A conversation with Miles Brundage 4 April 2014

Participants

- Miles Brundage—PhD Student in Human and Social Dimensions of Technology at Arizona State University
- Nick Beckstead—Research Fellow, Future of Humanity Institute (FHI); Trustee, Centre for Effective Altruism

Summary

Purpose of the call: I contacted Miles to learn about the field of responsible innovation and how it relates to potential global catastrophic risks from technologies like artificial intelligence, synthetic biology, and nanotechnology.

Why this person: My FHI college Daniel Dewey recommended that I speak with Miles Brundage because he is someone who works in responsible innovation and knows a lot about FHI's work.

We discussed subdivisions within these fields, the policy impact of responsible innovation and related fields, overlap between FHI concerns and concerns of people working in responsible innovation and related fields, and similarities and differences between approaches. FHI works with longer time horizons and focuses more on analyzing risks associated with specific technologies. Responsible innovation and related fields focus on shorter time horizons and focus more on building robust capacities to govern new technologies in general and engaging technologies in their early stages. Miles suggested that general frameworks for integrating social and ethical concerns might usefully be applied in fields like artificial intelligence, synthetic biology, and nanotechnology in order to reduce potential global catastrophic risks from these technologies but may be insufficient over a longer time horizon.

How is the field of responsible innovation similar to or different from fields like emerging technology governance and related fields?

Some terms used to describe different and overlapping areas of research in this space are "science and technology studies," "emerging technology governance," "technology assessment," "science and technology policy,"   "participatory technology assessment," and "responsible innovation." Much of the work in this area is done by social scientists and humanists, rather than natural scientists and engineers. The term "responsible (research and) innovation", sometimes abbreviated as RI or RRI, was popularized in the 2000s when a bureaucrat in the EU used it to pull together some of the aforementioned literatures into a cohesive framework. That's the term Miles uses to describe his work.

Some differences between these fields and FHI

Time horizons: FHI often focuses on 20+ year time horizons, whereas these fields tend to focus on shorter time horizons and current issues. In Miles' view, this is driven in part by differences in views about the feasibility of thinking so far out into the future.

These fields focus more on building institutional capacity to govern new technologies, whereas FHI is more interested in analyzing risks associated with specific future technologies and making recommendations themselves.

Unit of analysis: FHI tends to focus on the consequences of a few kinds of technology, whereas these fields focus on how to integrate social and ethical concerns into research and innovation in general regardless of existential risk considerations (though the analysis in both cases is often focused on specific scientific/technological domains).

When have people in this area influenced policy?

In the EU, paying attention to the ideas of this field has been the norm for a long time.

In the US, people working in this area influenced the National Nanotechnology Research Act, and that led to the creation of two Centers for Nanotechnology and Society, one at ASU and another at UC Santa Barbara. People who came from policy backgrounds and science and technology studies—such as Langdon Winter—testified before Congress in this context and affected policy.

However, this area hasn't been very successful at changing policymaking in the US. The BRAIN Initiative paid lip service to ethics and social implications, but didn't call for any robust integration of social and ethical concerns into the actual process of research. This is an illustration of how science in the US is largely run by scientists and politicians who don't take the concerns of this field very seriously and aren't aware of the ways in which the literature has moved past the simplistic models that have been used before, such as The Humane Genome Project, the model for which was essentially "have some ethicists think about things on the side." This field also didn't play a role in the Asilomar conference on recombinant DNA. There was once an Office of Technology Assessment in the US, but it has closed down.

Awareness of FHI's work in these areas

How much do you know about FHI's work?

Miles has read most of FHI's work on the future of AI, has attended an AGI workshop at FHI and talked with FHI staff, is generally familiar with MIRI's work on AI, and is familiar with the concept of "existential risk" and why FHI is interested in it.

How aware are people in this general area of FHI's work?

It's hard to say because there are many overlapping disciplines, and Miles isn't intimately familiar with all of them. Insofar as he fits into one discipline, it's science and technology studies, and responsible innovation is closely linked to that. People in science and technology studies emphasize the noninevitability and social context of developments in science and technology and the importance of being reflective about which technological trajectories we follow.

Miles would speculate that most people in this area haven't heard of FHI or lump it together with other varieties of future-oriented work, such as future studies, forecasting, "visioneering" (a term used by historian Patrick McCray to refer to, among others, Drexler) etc. that do 20+ year technological analysis. Their general perception is that forecasts/predictions have been unsuccessful and that (as the scenario planning literature also emphasizes) robustness against diverse, plausible scenarios and institutional capacity for adaptation is more important than knowledge of the specific future that ultimately comes to pass.

To what extent are researchers in these fields aware of and/or interested in issues related to potential global catastrophic risks from technologies like artificial intelligence, synthetic biology, and nanotechnology?

Synthetic biology

Miles didn't feel qualified to comment on specifics for synthetic biology. There are some people at ASU who work on this. E.g., Guston has published on this topic.

Artificial intelligence

People in this field are not very interested in the possibility of an intelligence explosion or existential risk from AI. People do write papers about machines as moral agents/moral patients, and the possible implications of sub-human-level and human-level intelligence. Some work in science and technology studies, responsible innovation, and related fields looks at the connections between science fiction, public understanding of science, and scientists' visions of the future. Apocalyptic visions of the future are typically seen as implausible by social scientists who see such visions as sociological objects of analysis to explain rather than serious risks to be engaged with. This may reflect a combination of ignorance about FHI claims and generalization from the history of failed attempts at technological foresight and the frequent recurrence of such fears throughout history.

Nanotechnology

Many people in this field are familiar with Drexlerian nanotechnology and debates between Drexler and Smalley. They see that historical episode as a cautionary tale against focusing on technological developments in the distant future. A paper on this issue is "A Critique of Speculative Nanoethics," though there is disagreement in these fields about how far to think ahead, the role of future-orientation in technological governance, etc. People who focus on nanotechnology governance do not emphasize Drexlerian molecular manufacturing, and focus instead on nearer term issues like nanoparticles. This is consistent with the field's general tendency to focus on what is currently happening in labs—or will soon be happening—rather than focusing on more speculative issues involving future technology. This partly driven by a general perception that it's very difficult to make progress on such questions.

Miles doesn’t feel qualified to comment on the plausibility of Drexlerian molecular manufacturing and its possible social impacts.

People in responsible innovations are sometimes interested in long-term distributive/social justice implications of technology, but are not very interested in existential risk.

Where could someone concerned about GCRs from emerging technology benefit by learning about this field?

There are a lot of literatures that might be relevant:

1. Science and technology studies might sharpen your understanding about what’s flexible or inflexible about technological change, how technological changes actually happens, and how technological change can depend on normative factors. It could generally give you a better sense of what is and isn’t inevitable in technological development.
2. Technology assessment has been going for decades, and they focus on methods for evaluating risks from future technologies.
3. Futures studies isn’t very respected academically, but they’ve written some interesting material about how to think about plausibility, probability, and possibility of future scenarios. Foresight and forecasting may be relevant. The most established method in this space is scenario planning, though it is may be more of a body of practice than a body of theory—its status as a field, practice, discipline, etc. is contested.
4. Technology ethics and the philosophy of technology.
5. Literature on “dual-use technologies.”

What about general science and technology policy relevant to making sure x-risk relevant tech goes well?

In 2013, Dan Sarewitz gave testimony before Congress. It included a synthesis of tips for integrating social concerns into scientific funding and governance. An issue is that existential risk is concerned with worst case scenarios, but a lot of technology assessment focuses on median cases. There aren’t existing frameworks that would handle risks that would allow very small, unorganized groups that might pose existential threats.

People do have general frameworks for responsible innovation. E.g. Rene von Schomberg has a framework and is heavily involved in EU policy-making around science. The gold standard framework within responsible innovation is “Developing a framework for responsible innovation” in the journal *Research Policy*. The key ideas are anticipation, reflexiveness, engagement, and responsiveness (AREA) and they are being institutionalized by the EPSRC in the UK. The four words don’t do justice to the richness of the underlying practices. This paper gives many examples of how this framework could be institutionalized. In Miles's view, frameworks are more promising as a tool for pointing us to good questions and some reasonable practices, but must be adjusted to particular cases or technologies.

Learning more

People to talk to

1. David Guston—Co-Director, Consortium for Science, Policy & Outcomes. Director, Center for Nanotechnology in Society. Professor, Political Science at ASU. Miles' advisor.

2. Dan Sarewitz—Senior Sustainability Scientist, Global Institute of Sustainability. Professor of Science and Society, School of Life Sciences, College of Liberal Arts and Sciences at ASU. Co-Director, Consortium for Science, Policy, and Outcomes.

3. Gary Marchant—Faculty Director and Faculty Fellow, Center for Law, Science & Innovation at ASU.

4. People who developed the responsible innovation framework in the EU: Jack Stilgoe, Richard Owen, and Phil McNaughton

5. Karlsruhe Institute for Technology Assessment and Systems Analysis (ITAS). Some people here are familiar with both technology assessment and responsible innovation.

6. People who work on "dual-use technologies"

7. Huw Price

8. Seth Baum

References

1. *Autonomous Technology* by Langdon Winner.
2. *The Whale and the Reactor* by Langdon Winner.
3. David Guston's paper "Understanding Anticipatory Governance" in the *Journal of Social Studies of Science*.
4. Browse the writings of Dan Sarewitz and David Guston at CSPO.org.
5. "On not forgetting futures" in the Journal of Responsible Innovation by Cynthia Selin. The references give a flavor of what foresight/future studies/forecasting might contribute to thinking about existential risk.
6. "Developing a framework for responsible innovation" in the journal *Research Policy.*

Questions sent to Miles prior to our conversation

1. How is the field of responsible innovation similar to or different from fields like emerging technology governance and related fields?

2. To what extent are researchers in these fields aware of and/or interested in issues related to potential global catastrophic risks from technologies like artificial intelligence, synthetic biology, and nanotechnology? (Even if they don't frame their interests the same way, I would be interested in the extent to which there is incidental overlap--e.g. proposed frameworks for

handling less extreme risks carrying over to global catastrophic risks, proposed general frameworks for the governance of emerging technology carrying over to global catastrophic risks from AI, synthetic biology, and nanotechnology.)

3. Who are the major figures in these fields? Who is working on topics most closely related to FHI's interests? Who is doing work that FHI might learn the most from?