RESPONSIBLE ARTIFICIAL INTELLIGENCE

Policy pathways to a positive AI future

CSPO
December 15, 2023

ANDREW MAYNARD
Professor of Advanced Technology Transitions
Arizona State University School for the Future of Innovation in Society
Artificial intelligence is not new, but …
Introducing ChatGPT

We’ve trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

OpenAI, November 30 2022
The year that **generative AI** changed the world

- A whole new way of interacting with tech has emerged
- We’ve gone from hardly anyone using generative AI, to a significant number of people using it in some way
- Students the world over have discovered innovative ways of completing assignments faster and better
- Established norms and approaches to teaching in higher education have been shaken to the core
- Unprecedented opportunities for AI-assisted learning at scale have emerged
- Researchers have been introduced to transformative new ways that AI can augment and accelerate discovery
- People who think for a living have begun to wonder if AI is going to replace them
- Institutions with no previous interest in AI have established artificial intelligence steering committees
- AI experts have become overnight celebrities
- Politicians and world leaders have met (repeatedly) with AI entrepreneurs

- National policies and regulations around AI have gone from near nothing to full blown pre-legislative discussion
- International governance around AI has equally been kicked into action
- There’s been a dramatic shift from talking about AI ethics to addressing AI risks
- There have been new and rapidly growing concerns around AI-driven threats to democracy and social stability
- Chatter around the exponential risks of AI has gone mainstream
- Responsible AI has become a thing
- Half the world seem to have become AI experts overnight (at least that’s how it sometimes feels)
- There’s been an acceleration in developing AI foundation models that are generalizable to different uses
- There’s been growing discussion of the possibility of artificial general intelligence
- And corporate politics at OpenAI have become more compelling than reality TV
AI Policy Analyst
Expert in AI policy analysis for US and EU, using factual and up-to-date info.
By Andrew Maynard

What are the latest updates on the EU AI Act?
How should the US respond to EU developments?
How do US and EU AI regulations differ?
What future trends are expected in AI policy?

Message AI Policy Analyst...

ChatGPT can make mistakes. Consider checking important information.
“In recent years, a new successful paradigm for building AI systems has emerged: Train one model on a huge amount of data and adapt it to many applications. We call such a model a foundation model.”

Stanford University Human-centered Artificial Intelligence (HAI) Center for Research on Foundation Models
Frontier AI models:
highly capable foundation models that could exhibit dangerous capabilities
Markus Anderljung et al. 2023
Artificial intelligence has the potential to **transform** how we …
Even without the more speculative projections of where AI may be taking us, novel, powerful, and potentially disruptive technologies are emerging that will require equally novel approaches to governance and oversight if they are to benefit society.
United Nations high level advisory board on Artificial Intelligence
October 26, 2023 (link)

US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence
October 30, 2023 (link)

EU Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI
December 8, 2023 (link)

OECD AI Principles
Updated November 7, 2023 (link)

United Kingdom AI Safety Institute
November 2, 2023 (link)

China Global AI Governance Initiative
October 20, 2023 (link)

Vatican AI & Peace
December 8, 2023 (link)
What is Artificial Intelligence?

... it’s complicated!
What is “Natural” Intelligence?
What is "Natural" Intelligence?
What is **Artificial Intelligence?**
What is **Artificial Intelligence**?

The ability of ... a machine to **deduce** how something works or behaves, based on information they collect or are given, their ability to retain and **build** on this knowledge, and their ability to **apply** this knowledge to bring about intentional change.

— Stuart Russell, as interpreted in Films from the Future (Maynard, 2018)
What is **Artificial Intelligence?**

An AI system is a machine-based system that, for explicit or implicit objectives, *infers*, from the input it receives, how to generate *outputs* such as predictions, content, recommendations, or decisions that can *influence* physical or virtual environments.

— OECD [Link]
What is **Artificial Intelligence**?

A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.

— 15 U.S.C. 9401(3) [Link]
What is “Artificial” Intelligence?
What are the risks?

It’s also complicated ...
What are the **risks**?

... there are capabilities associated with emerging and future AI that **could** be deeply disruptive to social, economic, and political prosperity.
The vast majority of today’s social, environmental and economic challenges are a **direct result** of past innovation.
INNOVATION

PROBLEM

SOLUTION
INNOVATION

PROBLEM

SOLUTION

CONSEQUENCES
“Progress”
Negative feedback loops
Intuitive/linear associations
Simple innovation models apply
Trial and error works
"Progress"

Time

$T_{\text{consequences}} \gg T_{\text{innovation}}$

Negative feedback loops
Intuitive/linear associations
Simple innovation models apply
Trial and error works

$T_{\text{consequences}} < T_{\text{innovation}}$

Positive feedback loops
Complex/nonlinear associations
Simple innovation models don’t apply
Unconventional thinking needed

?
Could advances in AI push society beyond the point where the timescale of consequences exceeds the timescale within which we can develop robust governance/policy solutions using conventional approaches?
**Tipping Points**

<table>
<thead>
<tr>
<th>Language</th>
<th>Money</th>
<th>Internet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture</td>
<td>Democracy</td>
<td>Electricity</td>
</tr>
<tr>
<td>Printing press</td>
<td>Fire</td>
<td>Harnessing the atom</td>
</tr>
<tr>
<td>Cyber space</td>
<td>Steam power</td>
<td>Artificial fertilizers</td>
</tr>
</tbody>
</table>

**Future-Transforming Possibilities**

<table>
<thead>
<tr>
<th>Artificial general intelligence?</th>
<th>Artificial awareness?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mimicking what makes us human</td>
<td>Beyond-human ability to leverage information</td>
</tr>
<tr>
<td>Artificial consciousness?</td>
<td>“Inaccessible islands” of discovery</td>
</tr>
</tbody>
</table>

**Unique Conditions**

- 8 billion people
- Extreme planetary coupling
- Limited resources
- Massive networking
- Ability to manipulate biological, material, and cyber systems

**Unprecedented Advances**

- “Inaccessible islands” of discovery
- Human-integrated technologies

**Novel and emergent risks**
Tipping Points
- Language
- Agriculture
- Printing press
- Cyber space

Money
Democracy
Fire
Steam power

Internet
Electricity
Harnessing the atom
Artificial fertilizers

Future-Transforming Possibilities
- Artificial general intelligence?
- Artificial consciousness?
- Artificial reasoning?

Unique Conditions
- 8 billion people
- Extreme planetary coupling
- Limited resources

Massive networking
Ability to manipulate biological, material, and cyber systems

Unprecedented Advances
- Mimicking what makes us human
- Beyond-human ability to leverage information
- “Inaccessible islands” of discovery

Navigating the AI Technology Transition
Navigating the AI Technology Transition

- Learning from past technology transitions
- Responsible Innovation
- Risk Innovation
- Agile Regulation
- Thinking Differently
Navigating advanced technology transitions: using lessons from nanotechnology

Andrew D. Maynard & Sean M. Dudley

As researchers, developers, policymakers and others grapple with navigating socially beneficial advanced technology transitions—especially those associated with artificial intelligence, DNA-based technologies, and quantum technologies—there are valuable lessons to be drawn from nanotechnology. These lessons underscore an urgent need to foster collaboration, engagement and partnerships across disciplines and sectors, together with bringing together people, communities, and organizations with diverse expertise, as they work together to realize the long-term benefits of transformative technologies.

In November 2020, OpenAI released ChatGPT—a public interface with the company's large pretrained transformer (GPT) large language model (LLM). To researchers working on foundation models and artificial intelligence more broadly, it was just one more step toward developing the next iteration of AI-based technologies. In this article, we consider ChatGPT and other forms of generative AI to illustrate the public's attention. It became clear that the world was in the thick of an advanced technology transition that few were prepared for.

What is perhaps not as clear is that there are insights from nanotechnology that are deeply relevant to navigating this and other advanced technology transitions, and lessons that developers, decision makers, and others would be wise to heed as they work toward realizing the emergence of new technologies that substantially benefit society.

The current wave of innovation around artificial intelligence has highlighted the profound changes associated with successfully realizing the benefits of advanced technologies. Despite their considerable potential to bring about positive change, emerging AI-based technologies threaten to displace jobs, disrupt education, undermine social norms, destabilize the economy, and even throw democratic principles and processes into disarray. As a result, developers and policymakers have been scrambling to navigate this increasingly complex landscape that is emerging between possible benefits and risks.

Generative AI, through, is only one of a growing number of transformative technologies that are on the horizon. These include, but are far from limited to, quantum technologies, advanced DNA-based technologies, neurotechnologies, and even the possibility of artificial general intelligence. Successfully navigating the transitions that these and other technologies are anticipated to bring about will require increasingly innovative approaches to technology policy, multi-sector, and multi-stakeholder collaborations and partnerships. In many ways, this is a landscape that is reminiscent of the early days of nanotechnology—and one that led to substantial advances in how advanced technology transitions can be navigated successfully.

Between the mid 1990s and the late 2000s, nanotechnology went from an esoteric concept to a driver of technological change that spanned domains ranging from materials science and biotechnology to risk, policy, and even philosophy. It represented a profound advanced technology transition, and one that remains ongoing. Nanotechnology—on the emergence of nanoscale science and engineering as a distinct field to be more precise—is not the first advanced technology transition to have deeply impacted global society. It was preceded by earlier generative engineering, the advent of the internet, digital computing, nuclear energy, electricity, steampower, and other past technological "revolutions." Yet in many ways it stands apart from those previous transitions in the nature, scope, and interconnectedness of the changes that were associated with it.

While the concept of intentionally manipulating and engineering matter at the atomic scale can be traced back at least as far as Richard Feynman's now famous talk, There's Plenty of Room at the Bottom given at the California Institute of Technology in 1959 (ref. 1), it wasn't until the 1990s that Eric Drexler popularized the idea of atomically precise manufacturing in his book Engines of Creation. At the time, Drexler's ideas were creative and inspirational. But they remain on the fringe of mainstream science until a group of US agencies began looking for a compelling concept to drive investment in research and development in the 1990s (ref. 2). What emerged was a vision for "the next industrial revolution" that took on a more prismatic and plausible set of ambitions for use on tuning novel properties and functionalities out of materials through their intentional design and manufacture at the nanoscale.
Responsible Innovation

- Anticipation
- Reflexivity
- Inclusion
- Responsiveness

There is currently a knowledge and practice gap between responsible innovation and responsible AI.

Developing a framework for responsible innovation

Jack Stilgoe, Richard Owens, Phil Macnaghten (2013). Research Policy
“... risk innovation frames risk as a threat to existing or future ‘value’ where value is broadly and multiply defined within personal, societal and organizational contexts.”

Why we need Risk Innovation
Orphan Risks

Social and Ethical Factors

Unintended Consequences of Emerging Technologies

Organizations & Systems

riskinnovation.org
Orphan Risks

Privacy
Social Justice & Equity
Social Trends

Social and Ethical Factors

Ethics
Worldview
Perception

riskinnovation.org
Orphan Risks

Bad Actors
Geopolitics
Governance & Regulation

Organizational Values & Culture
Reputation & Trust
Standards

organizations and systems

riskinnovation.org
Orphan Risks

- Black Swan Events
- Co-opted Tech
- Health & Environment

Unintended Consequences of Emerging Technologies

- Intergenerational Impacts
- Loss of Agency
- Product Lifecycle

riskinnovation.org
The Risk Innovation Planner helps identify and strategically address “orphan risks” – often-overlooked risks to success for which there are no agreed upon tools, standards, or mitigations already in place, and which if not planned for can easily blind-side an enterprise down the pike.

The Planner provides a quick yet effective way to identify, plan for, and evaluate progress against orphan risks which are relevant to your enterprise. With regular use of the Planner, your team will create strategies for success, building value and creating positive outcomes.

1. Identify three areas of value for your enterprise, your investors, your customers, and your community.

2. Circle the orphan risks that have the potential to impact, or pose a threat to, your priority areas of value. For reference, consult the Definition and Scenario cards.

3. Circle the 2-3 areas of highest value to focus on over the next few months.

Risk innovation approaches risks as a threat to value, or a threat to something of importance to your enterprise, your investors, your customers, or your community. Whether tangible or intangible, a current product or a future success, if it’s worth something to you or your stakeholders, it’s a value of yours. By starting with what’s more valuable to each of these areas, you can begin to more clearly see where and when orphan risks might have the most blindsiding impact.

Describe the specific way in which these risks threaten your priority areas of value and, by extension, your enterprise, investors, customers, and/or community.

Consider a few actions you can take throughout the next quarter to begin planning for your specific risks.

Taking small steps now will add up, helping you build strategies to plan for orphan risks and avoid blindsiding impact. Each action should address: What am I going to do, why am I going to do it, and how will I accomplish it? Actions should be specific, enough to complete within 2-4 weeks. For instance, read a book, plan to host a public engagement or other members of your organization, work on your orphan risk strategy, draft an orphan risk policy.

Quarterly Reflection: What actions were effective and worth the time and resources? How can you begin to integrate these actions into your risk planning strategy?

Thank you for completing The Risk Innovation Planner! For more information, please visit us at www.riskinnovation.org or email us at info@riskinnovation.org.

riskinnovation.org
Hypothetical
An AI company governed by a not for profit organization

**ENTERPRISE**
- Innovation and Research Excellence
- Safety and Ethical Standards
- Sustainability and Growth

**INVESTORS**
- Financial Returns
- Impact and Legacy
- Risk Mitigation

**CUSTOMERS**
- Access to Advanced Tools
- Reliability and Trust
- Empowerment and Capability Expansion

**COMMUNITIES**
- Equitable Access and Inclusion
- Societal Advancement
- Responsible Innovation

riskinnovation.org
Hypothetical
An AI company governed by a not for profit organization

**ENTERPRISE**
- Innovation and Research Excellence
- Safety and Ethical Standards
- Sustainability and Growth

**INVESTORS**
- Financial Returns
- Impact and Legacy
- Risk Mitigation

**CUSTOMERS**
- Access to Advanced Tools
- Reliability and Trust
- Empowerment and Capability Expansion

**COMMUNITIES**
- Equitable Access and Inclusion
- Societal Advancement
- Responsible Innovation

**VALUE**
- Ethics
- Social Justice & Equity
- Perception
- Black Swan Events
- Loss of Agency
- Governance & Regulation
- Organizational Values & Culture
- Reputation & Trust

riskinnovation.org
“A more agile, flexible approach to regulation is needed in order to seize the potential of the Fourth Industrial Revolution to change lives for the better.”

Agile Regulation for the Fourth Industrial Revolution A Toolkit for Regulators
World Economic Forum (2020)
Agile Regulation

- Anticipatory regulation
- Outcomes-focused regulation
- Experimental regulation
- Data-driven regulation
- Self- and co-regulation
- Joined up regulation
- International regulatory collaboration

Agile Regulation for the Fourth Industrial Revolution A Toolkit for Regulators
World Economic Forum (2020)
Measures to support innovation and SMEs

MEPs wanted to ensure that businesses, especially SMEs, can develop AI solutions without undue pressure from industry giants controlling the value chain. To this end, the agreement promotes so-called regulatory sandboxes and real-world-testing, established by national authorities to develop and train innovative AI before placement on the market.
Thinking differently about AI and **Decision Making**
There’s a growing need to develop a sophisticated understanding of near and far term threats and opportunities associated with AI that guides decisions with far-reaching consequences.
Some base assumptions

- We cannot stop the emergence of transformative AI — we can only guide it.
- There is tight coupling between people, society, and the future of AI.
- AI development without forethought and responsibility is more likely to cause harm than good.
- Intelligence is “just bits, all the way down.”
- Understanding risk as a threat to present and future “value” can help navigate complex technology transitions.
Uninformed, naive and irresponsible development of AI potential threatens value that’s important to us now.

Slow or throttled AI development could potentially threaten to value that we aspire to.
Advanced Technology Transitions
Advanced Technology Transitions

Long term threat
Long term opportunity
Near term threat
Near term opportunity
Education and AI

<table>
<thead>
<tr>
<th>LONG TERM</th>
<th>NEAR TERM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diminished critical thinking</td>
<td>Unfair advantages</td>
</tr>
<tr>
<td>Personalized learning at scale</td>
<td>Rethinking learning</td>
</tr>
</tbody>
</table>

THREAT

OPPORTUNITY
Education and AI

**Long Term**
- Diminished critical thinking
- Personalized learning at scale

**Near Term**
- Unfair advantages
- Rethinking learning

**Threat**

**Opportunity**
Education and AI

Mechanisms

1. Checks and balances
2. Responsible innovation
3. Principled innovation
4. Mindset change
5. Human-centric innovation

LONG TERM

Diminished critical thinking

PERSONALIZED LEARNING AT SCALE

NEAR TERM

Unfair advantages

RETHERINKING LEARNING
Research and AI

Mechanisms
1. New science of AI discovery
2. New discovery mindset
3. New tools
4. Governance
5. Responsible research
6. Principles innovation
7. Transdisciplinary innovation
8. Multistakeholder engagement
Social Cohesion and AI

Mechanisms
1. AI literacy
2. Countering complacency
3. Counter influence operations
4. Inclusive innovation
5. Transdisciplinary innovation
6. Responsible innovation
7. Principled innovation

Social collapse

Futures we aspire to

Mis- and disinformation

Healthy society

1. Countering complacency
2. Counter influence operations
3. Inclusive innovation
4. Responsible innovation
5. Transdisciplinary innovation
6. Principled innovation
7. AI literacy
AI is a transformative technology where the promise is profound and the risks are unclear. What is clear is that emerging approaches to policy will need to be as innovative as the technologies they address if we’re to succeed in ensuring a positive AI future.