Artificial Intelligence and Responsible Innovation

Miles Brundage

Consortium for Science, Policy, and Outcomes

Virtual Institute of Responsible Innovation

Overview

Responsible AI innovation in context

Framework for responsible innovation in general

Application to AI and lessons from other scientific/technological fields

Open questions and role of philosophy and theory

Attention to the topic

"As creators of the new science and technology of AI, it is our joint responsibility to pay serious attention [to its social consequences]."

-Poole and Mackworth (2010)

"[W]e cannot divorce AI research from its ethical consequences."

- Stuart Russell and Peter Norvig (2009)

IJCAI panel on "What if we succeed?"

But existing work/literature has various limitations

Not generally operationalized(able?) in undergraduate/graduate training

Often focused on what **not** to do as opposed to how to benefit society maximally

Often robot-centric

Often focused on "downstream" issues
(Though with important exceptions, such as the EPSRC Principles of Robotics)

My argument

Lessons can be learned from the theoretical framework for responsible innovation developed by scholars in science and technology policy, science and technology studies (STS), philosophy of science, and other fields

As well as case studies in nanotechnology, synthetic biology, and geoengineering over the past decade...

In order to:

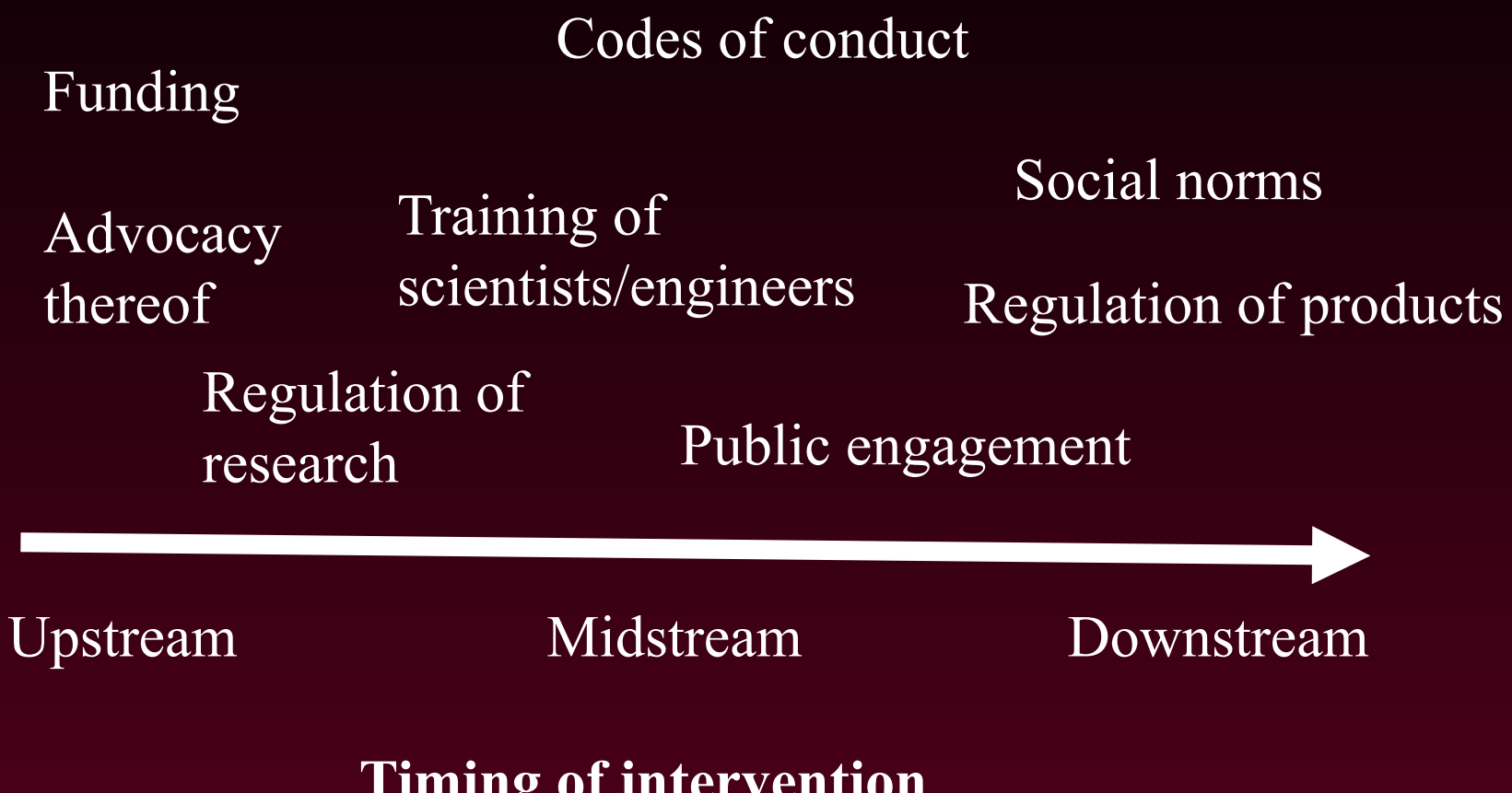
- **Clarify the nature of/relationships between work already done**

- **Identify future opportunities**

Responsible Innovation & Technological Governance

Timing of intervention

Responsible Innovation & Technological Governance



Responsible Innovation & Technological Governance (AI)

Advocacy for AI funding in general

Advocacy of funding specific research paradigms/programs

Teaching students about ethical issues

Machine ethics, Friendly AI

Robot ethics, Regulation re: responsibility attribution

Application-specific concerns/advocacy

Upstream Midstream Downstream

Timing of intervention

Why does responsible innovation (and this talk) focus on the midstream?

Collingridge's dilemma

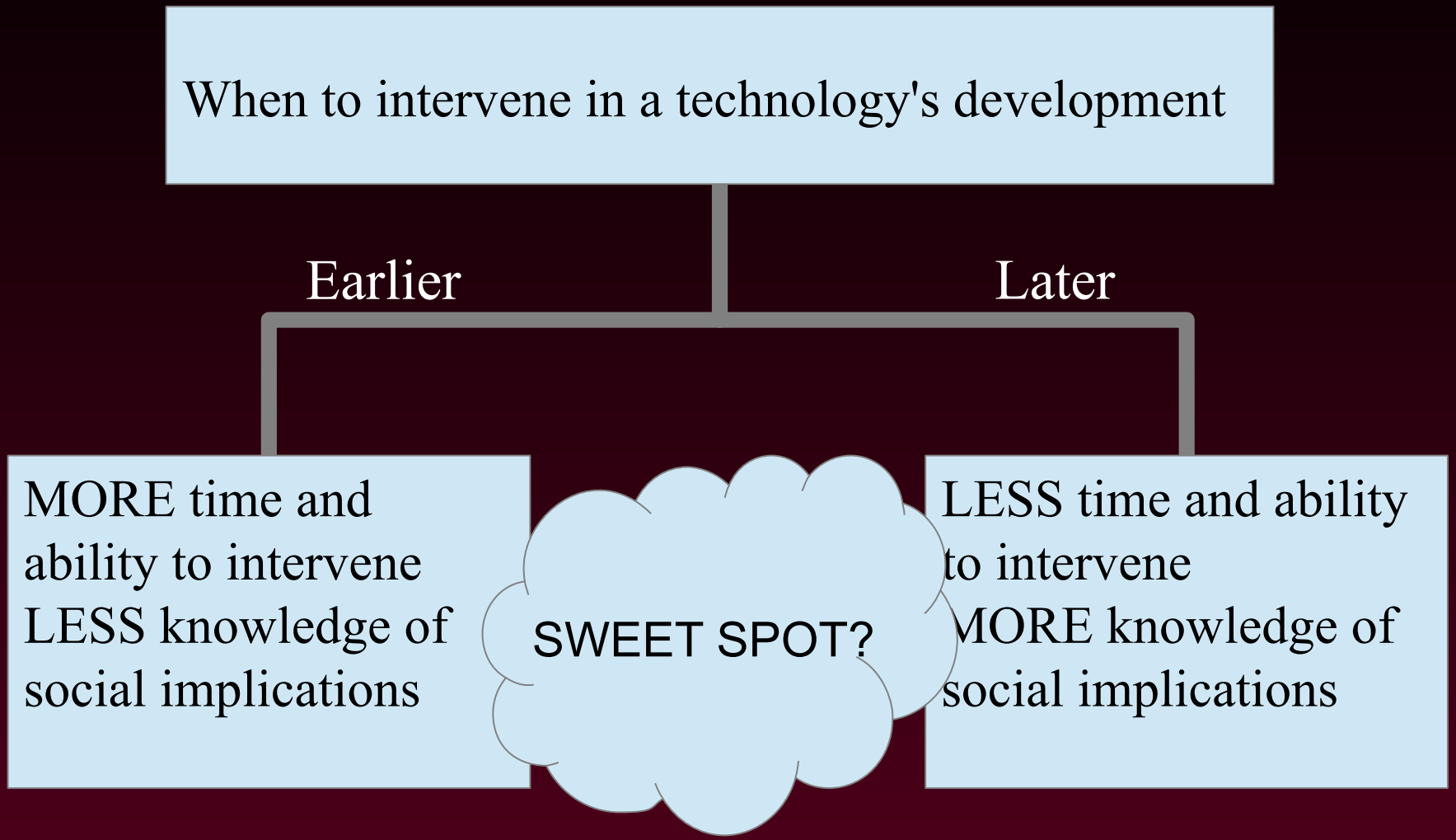
When to intervene in a technology's development

Collingridge dilemma

When to intervene in a technology's development

Earlier	Later
MORE time and ability to intervene **LESS** knowledge of social implications	**LESS** time and ability to intervene **MORE** knowledge of social implications

Collingridge dilemma



Researchers have concentrated, specialized knowledge

About AI capabilities, plausible future trajectories, risks/opportunities, etc.

Vast space of possible AI designs, paradigms, goals, etc.

Scientific curiosity/rigor and current funding regimes underconstrain the development of AI, opening a space for explicit normative debate by experts and the public

Responsible Innovation: One Definition

"Responsible innovation means taking care of the future through **collective** stewardship of science and innovation in the present." - Jack Stilgoe et al. 2013 (emphasis added)

Responsible Innovation: 4 Dimensions

Anticipation

- "Anticipation prompts researchers and organizations to ask 'what if...?' questions (Ravetz, 1997), to consider contingency, what is known, what is likely, what is plausible and what is possible." - Stilgoe et al. 2013

Responsible Innovation: 4 Dimensions

Reflexivity

- "Reflexivity, at the level of institutional practice, means holding a mirror up to one's own activities, commitments and assumptions, being aware of the limits of knowledge and being mindful that a particular framing of an issue may not be universally held." Ibid.

Responsible Innovation: 4 Dimensions

Inclusion

- "[I]ncluding new voices in discussions of the ends as well as the means of innovation." Ibid.

Responsible Innovation: 4 Dimensions

Responsiveness

- "Responsible innovation requires a capacity to change shape or direction in response to stakeholder and public values and changing circumstances." Ibid.

Many tools to choose from...but it's not obvious what “responsibility” entails in a given field

Anticipation
- Foresight
- Technology assessment
- Horizon scanning
- Scenarios
- Vision assessment
- Socio-literary techniques

Inclusion
- Consensus conferences
- Citizens' juries and panels
- Focus groups
- Science shops
- Deliberative mapping
- Deliberative polling
- Lay membership of expert bodies
- User-centered design
- Open innovation

Reflexivity
- Multidisciplinary collaboration and training
- Embedded social scientists and ethicists in laboratories
- Ethical technology assessment
- Codes of conduct
- Moratoriums

Responsiveness
- Constitution of grand challenges and thematic research programs
- Regulation
- Standards
- Open access and other mechanisms of transparency
- Niche management
- Value-sensitive design
- Moratoriums
- Stage-gates
- Alternative intellectual property regimes

Responsible Innovation in AI: 3 Questions for Researchers

First, how could different kind(s) of AI affect society, and how should this affect research goals in the present?

Responsible Innovation in AI: 3 Questions for Researchers

First, how could different kind(s) of AI affect society, and how should this affect research goals in the present?

Second, what domains should AI technology be applied to/how urgently, and what mix of basic and applied research is optimal from a societal perspective?

Responsible Innovation in AI: 3 Questions for Researchers

First, how could different kind(s) of AI affect society, and how should this affect research goals in the present?

Second, what domains should AI technology be applied to/how urgently, and what mix of basic and applied research is optimal from a societal perspective?

Third, what does the public want from AI, and what do they and various stakeholders such as policy-makers, educators, etc. need to know?

Desired AI Characteristics

How could different kind(s) of AI affect society, and how should this affect research goals in the present?

"Minimally, we are morally responsible for those things we intend to bring about. … While this is widely accepted, it is a difficult question under which circumstances and to what extend we should be responsible for unintended consequences." - Douglas 2009

Desired AI Characteristics (cont'd)

Thinking humanly
Acting humanly
Thinking rationally
Acting rationally (Russell & Norvig 2009)

Transparent
Predictable
Not deceptive
Robust against manipulation (EPSRC 2010, Bostrom & Yudkowsky 2013)

Human-level intelligence (Nilsson 2005)
Augment human intelligence (Hoffman et al. 2011)

Affective,
Moral (Wallach and Allen 2010, Scheutz today)
Moral patient (Scheutz 2012) or not (Bryson 2013)

Applications

What domains should AI technology be applied to/how urgently, and what mix of basic and applied research is optimal from a societal perspective?

Finite time/resources
Exploration/exploitation
Control vs. breadth of impact
Path dependence
Shaping public expectations/demands
But...constraints (Nourbhakhsh 2013)

Applications (cont'd)

What domains should AI technology be applied to/how urgently, and what mix of basic and applied research is optimal from a societal perspective?

Search and rescue
Elderly assistance
Lethal autonomous robots
Socially assistive robots
Medical assistant systems
Sustainability

Engagement

What does the public want from AI, and what do they and various stakeholders such as policy-makers, educators, etc. need to know?

Why engage?

Intrinsic (publicly funded research, influence on people's lives, educate policy-makers/educators/citizens)

Instrumental (head off negative reactions, get useful feedback to maximize impact)

Engagement (cont'd)

What does the public want from AI, and what do they and various stakeholders such as policy-makers, educators, etc. need to know?

Three historical stages of research on public engagement (Wilsdon and Willis 2004):

Public understanding of science
From deficit to dialogue
Moving engagement upstream

Higher-level questions

- Who is responsible for ensuring (which aspects of) responsible innovation?
- Role of professional associations vs. researchers vs. funding agencies, etc.
- Tensions between funding and responsibility (Nourbakhsh 2013)
- Best level(s) of abstraction
- Obligation vs.supererogation
- Best tools for anticipation, and their limitations
- Relationships between stages of development

Philosophy, Theory, and Responsible Innovation in AI

- Consciousness
- Plausibility/antecedents/nature of an intelligence explosion
- Possible sequences, if not timeline, of occupations being automated
- Space of possible minds and possible long-term convergence between short-term research paradigms

Conclusion

There is an extensive literature on responsible innovation and related concepts (technology assessment, engineering ethics, anticipatory governance, socio-technical integration, etc.)

Applying some of the findings of this literature to AI suggests some fruitful possible areas of research and public engagement, though there are many big open questions

Bibliography

Bostrom, N., Yudkowsky, E. 2013 The Ethics of Artificial Intelligence, in *Cambridge Handbook of Artificial Intelligence*.

Bryson, J. 2013. “Patiency is Not a Virtue: Intelligent Artefacts and the Design of Ethical Systems.”

Douglas, H. 2009. *Science, Policy, and the Value-Free Ideal*.

EPSRC panel, 2010, Principles of robotics, http://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/Pages/principlesofrobotics.aspx

Hoffman, R. et al. 2012. *Collected Essays on Human-Centered Computing, 2001-2011* (Washington, DC: IEEE Computer Society Press).

Nilsson, N. 2005. “Human-Level Artificial Intelligence? Be Serious!,” *AI Magazine*, 25th Anniversary Issue (American Association for Artificial Intelligence). .

Norvig, P. and Russell, S. 2009. *Artificial Intelligence: A Modern Approach (3rd edition)* (Upper Saddle River: Prentice Hall).

Nourbakhsh, I. 2013. *Robot Futures* (Cambridge: MIT Press).

Poole, D., Mackworth, A. 2010. *Artificial Intelligence: Foundations of Computational Agents.* http://artint.info/

Scheutz, M. 2012. The affect dilemma for artificial agents: should we develop affective artificial agents?” IEEE Transactions on Affective Computing.

Stilgoe et al 2013. Developing a framework for responsible innovation. *Research Policy*.

Wallach, W., Allen, C. 2010 *Moral Machines: Teaching Robots Right from Wrong*

Wilsdon, J. and Willis, R. 2004. *See-through Science*. http://www.demos.co.uk/publications/paddlingupstream

Thanks!

miles.brundage@asu.edu