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, 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

Superintelligence: Paths, Dangers, Strategies

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 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 transformative nature of previous technologies. However, this remains a rapidly-developing area well worth following.¹

A Relevance of progress in artificial intelligence

A.1 AI systems may enable and motivate more effective surveillance . 90
A.2 AI systems may help to make privacy-preserving surveillance feasible . 90
A.3 AI systems may increase the need for anti-forgery schemes . 91
A.4 Methods of computing on confidential data could help to decentralize the training of AI systems . 91
A.5 AI systems could be designed to interact with decentralized applications . 92
A.6 The problems of safe AI design and safe smart contract design may be connected . 93
A.7 New coordination and verification mechanisms may be useful for governing AI systems . 94
A.8 Changes to the political landscape, generally, may impact the governance of AI systems . 94
A.9 Fully homomorphic encryption may have applications in AI safety . 94
Strategic properties of artificial intelligence

[Ben Garfinkel, Allan Dafoe]
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
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

<table>
<thead>
<tr>
<th>Age 18-33</th>
<th>Age 34-49</th>
<th>Age 50-64</th>
<th>65 and older</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>18</td>
<td>28</td>
<td>11</td>
</tr>
<tr>
<td>Female</td>
<td>8</td>
<td>30</td>
<td>23</td>
</tr>
<tr>
<td>White</td>
<td>12</td>
<td>31</td>
<td>27</td>
</tr>
<tr>
<td>Non-white</td>
<td>14</td>
<td>24</td>
<td>29</td>
</tr>
<tr>
<td>HS or less</td>
<td>8</td>
<td>21</td>
<td>32</td>
</tr>
<tr>
<td>Some college</td>
<td>12</td>
<td>28</td>
<td>32</td>
</tr>
<tr>
<td>College+</td>
<td>18</td>
<td>39</td>
<td>23</td>
</tr>
<tr>
<td>Not employed</td>
<td>16</td>
<td>26</td>
<td>28</td>
</tr>
<tr>
<td>Employed (full- or part-time)</td>
<td>15</td>
<td>31</td>
<td>28</td>
</tr>
<tr>
<td>Income less than $30K</td>
<td>8</td>
<td>25</td>
<td>33</td>
</tr>
<tr>
<td>Income $30-70K</td>
<td>11</td>
<td>32</td>
<td>29</td>
</tr>
<tr>
<td>Income $70-100K</td>
<td>21</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>Income more than $100K</td>
<td>21</td>
<td>38</td>
<td>21</td>
</tr>
<tr>
<td>Prefer not to say income</td>
<td>10</td>
<td>18</td>
<td>23</td>
</tr>
</tbody>
</table>

### Support for developing HLMI

<table>
<thead>
<tr>
<th>Age 18-33</th>
<th>Age 34-49</th>
<th>Age 50-64</th>
<th>65 and older</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>11</td>
<td>28</td>
<td>19</td>
</tr>
<tr>
<td>Female</td>
<td>12</td>
<td>30</td>
<td>25</td>
</tr>
<tr>
<td>White</td>
<td>14</td>
<td>24</td>
<td>27</td>
</tr>
<tr>
<td>Non-white</td>
<td>12</td>
<td>22</td>
<td>29</td>
</tr>
<tr>
<td>HS or less</td>
<td>8</td>
<td>21</td>
<td>32</td>
</tr>
<tr>
<td>Some college</td>
<td>13</td>
<td>28</td>
<td>32</td>
</tr>
<tr>
<td>College+</td>
<td>12</td>
<td>31</td>
<td>23</td>
</tr>
<tr>
<td>Not employed</td>
<td>10</td>
<td>26</td>
<td>28</td>
</tr>
<tr>
<td>Employed (full- or part-time)</td>
<td>15</td>
<td>31</td>
<td>28</td>
</tr>
<tr>
<td>Income less than $30K</td>
<td>8</td>
<td>25</td>
<td>33</td>
</tr>
<tr>
<td>Income $30-70K</td>
<td>11</td>
<td>32</td>
<td>29</td>
</tr>
<tr>
<td>Income $70-100K</td>
<td>21</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>Income more than $100K</td>
<td>21</td>
<td>38</td>
<td>21</td>
</tr>
<tr>
<td>Prefer not to say income</td>
<td>10</td>
<td>18</td>
<td>23</td>
</tr>
</tbody>
</table>
Expected impact of high-level machine intelligence

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

<table>
<thead>
<tr>
<th>Response</th>
<th>Percentage of respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.</td>
<td>12</td>
</tr>
<tr>
<td>-1.</td>
<td>22</td>
</tr>
<tr>
<td>0.</td>
<td>21</td>
</tr>
<tr>
<td>1.</td>
<td>21</td>
</tr>
<tr>
<td>2.</td>
<td>5</td>
</tr>
<tr>
<td>Don't know</td>
<td>18</td>
</tr>
<tr>
<td>Skipped</td>
<td>&lt;1</td>
</tr>
</tbody>
</table>
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]

The Vulnerable World Hypothesis

(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

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

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
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 Approach

(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.
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
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.
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