.): *Ethics of Artificial Intelligence* (Oxford University Press, 2019)]

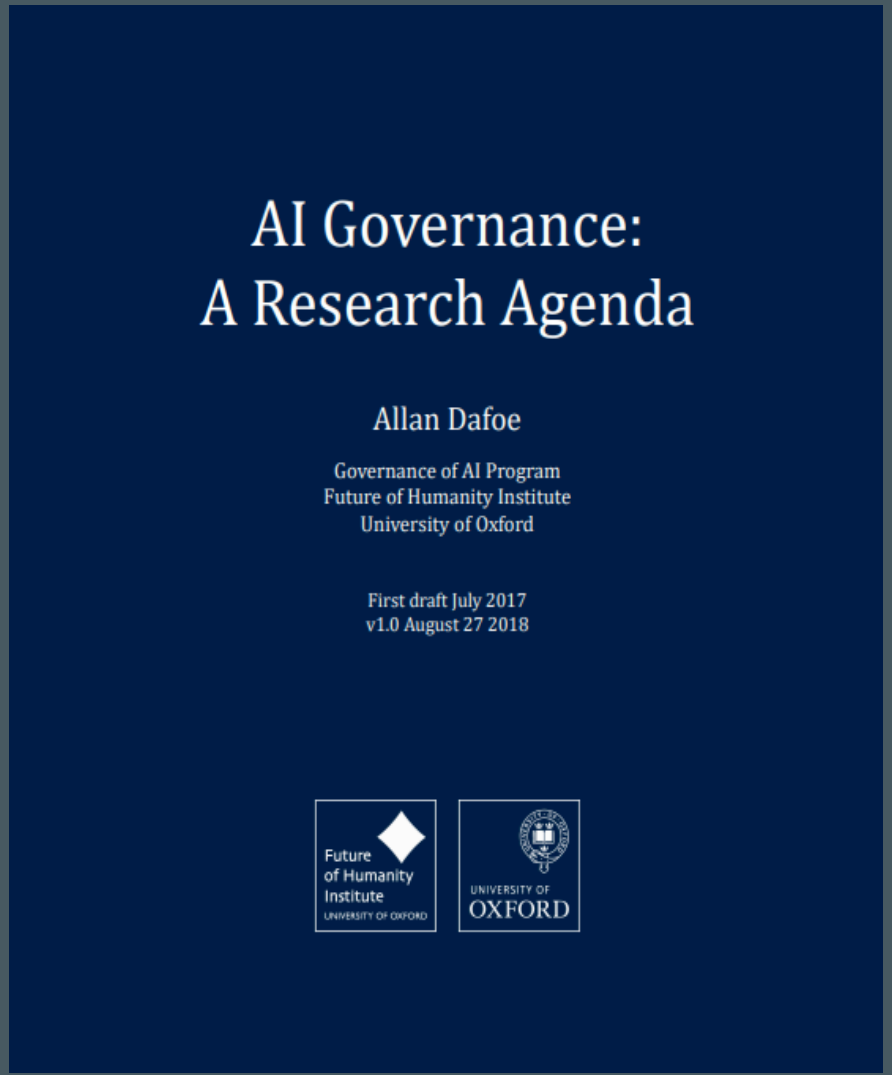
[www.nickbostrom.com/papers/aipolicy.pdf]

ABSTRACT

We consider the speculative prospect of superintelligent AI and its normative implications for governance and global policy. Machine superintelligence would be a transformative development that would present a host of political challenges and opportunities. This paper identifies a set of distinctive features of this hypothetical policy context, from which we derive a correlative set of policy desiderata—considerations that should be given extra weight in long-term AI policy compared to in other policy contexts. Our contribution describes a desiderata "vector field" showing the *directional change* from a variety of possible normative baselines or policy positions. The focus on directional normative change should make our findings relevant to a wide range of actors, although the development of concrete policy options that meet these abstractly formulated desiderata will require further work.

Efficiency	
Technological opportunity AI risk Possibility of catastrophic global coordination failures	**Expeditious progress.** This can be divided into two components: (a) The path chosen leads with high probability to the development of superintelligence and its use to achieve technological maturity and to unlock the cosmic endowment. (b) The progress in AI is speedy, and socially beneficial products and applications are made widely available in a timely fashion. **AI safety.** Techniques are developed that make it possible (without excessive cost, delay, or performance penalty) to ensure that advanced AIs behave as intended. A good alignment solution would enable control of both external and internal behaviour (thus making it possible to avoid intrinsically undesirable types of computation without sacrificing much in terms of performance. **Conditional stabilization.** The path is such that if avoiding catastrophic global coordination failure requires that temporary or permanent stabilization is undertaken or that a singleton is established, then the needed measures are available and are implemented in time to avert the catastrophe. **Non-turbulence.** The path avoids excessive efficiency losses from chaos and conflict. Political systems maintain stability and order, adapt successfully to change, and mitigate any socially disruptive impacts.
Allocation	
Risk externalities Reshuffling Cornucopia Veil of ignorance	**Universal benefit.** All humans who are alive at the transition get some share of the benefit, in compensation for the risk externality to which they were exposed. **Magnanimity.** A wide range of resource-satiable values (ones to which there is little objection aside from cost-based considerations), are realized if and when it becomes possible to do so using a minute fraction of total resources. This may encompass basic welfare provisions and income guarantees to all human individuals. It may also encompass many community goods, ethical ideals, aesthetic or sentimental projects, and various natural expressions of generosity, kindness, and compassion. **Continuity.** The path affords a reasonable degree of continuity such as to (i) maintain order and provide the institutional stability needed for actors to benefit from opportunities for trade behind the current veil of ignorance, including social safety nets; and (ii) prevent concentration and permutation from racially exceeding the levels implicit in the current social contract.
Population	
Interests of digital minds Population dynamics	**Mind crime prevention.** AI is governed in such a way that maltreatment of sentient digital minds is avoided or minimized. **Population control.** Generational choices, concerning what new beings to bring into existence, are made in a coordinated manner and with sufficient foresight to avoid unwanted Malthusian dynamics and political erosion.
Mode	
Context transformation	**Responsibility and wisdom.** The seminal applications of advanced AI are shaped by an agency (individual or distributed) that has an expansive sense of responsibility and the practical wisdom to see what needs to be done in radically unfamiliar circumstances.


Perceptions of AI governance challenges around the world
Issue importance
(0 = Not at all important; 3 = Very important)
2.65
2.60
2.55
2.50
2.45
Data privacy
Cyber attacks
Surveillance
Digital manipulation
Autonomous weapons
Autonomous vehicles
Hiring bias
Value alignment
U.S.-China arms race
Disease diagnosis
Technological unemployment
Criminal justice bias
Critical AI systems failure
60.0%
65.0%
70.0%
Likelihood of impacting large numbers of people around the world within 10 years
Source: Center for the Governance of AI



Technical landscape: capabilities, mapping, forecasting, safety

Politics: international geopolitics, domestic and mass politics, IPE, international security

Ideal Governance: values, principles, appealing positive visions, institutional design, norm building

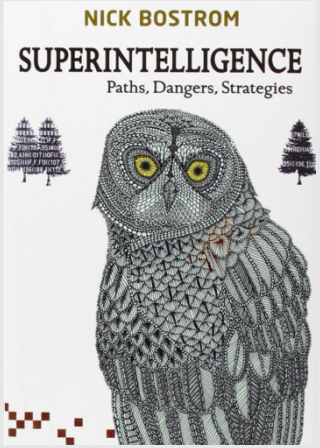
Policy: translation of long-term goals into concrete near-term policy actions

Windfall Clause

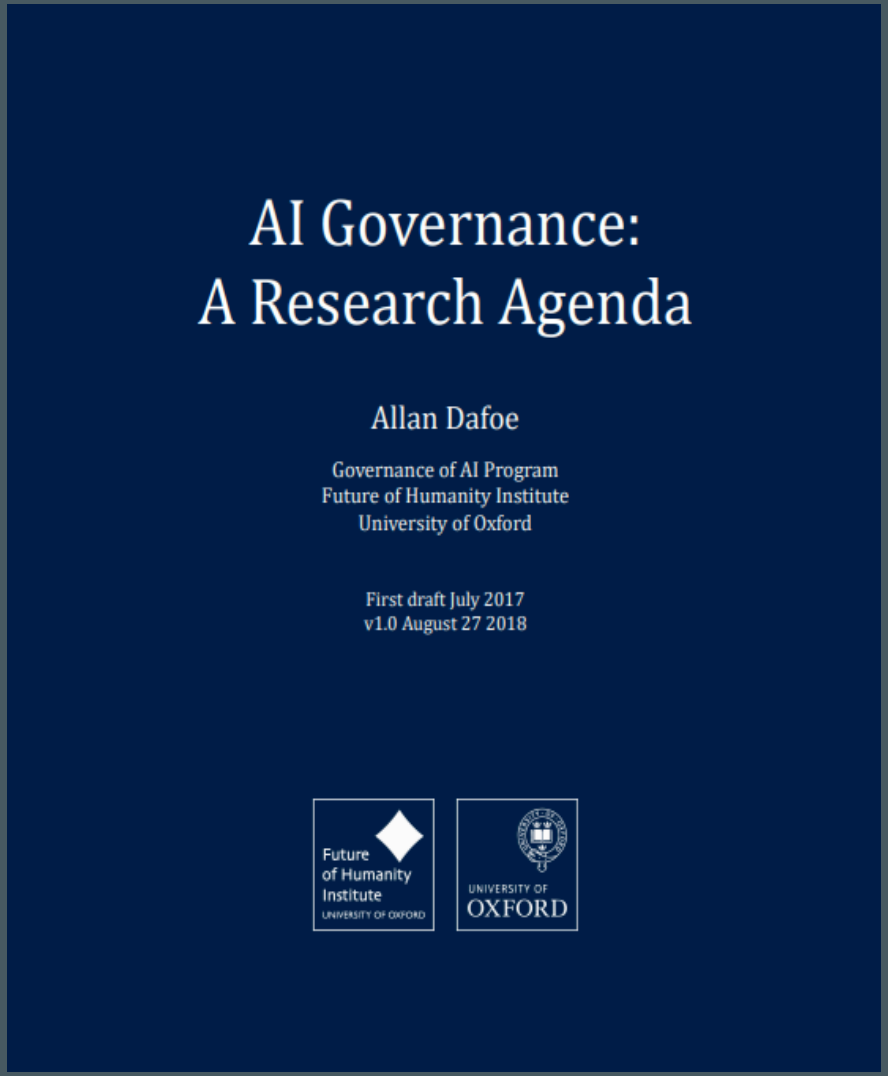
[Cullen O'Keefe, Carrick Flynn, Ben Garfinkel, Peter Cihon, Allan D

The common good principle: Superintelligence should be developed only for the benefit of all of humanity and in the service of widely shared ethical ideals.

A "windfall clause" to the effect that … profits in excess of [a very high threshold, say a trillion dollars annually] would be distributed to all of humanity… Adopting [it] should be substantially costless … its widespread adoption would give humankind a valuable guarantee … [that] everybody would share in most of the benefits.



I. **Motivation**
II. **Legal Permissibility**
III. **Recipients**



Technical landscape: capabilities, mapping, forecasting, safety

Politics: international geopolitics, domestic and mass politics, IPE, international security

Ideal Governance: values, principles, appealing positive visions, institutional design, norm building

Policy: translation of long-term goals into concrete near-term policy actions

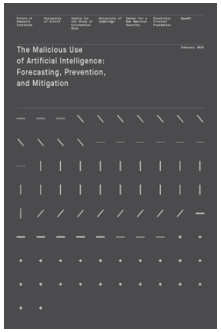
Future of Humanity Institute
UNIVERSITY OF OXFORD

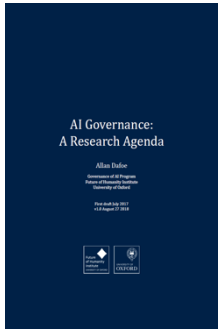
Center for the Governance of AI

UNIVERSITY OF OXFORD

Team

Research





Public

FOREIGN AFFAIRS

nnt NTI PS

BBC Machine Learning Fireside Chats

MIT Technology Review

Yale JACKSON INSTITUTE FOR GLOBAL AFFAIRS

The Economist

Effective Altruism Global

aippg appg All Party Parliamentary Group on Artificial Intelligence

CYBERUK 2018

PUBLIC POLICY FORUM
FORUM DES POLITIQUES PUBLIQUES
PPF

CISAC Stanford

Learn more at: governance.ai