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
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 processes. Good governance usually means that it is effective, legitimate, inclusive, and 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

**Deciphering China’s AI Dream**

The context, components, capabilities, and consequences of China’s strategy to lead the world in AI

Jeffrey Ding
Governance of AI Program,
Future of Humanity Institute, University of Oxford
March 2018
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 assume 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 transformiveness of previous technologies. However, this remains a rapidly-developing area well worth following.
Strategic properties of artificial intelligence

[Ben Garfinkel, Allan Dafoe]

How Does the Offense-Defense Balance Scale? *

Ben Garfinkel, Allan Dafoe†

May 15, 2018

[Diagram showing a heatmap with axes labeled 'Zero day acquisition' and 'Defensive investments']
Expert surveys
Katja Grace, John Salvatier, Baobao Zhang, Allan Dafoe, Owain Evans
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
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 machine intelligence

**Mean:** $-0.17$ (MOE: $+/0.05$); **N:** 2000

<table>
<thead>
<tr>
<th>Responses</th>
<th>Percentage of respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2. Extremely bad, possibly human extinction</td>
<td>12%</td>
</tr>
<tr>
<td>-1. On balance bad</td>
<td>22%</td>
</tr>
<tr>
<td>0. More or less neutral</td>
<td>21%</td>
</tr>
<tr>
<td>1. On balance good</td>
<td>21%</td>
</tr>
<tr>
<td>2. Extremely good</td>
<td>5%</td>
</tr>
</tbody>
</table>

| Don't know/Missing | 18% |
| Skipped | <1% |
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

The Vulnerable World Hypothesis
[Nick Bostrom]

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

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 Intelligence

[Remco Zwetsloot, Allan Dafoe]

Accidents

Misuse
Structural Risks from Artificial Intelligence

[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
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 technology

[Allan Dafoe, Waqar Zaidi]

Lessons
1. Scientists can be politically powerful.
2. Scientists can play crucial role enabling cooperation.
3. Radical proposals are possible.
6. Ugly decisions made under “necessity”.
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 for Superintelligent AI: A Vector Field Approach¹

(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/ai/policy.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.
Perceptions of AI governance challenges around the world

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 Dafoe]

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
Narrow Transformative Capabilities

Most likely where:

- Data rich
- Dan simulate environment
- Narrow domains
- Ripe technical problem
- Fast decisions
- Many variables
- High stakes

Candidate Application Areas

- Engineering, science, math, drug
discovery, material science.
- Cyber.
- Surveillance.
- Profiling (lie detection, emotion
detection, psychological insight,
DNA). Personal assistants/
advertising.
- Social network mapping and
manipulation.
1. Achieve common knowledge that those who develop AI or profit from AI bear social responsibility for AI.
2. Commit to support prosocial actors and initiatives.
3. Inclusive and expert deliberation to refine our understanding of what AI social responsibility entails.
4. Patience and humility regarding object level norms.
5. Critical third-party assessment of compliance with that responsibility.

It is our collective responsibility to expect AI companies to act for the common good, and to promote those that lead the way.
Strategic Implications of Openness/Closure

Nick Bostrom

Global Policy

Strategic Implications of Openness in AI Development

Abstract
This paper attempts a preliminary analysis of the global desirability of different forms of openness in AI development (including openness about source code, science, data, safety techniques, capabilities, and goals). Short-term impacts of increased openness appear mostly socially beneficial in expectation. The strategic implications of medium and long-term impacts are complex. The evaluation of long-term impacts, in particular, may depend on whether the objective is to benefit the present generation or to promote a time-neutral aggregate of well-being of future generations. Some forms of openness are plausibly positive on both counts (openness about safety measures, openness about goals). Others (openness about source code, science, and possibly capability) could lead to a tightening of the competitive situation around the time of the introduction of advanced AI, increasing the probability that winning the AI race is incompatible with using any safety method that incurs a delay or limits performance. We identify several key factors that must be taken into account by any well-founded opinion on the matter.

Policy Implications
- The global desirability of openness in AI development – sharing e.g. source code, algorithms, or scientific insights – depends on complex tradeoffs.
- A central concern is that openness could exacerbate a racing dynamic; competitors trying to be the first to develop advanced (superintelligent) AI may accept higher levels of existential risk in order to accelerate progress.
- Openness may reduce the probability of AI benefits being monopolized by a small group, but other potential political consequences are more problematic.
- Partial openness that enables outsiders to contribute to an AI project’s safety work and to supervise organizational plans and goals appears desirable.

Private Closing of AI R&D Today: Proposal and Considerations

Summary
Motivation and Proposals
Scope of the Bracket
Recommendations
Appendix: Considerations
- Source of Risk: Close-Race, Non-Liberal AI Programs, or Bad Actors
- Detectability
- Closure May Attract One’s Government
- Government Is Better Positioned to Classify and Control Dangerous Information
- Justifications: Importance, Persuasiveness, Sincerity, Leaks.
- Benefits of Experimenting with Closed Policies
- Risk of Researcher Backlash
- Charges of Anti-Competitive Behavior and Protectionism
- Distance to TAI
- Innovation Concentration and Optimal Bracket Size
- Fungibility of Resources
- Risk of Racing
- Role of State in AI Race
- Optimal Racing Strongly Benefits from Coherent Grand Strategy
Superintelligence Problems

[The governance problem when imagining AGI] One way into this problem is to imagine a world with AGI or superintelligence, which immediately makes several important issues clear.

[Safety/Alignment problem.] How safe is the system? How aligned? How sure are we that it is aligned? What is the epistemic organizational process for convincing ourselves that it is sufficiently aligned?

Technical solutions and Organizational Solutions:
Much of safety is in the organization, the incentives. What are the best org structures, let’s make sure: Whistleblowing, safety through debate, monitoring and surveillance, correct incentives.

Value Selection Problem (moral philosophy, politics). But then, aligned to what values? Whose values? If we’re going to aggregate, how do we do that?

Purpose Problem:

Political Problem. How do we get there? Or if there are multiple power centers, how do we make sure their interactions are not harmful?
AI Governance: A Research Agenda

Allan Dafoe
Governance of AI Program
Future of Humanity Institute
University of Oxford

First draft July 2017
v1.0 August 27 2018
Perceived trust in various actors to develop and manage AI

Outcome measures
- Trust in various actors to develop AI in the interest of the public
- Trust in various actors to manage AI in the interest of the public

Source: Center for the Governance of AI
Capability B (eg physics research)

Superhuman

Human Level

AGI

Human Level

Superintelligence

Capability A (eg understanding humans)
Capability B (eg physics research)

Superhuman

Human Level

Capability A (eg understanding humans)

AGI

Human Level

Superintelligence