Bridging the gap:

the case for an 'Incompletely Theorized Agreement' on AI policy

June 7th, 2021 | ***Quo Vadis, AI Ethics?*** |
Gothenburg – Chalmers AI Research Centre

Charlotte Stix & **Matthijs Maas**

UNIVERSITY OF CAMBRIDGE

CENTRE FOR THE STUDY OF EXISTENTIAL RISK

An argument made with Charlotte Stix (TU Eindhoven)

Springer Link

Original Research | Open Access | Published: 15 January 2021

Bridging the gap: the case for an 'Incompletely Theorized Agreement' on AI policy

Charlotte Stix & Matthijs M. Maas

AI and Ethics (2021) | Cite this article

1873 Accesses | **35** Altmetric | Metrics

Abstract

Recent progress in artificial intelligence (AI) raises a wide array of ethical and societal concerns. Accordingly, an appropriate policy approach is urgently needed. While there has been a wave of scholarship in this field, the research community at times appears divided amongst those who emphasize 'near-term' concerns and those focusing on 'long-term' concerns and corresponding policy measures. In this paper, we seek to examine this alleged 'gap', with a view to understanding the practical space for inter-community collaboration on AI policy. We propose to make use of the principle of an 'incompletely theorized agreement' to bridge some underlying disagreements, in the name of important cooperation on addressing AI's urgent challenges. We propose that on certain issue areas, scholars working with near-term and long-term perspectives can converge and cooperate on selected mutually

In short:

Why we should

‘Bridge the gap’

in AI policy

(and one way how)

- **The Gap**
- **The Bridge**

Bridging… What gap?

Why all these AI ethics debates? AI gives rise to a range of existing, emerging, and anticipated challenges

We need a well-coordinated AI expert epistemic community to take the lead on these challenges

- …these challenges will need **informed policy responses** at both national and international level
- This requires a well-coordinated AI expert 'epistemic community' able to:
  - shift researcher community norms (where necessary),
  - scrutinize AI principals (tech developers and users),
  - engage public and affected stakeholders in debate, and in policy formation around AI
  - advocate governments on specific policies, minimum standards, etc.
  - Etc. etc. …

A ‘house divided’ in the ‘responsible AI policy’ community?

- …in recent years, there has been an *apparent* separation in the responsible AI community—between those focusing on AI’s evident problems in the ‘near-term’, or on more uncertain challenges in the ‘long-term’
  - NT: (algorithmic bias, facial recognition, self-driving cars, DeepFakes, lethal autonomous weapon systems, digital humanitarianism, environmental impact…)
  - LT: (far-reaching unemployment; ‘transformative’ societal impacts including risks to ‘epistemic security’; risks to military strategic stability; extreme risks from very advanced AI systems (whether AGI or other), if not aligned with human values, …)

Recent work has begun to question and nuance the meaningfulness of this distinction…

RESEARCH-ARTICLE

Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society

Authors: Carina Prunkl, Jess Whittlestone Authors Info & Affiliations

Publication: AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society • February 2020 • Pages 138–143 • https://doi.org/10.1145/3375627.3375803

2 136 Get Access

ABSTRACT

One way of carving up the broad 'AI ethics and society' research space that has emerged in recent years is to distinguish between 'near-term' and 'long-term' research. While such ways of breaking down the research space can be useful, we put forward several concerns about the

Open Access Article

Medium-Term Artificial Intelligence and Society

by Seth D. Baum

Global Catastrophic Risk Institute, P.O. Box 40364, Washington, DC 20016, USA

Information **2020**, *11*(6), 290; https://doi.org/10.3390/info11060290

Received: 16 February 2020 / Revised: 25 May 2020 / Accepted: 26 May 2020 / Published: 29 May 2020

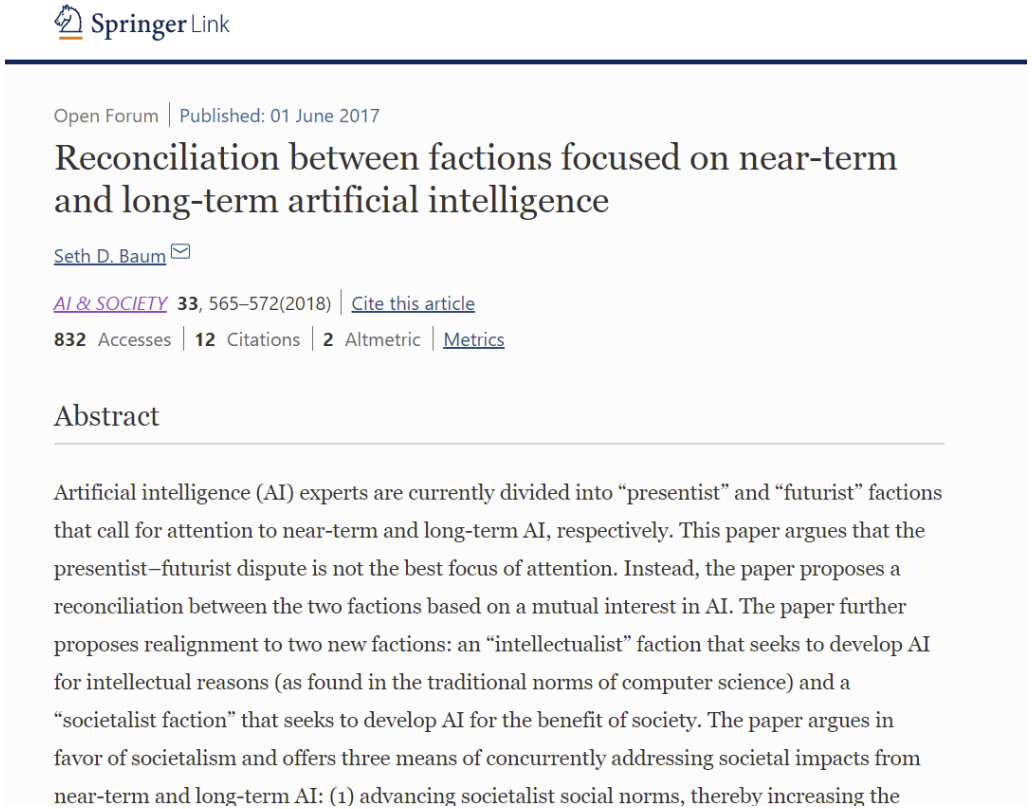
(This article belongs to the Section Artificial Intelligence)

View Full-Text Download PDF Browse Figures Citation Export

Abstract

There has been extensive attention to near-term and long-term AI technology and its accompanying societal issues, but the medium-term has gone largely overlooked. This paper develops the concept of medium-term AI, evaluates its importance, and analyzes some medium-term societal issues. Medium-term AI can be important in its own right and as a topic that can bridge the sometimes acrimonious divide between those who favor attention to near-term AI and those who prefer the long-term. The paper proposes the medium-term AI hypothesis: the medium-term is important from the perspectives of those who favor attention to near-term AI as well as those who favor attention to long-term AI. The paper analyzes medium-term AI in terms of governance institutions, collective action, corporate AI development, and military/national security communities. Across portions of these four

...as well as call for greater reconciliation and collaboration between these communities



Springer Link

Open Forum | Published: 01 June 2017

Reconciliation between factions focused on near-term and long-term artificial intelligence

Seth D. Baum

AI & SOCIETY **33**, 565–572(2018) | Cite this article

832 Accesses | **12** Citations | **2** Altmetric | Metrics

Abstract

Artificial intelligence (AI) experts are currently divided into "presentist" and "futurist" factions that call for attention to near-term and long-term AI, respectively. This paper argues that the presentist–futurist dispute is not the best focus of attention. Instead, the paper proposes a reconciliation between the two factions based on a mutual interest in AI. The paper further proposes realignment to two new factions: an "intellectualist" faction that seeks to develop AI for intellectual reasons (as found in the traditional norms of computer science) and a "societalist faction" that seeks to develop AI for the benefit of society. The paper argues in favor of societalism and offers three means of concurrently addressing societal impacts from near-term and long-term AI: (1) advancing societalist social norms, thereby increasing the

nature machine intelligence

Explore Content | Journal Information | Publish With Us | Subscribe

nature > nature machine intelligence > comment > article

Comment | Published: 07 January 2019

Bridging near- and long-term concerns about AI

Stephen Cave & Seán S. ÓhÉigeartaigh

Nature Machine Intelligence **1**, 5–6(2019) | Cite this article

5171 Accesses | **7** Citations | **55** Altmetric | Metrics

Debate about the impacts of AI is often split into two camps, one associated with the near term and the other with the long term. This divide is a mistake – the connections between the two perspectives deserve more attention, say Stephen Cave and Seán S. ÓhÉigeartaigh.

Clarifying our aims with this argument:

NOT to take a position on underlying debates over AI, its future trajectories, its risks, or underlying ethics -- (e.g. not 'who is right about AI?'; or 'what is more urgent?')

RATHER to reflect on social dynamics and self-facing narratives in 'AI ethics' epistemic community, to map the practical space for collaboration on AI policy:

- ***Emphasize importance*** of the community's coherence to its pursuit of policy impact in coming years
- ***Nuance debate, challenge the perception*** of unbridgeable disagreement:
  - Isolated spats do not represent widespread respectful engagement
  - 'Near-term vs. long-term' narrative overstates the diverse heterogeneity in people's actual positions on underlying sub-questions (cf. Prunkl & Whittlestone 2020)
  - Even where they exist, sources of disagreement need not be barriers to collaboration
- Propose a **principle for justifying and grounding more productive collaboration on AI policy,** wher there are shared policy interests

- The Gap
- The Bridge

- **The Gap**
- **The Bridge**
  - Why does this matter? The importance of the epistemic community organization
  - Potential grounds for divergence
  - Towards an ‘Incompletely Theorized Agreement’ on AI policy

- **The Gap**
- **The Bridge**
  - Why does this matter? The importance of the epistemic community organization
  - Potential grounds for divergence
  - Towards an ‘Incompletely Theorized Agreement’ on AI policy

Why does this matter? The importance of epistemic community organization for AI policy

- As noted, many of AI's challenges need urgent and responsive policies…
- Is the current 'responsible AI policy' community in a good position to deliver or advocate these?
  - The community is still young… In recent years, has seen…
  - Some policy **successes** (e.g. company biased dataset retraction, policing facial recognition moratoria [IBM,], etc. …),
  - Some policy **gridlocks** (e.g. lethal autonomous weapon systems)…
  - Sustained policy shifts take time and coherent epistemic community action on issue framing, consolidation of policy options, and implementation
  - Continued community fragmentation can undercut policy access or efficacy at key time

Historical lessons: structure of a field or community affects the ability to shape and influence policy downstream

- **Nanotechnology**:
  - High-profile adversarial 2003 'Drexler-Smally' debate publicly cemented an oversimplified caricature of the field, and a self-fulfilling split,
  - Was fuelled by exaggerated and constructed controversy: *"para-scientific' media created polarizing controversy that attracted audiences and influenced policy and scientific research agendas. [···] bounding nanotechnology as a field-in-tension by* ***structuring irreconcilable dichotomies out of an ambiguous set of uncertainties****."* (Kaplan & Radin 2011)
- **Ballistic missile defense arms control:**
  - Coordinated epistemic community, working on theoretical accounts of nuclear strategy, staked a position on the destabilizing risks of deploying ballistic missile defense.
  - Disseminated this understanding to both US and Soviet leadership, laid the foundations of 1972 ABM Treaty, first of its kind

The situation for AI policy—

- May soon face a key window to consolidate greater cooperation in responsible AI community:
- *Closing Window of Opportunity*
  - AI policy currently may allow some flexibility in terms of problem framings, governance instrument choice and design, and community organization…
  - …but field has a high likelihood of becoming more rigid as framings, public perceptions, and stakeholder interests in AI crystallize.
- *Risks are concrete and timely*
  - Global framings of AI at times seeing policymaker narratives of competition and 'arms races' –mishandling early AI cases today could yield decline in voice or policy influence

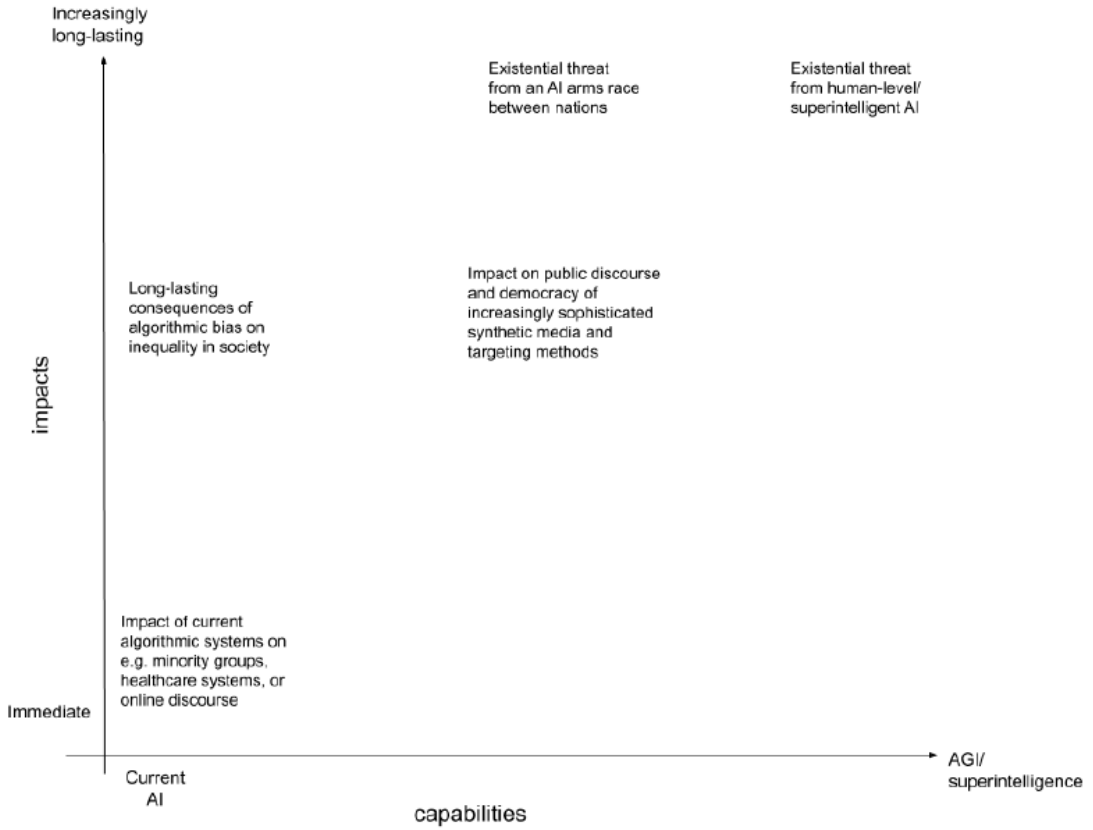
- **The Gap**
- **The Bridge**
  - Why does this matter? The importance of the epistemic community organization
  - Potential grounds for divergence
  - Towards an 'Incompletely Theorized Agreement' on AI policy

Examining potential grounds for divergence

- *Different epistemic or methodological commitments?*
  - Varying levels of tolerance for scientific uncertainty, and/or probability threshold for concern (e.g. 'how sure should we be about a problem to work on it?')
  - Differential interpretations of whether to admit, or how to weight, various forms of evidence (e.g. extrapolating from observed failure modes in existing ML systems; philosophical arguments; historical comparisons;...)
  - **BUT: while these disagreements are scientifically relevant... They don't foreclose pragmatic collaboration on issue areas where communities converge on shared AI policy goals**

Examining potential grounds for divergence – different pragmatic judgments?

- Different perceptions of dynamics of AI: (Prunkl & Whittlestone 2020)
  - *Capabilities*
  - *Impacts*
  - *Certainty*
  - *Extremity*
  - **But: on examination, experts' positions on these questions can be more complex and nuanced, doesn't simply breaking along 'near-term/long-term' axis (cf. Prunkl & Whittlestone)**



Examining potential grounds for divergence – different pragmatic judgments?

- Different perceptions of dynamics and path-dependencies of *governing* AI, e.g.
  - *How long-lasting are the consequences of near-term AI issues?* Long-termist may think 'short' → **on examination, many near-term issues likely to scale up, and/or cause 'turbulence' for long-term**
  - *How much leverage do we have today to meaningfully shape long-term AI policies?* Near-termist may say 'none' → **but on reflection, likely important path-dependencies**

Other sources of conflict also overstated: is there really inter-community competition over resources?...

- *Perceptions of resource competition?* (cf. Krakovna 2018)
  - ...Media attention? – 'long-term' researchers equally aversive to 'terminator' coverage
  - ...Funding? – not much current overlap in area-specific funding streams
  - ...Talent? -- Does not seem any meaningful 'brain drain' of AI researchers
- **In sum**: many perceived 'barriers' or trade-offs between these communities are arguably **overstated**, **contestable**, or simply **not so relevant to precluding policy collaboration** on joint policy programs

- **The Gap**
- **The Bridge**
  - Why does this matter? The importance of the epistemic community organization
  - Potential grounds for divergence
  - Towards an ‘Incompletely Theorized Agreement’ on AI policy

Towards an 'Incompletely Theorized Agreement' on AI policy

- Just because these barriers are not strong and/or may not *preclude* cooperation…
- …is there also space for positive, mutually productive opportunities for both communities to work on?
- And if so, how could such cooperation be justified and organized?

What is an Incompletely Theorized Agreement (ITA)?

- An '**Incompletely Theorized Agreement'** (ITA) is a principle from constitutional law (Cass Sunstein).
- An ITA allows a community to bypass or suspend any theoretical disagreement on matters where…
  - (1) the underlying disagreement appears relatively intractable [either given current information;] and;
  - (2) there is an urgent need to address certain shared practical issues.

Towards an ‘Incompletely Theorized Agreement’ on AI policy

“**Incompletely theorized agreements are a way to allow people to live together. Without them, social order would break down. They also show a form of mutual respect.**

By refusing to tackle people’s foundational commitments, citizens announce to one another: ‘Let us find a way forward, with civility and respect, while acknowledging, and making space for, uncertainty or profound differences on life’s deepest questions.’

– (Sunstein 2018)

ITAs have underpinned landmark achievements of global cooperation in history, including human rights…

"The philosopher Jacques Maritain relates how, at an 1948 meeting to present the **Universal Declaration of Human Rights** to the public, a member of the public expressed astonishment, that champions of extremely opposed ideologies had been able to agree on this list or rights. The committee's response to which was: "***Yes, we agree about the rights, but on condition that no one asks us why***"."

[Sunstein 2018]

An ITA in AI Policy? Shared areas of strategy-relevant *research*

- Into the **general levers of (government) policy formation** around AI
- Into the **relative efficacy of various policy levers** for AI governance (e.g. codes of ethics; naming-and-shaming; hard law…).
- Into better ways to **scrutinize and mitigate undue influence** of various stakeholders (private, government) on AI ethics (research) programs
- Into the question of '**social value alignment**'—
  - "Drawing important continuities between the work of the fairness, accountability, transparency and ethics community, and work being done by technical AI safety researchers, we suggest that more attention needs to be paid to the question of **'social value alignment' - that is, how to align AI systems with the plurality of values endorsed by groups of people, especially on the global level."** (Gabriel & Ghazavi 2021)

An ITA in AI Policy? Shared areas for joint *policy actions* & *proposals*

- Shape debates over the **appropriate scientific culture or norms** around considering the impact and dissemination of AI research in advance
- Early global regulation of-/ bans on **military uses of AI**
- Policy interventions aimed at **preserving the integrity of public discourse** and informed decision-making in the face of AI systems.
- Policies to secure citizens' (political) autonomy and independence from unaccountable '**perception control**' [e.g. computational propaganda; 'value lock-in']
- Institutional design choices to ensure **'scalability' or 'adaptability' of governance institutions to changes** in AI capability and use over both the near- and long-term

Pursuing an ITA on AI policy: benefits

- Improve research community cohesion and cross-fertilization
- Multi-faceted but united 'responsible AI' community: to present policymakers with an epistemic community delivering integrated and aligned policy proposals, and prevent politicization of research;
- Longer policy shelf life: to tailor more 'general' policies which need not assume further advances in AI capabilities, but which are also not susceptible to 'obsolescence' if or when such advances do occur;

Pursuing an ITA on AI policy: limitations

- Incompletely theorized agreements could prove brittle, if AI or its challenges change, in ways that misalign community's preferred policies.
- An incompletely theorized agreement is a 'stopgap' measure, not a general ideal or permanent fix.
- A sloppily formulated ITA on a given AI policy issue could inhibit effective action rather than enable it, e.g.--
  - Solidifying surface-level agreement on vague general principles or values, rather than particular policies (e.g. 'certification scheme for AI products with safety tests X, Y, Z').

In sum

- We do not propose ITAs as an unambiguously valuable tool across all AI policy cases…
- ITAs do have potential drawbacks or trade-offs, which the community should consider before invoking them in any particular policy area
- However, we propose sober scrutiny of perceived barriers that would stand in the way of productive collaboration
- It also encourages the exploration of opportunities for shared research or policy action, on the argument that such work can be justified and grounded in an 'incompletely theorized agreement'

Thank you!

Springer Link

Original Research | Open Access | Published: 15 January 2021

Bridging the gap: the case for an 'Incompletely Theorized Agreement' on AI policy

Charlotte Stix & Matthijs M. Maas

AI and Ethics (2021) | Cite this article

1873 Accesses | **35** Altmetric | Metrics

Abstract

Recent progress in artificial intelligence (AI) raises a wide array of ethical and societal concerns. Accordingly, an appropriate policy approach is urgently needed. While there has been a wave of scholarship in this field, the research community at times appears divided amongst those who emphasize 'near-term' concerns and those focusing on 'long-term' concerns and corresponding policy measures. In this paper, we seek to examine this alleged 'gap', with a view to understanding the practical space for inter-community collaboration on AI policy. We propose to make use of the principle of an 'incompletely theorized agreement' to bridge some underlying disagreements, in the name of important cooperation on addressing AI's urgent challenges. We propose that on certain issue areas, scholars working with near-term and long-term perspectives can converge and cooperate on selected mutually