THE CHANGING LANDSCAPE OF AI: LESSONS FROM A YEAR OF POLICY DISCOVERY

CONTENTS

INTRODUCTION

During 2023, Google DeepMind worked with various civil society and social impact organisations to co-design and deliver nine roundtables[1] that explored the opportunities and risks presented by the deployment of AI in key sectors. Each roundtable resulted in a written report.

Participants in the roundtables included representatives from civil society, academia, advocacy groups, governments, startups and the private sector. Some roundtables took place in person, while others were virtual or hybrid, with participants drawn from all over the world. Themes ranged from the broad to the specific, encompassing national security, disability and the future of work, education and more.

We're far from alone in convening these kinds of multi-stakeholder discussions. Rapid advances in AI-powered technologies have sharpened the focus of policymakers, civil society and the public at large on AI and its societal impacts. As an industry organisation working at the leading edge of AI research, Google DeepMind has a responsibility to demystify AI and provide insights in forums for policy discovery that include a broad range of voices. These insights must be reflective of the diversity of communities and sectors affected by AI and inclusive of the expertise of civil society, industry and academia. We're part of a larger ecosystem and our work – both the technical development of AI systems and our contribution to AI policy debates – must be reflective of and responsive to the rest of the ecosystem and the world beyond it.

This summary aims to surface and share the most pertinent questions, insights and perspectives that emerged from these roundtables, in the hope that they will serve policymakers' shaping of AI regulation. We also share what we learned through working with civil society and social impact organisations to convene these discussions and our plans to develop the programme further.

[1]In addition to those listed here, we contributed to a workshop with Brookings Institute on global governance, and sponsored an exhibition with UAL: Central Saint Martins (CSM) to design an experiential public exhibition imagining AI Futures. See more on both below. The IAS roundtable formed a Working Group, which has met twice to date and produced two outputs, with plans to continue its work.

KEY INSIGHTS

OPPORTUNITIES AND RISKS

- Now is the time to develop policy frameworks and ideas that work
- AI is most often an enabler, not the solution itself
- AI systems risk replicating and exacerbating existing biases, power imbalances and systemic inequalities…
- …But AI can also address inequalities, if designed with that goal in mind
- Safety risks associated with emergent capabilities are inherently challenging to manage
- AI accountability and oversight infrastructure is nascent

PATHWAYS AND IDEAS

- Equitable data is a prerequisite for equitable AI
- AI-related expertise and skills are needed across all sectors of society
- Broader participation in the development of AI systems is crucial, albeit hard to achieve at scale
- Building public trust in AI is essential for delivering benefits at scale
- Agile governance is needed for long term ecosystem alignment and accountability

WHAT WE LEARNED

- Working with organisations with specific expertise in each topic enriched the discussions
- Pre-reading was influential in how conversations were framed
- A range of organisations had advantages and trade-offs
- Imagining the future productively is hard
- Recency bias means that generative AI can dominate conversations
- Connections were forged and ideas generated between groups

ORGANISATIONS & REPORTS

Untapping the Potential of AI in Science
AI offers enormous potential to transform and accelerate scientific research. This report explores what steps AI labs, research organisations, and policymakers could take to untap the potential of AI in science.

Skating to Where the Puck is Going: Anticipating and Managing Risks from Frontier AI Systems
This report assesses the current trajectory of AI development and the emergent risks that come with augmented capabilities and increasingly general purpose models. It discusses measures that industry and governments should consider to guide these technologies in a positive and beneficial direction.

Data Equity and Data Governance in AI
Datasets are fundamental to training and operating AI systems. The recent acceleration in the pace of AI development and use has implications for data – how much is needed, and the extent to which its quality impacts society. This report explores the challenges of data equity and the kind of data governance that would be needed for equitable outcomes from AI.

Using AI to Improve Employment Outcomes for People with Disabilities
This report explores the current barriers to employment for people with disabilities and shares ideas for AI tools that can improve these outcomes, specifying considerations that must be incorporated into the development of such tools.

AI and the Future of Food Security
This report explores the potential for AI to help increase global food security by enabling businesses, governments and NGOs to develop more resilient, crisis-proof food systems in the face of an increasingly changed climate.

AI and the Future of Learning
Could artificial intelligence help unlock a revolution in how and when we learn across our lifetimes? This report explores the dynamic landscape of lifelong learning and AI, the opportunities and challenges of AI-enabled personalised learning, and the exciting future developments that have the potential to reshape education.

Model Machines: Exploring AI Futures Through Art and Design
What might a world with AI look and feel like, and how AI might benefit humanity? Google DeepMind worked with UAL: Central Saint Martins to design an experiential public exhibition entitled *Model Machines* to explore these questions.

AI Policy and Governance
The AI Policy and Governance Working Group (AIPGWG) represents a mix of sectors, disciplines, perspectives, and approaches, and organizes events that engage the public in discussion of the societal implications of AI.

In June, the AIPGWG held a roundtable at the Institute for Advanced Study in Princeton, NJ, and provided recommendations as a joint response to the National Telecommunications and Information Administration's (NTIA) request for comment on AI accountability policy. In September, it submitted recommendations[2] concerning global AI governance to the United Nations Secretary-General's Envoy on Technology.

Scan the QR Code to read each of the individual roundtable reports:

PROGRAMME BACKGROUND AND PROCESS

Background

Advances in AI don't automatically benefit those who need them most. In the case of a scientific breakthrough like AlphaFold,for example, it took intentional partnership with the Drugs for Neglected Diseases Initiative (DNDi) to understand how and where AlphaFold's predictions could help those suffering from neglected diseases, which disproportionately affect people in the Global South.

Our experience of working together to share the benefits of breakthroughs like AlphaFold has, in part, contributed to our practice of meaningfully engaging with civil society, experts, practitioners and advocates so that we can learn from one another. This kind of collaboration is essential to designing, building and deploying safe and beneficial AI in different contexts and geographies.

It was with this in mind that, in 2022, we worked with the Aspen Institute to convene two multi-disciplinary discussions about how to think about and enable equitable AI. A summary of those conversations was published as the report 'A Blueprint for Equitable AI' in early 2023. Our goal in exploring this concept was to see **how AI might facilitate and drive inclusion**, as opposed to reinforcing bias and historical patterns of injustice. The roundtables we worked on this year all build on questions and themes that emerged from collective consideration of how the benefits of AI might be distributed equitably.

The Process

As **AI will ultimately affect all sectors and societies**, there is a vast range of topics to explore. With a focus on learning through policy discovery, we were guided by the organisations we worked with to prioritise sectors where AI holds significant promise, and where collaboration and thoughtful policy design are needed to ensure benefits are distributed equitably. Additionally, we sought to understand pressing governance questions that must be answered in order for AI to have a positive impact in the world.

For each roundtable, we worked with an organisation with deep expertise in the selected subject and connections to a diverse range of experts and practitioners, including those working on the ground. Designing the roundtables was a collaborative process, though the organisations took the lead given their embedded expertise.

Representatives from Google DeepMind observed and participated in each roundtable, always aiming to foreground others' perspectives as well as contributing our own, especially in the case of voices that are sometimes overlooked or excluded. We also contributed to the reports organisations crafted to capture the learnings.

While each conversation took a slightly different form, discussions overall explored the following questions:

How should AI be built?

Where should AI be used?

How should AI be governed?

Who should be accountable?

OPPORTUNITIES AND RISKS

Navigating the rapidly changing landscape of AI

Any discussion about AI must consider the risks presented by the technology. Proven and potential benefits of AI are also an increasingly central part of national and international conversations about major policy and societal challenges and how we might address them. The ways in which key opportunities and risks were presented in the roundtables were illuminating and suggested certain shifts and reframings that could strengthen governance and collaboration.

Now is the time to develop policy frameworks and ideas that work

The sense that there's a **window of opportunity to influence policymaking** emerged strongly from almost every roundtable. To an unusual degree, government officials at the highest level are deeply engaged in questions of AI governance. The call to action to civil society, academia and those creating the technology was clear.

At the same time, **balancing appropriate urgency with interrogation** will be key to creating strong and inclusive policies. Particularly in the case of the sector- and topic-specific roundtables, participants emphasised that rushing to apply AI to complex systems without fully understanding the historic challenges and dynamics that characterise these systems is a recipe for disaster. Education, food security and disability in the workforce are complicated topics, with challenges and intricacies that predate the advent of AI-powered technologies. Likewise, the forces that make the global data landscape unequal, or contribute to a possible stalling in 'disruptive' scientific research,are multifaceted and can't be entirely addressed by AI. Additionally, while many of the challenges and opportunities of AI are new, sectors like healthcare and education have grappled with the integration of technological advances before. It's important **to engage with and learn from the past, as well as recognising the ways in which AI is novel**.

> "AI needs to help reframe the paradigm, not support it."

Royal Society of Arts, AI and the Future of Learning Roundtable Report

British Science Association, Untapping the Potential of AI in Science Roundtable Report

AI is most often an enabler, not the solution itself

For most individuals and groups, **AI is one strand in a web of social systems** that impact their lives and opportunities. Applying AI to systems that are fundamentally broken won’t lead to equitable outcomes. As one roundtable participant put it, “AI needs to help reframe the paradigm, not support it.”

In several contexts, participants advised that we focus on the potential of AI to enable solutions and fuel progress, rather than viewing it as the answer to long-standing and complex policy challenges. Another roundtable participant emphasised that, “AI is only a tool, not a magic bullet.”[2] Likewise, **AI was often framed as a complement to human intelligence and capabilities**, as opposed to a replacement. One roundtable participant suggested that, in the context of scientific discovery, “greater interoperability between human intelligence and machine intelligence would unlock the potential to scale AI research projects and embed them within the research process.”

In education, **AI-powered tools can help teachers** by freeing them from administrative burdens and allowing them to focus on the aspects of education that they alone can provide to students.

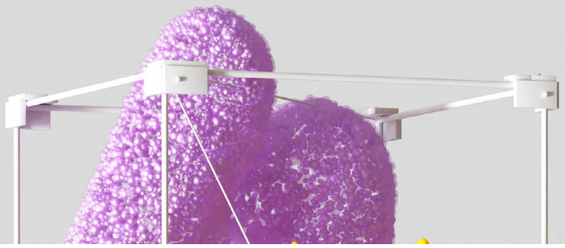
By **improving accessibility and removing existing barriers to participation**, AI could enable a world in which everyone can bring their unique skills and perspectives to the workplace, regardless of disabilities.

In science, AI can be a **transformative tool for scientists, especially when thought of as a ‘co-pilot.’** While the potential impact of AI on scientific research is significant, it’s still arguably most productive to think of AI as a complement to human scientists and to focus on maximising the complementary nature of the relationship.

In climate, AI could play a pivotal role in **forecasting early warning systems and monitoring complex dynamics, from natural disasters to food security challenges**. Working alongside governments and civil society organisations, these insights could lead to the development of better response policies and processes for prevention and mitigation of environmental, market and systemic challenges.

AI developers, policymakers and civil society organisations alike could further orient their ‘north stars’ to the societies they want to help build and contribute to, as well as the technologies they can discover and the risks they need to manage. This paradigm shift would help optimise for beneficial outcomes for all.

[2] Gro Intelligence., ‘AI and the Future of Food Security’



AI systems risk replicating and exacerbating existing biases, power imbalances and systemic inequalities…

Examples of biassed machine-learning outputs and the harm they cause, particularly to historically marginalised groups, are well established. Broad societal risks, including **bias, misinformation, surveillance and inequitable access to benefits**, were prominent in all the conversations. For example:

- In workplace contexts, well-intentioned **deployment of AI systems and tools could exacerbate existing barriers for people with disabilities** if not rigorously tested with the involvement of those they are intended to help.
- In scientific research, **disparities in access to data and training will lead to unequal benefits from scientific progress** - among scientists and societies alike.
- Education practitioners highlighted the risks of **applying AI too hastily to student assessment**, as automated systems may reinforce biases inherent in the training data and impact students' university prospects.

Throughout the discussions, participants pointed to the continued **existence of a global digital divide**. At the food security roundtable, it was emphasised that the economics of AI do not always promote equity of access. One participant asked how we might create the right market to **scale solutions to lower costs for participating in the use of AI**.

Risks that AI will exacerbate existing inequalities are closely linked to questions of data equity. The **datasets currently available for training AI systems are not fully representative of the global population** and historical biases are often baked into the data used to train models. Data is neither generated nor collected at the same rate, or to the same standard, globally. **Access to data, as well as its quality**, can create and reinforce power imbalances. Not all data is publicly available and even when it is, **availability does not guarantee accessibility**. From nonstandard data schemas to lack of data legibility and literacy, there are lots of reasons why data may not be usable. Many existing initiatives designed to diversify datasets are currently operating only on small scales, while other collaborative data governance efforts only go some way to solving the issue.

Participants noted that **loss of privacy is a significant near-term risk** from higher participation in AI technologies. The risk that data might be misused in a way that betrays individual characteristics to that person's detriment, or to a company's outsized gain, surfaced throughout our discussions.

DEFINING DATA EQUITY

There are many definitions of data equity. One way to conceptualise it is as "a set of principles and practices to guide anyone who works with data (especially data related to people) through every step of a data project through a lens of justice, equity, and inclusivity. And equity is not just an end goal, but also a framing for all data work from start to finish."

…But AI can also address inequalities, if designed with that goal in mind

As noted, the fact that AI-powered systems and technologies entrench existing biases and inequalities featured prominently in the roundtable discussions.

Conversely, it emerged that there are ways in which **AI could help increase equity** – but only if developers, policymakers and citizens make that a goal. With good intentions, all groups risk misrepresenting the hopes and concerns of affected communities if they fail to engage directly with those communities. A propensity to make assumptions, and for those assumptions to gain traction in the debate, was evident. Sometimes the policies that would actually promote equity are counterintuitive. **Conscious and active efforts must be made to realise the potential for AI to support, rather than undermine, equity.** For example:

- AI could **increase employment and earnings for people with disabilities**, making the workforce more inclusive. By developing tools that increase targeted recruiting of people with disabilities or that help disabled job-seekers find the right opportunities, AI-powered tools could also improve a disabled employee's experience on the job.
- **Personalised learning** could help level the playing field for children and adults alike, improving educational outcomes across the board. Generative AI opens up the potential for more genuinely personalised learning than we've seen in the past, while large AI models offer scope to improve the capabilities of AI-powered tools for teachers.
- **Re-thinking the collection, governance, architecture and management of data** could unlock benefits across virtually all applications of AI. Within **science**, for example, data is the key difference between fields that use AI and those that don't. To date, structural biology and genomics have led the way in terms of AI-enabled advances, partly because the life sciences have more established experience and frameworks for dealing with data. Many other domains, from materials science, physics and chemistry, to healthcare and criminology, have lots of unstructured data and making that data accessible and usable could hold the key to AI-driven advances in those fields and more.
- AI could help **democratise access to expert knowledge**, ensuring that no one is left behind due to the inability to afford expert support. A participant in the food security roundtable highlighted the example of small scale farmers in India applying for subsidies. To apply, farmers need to read and complete long legal documents, which they often don't have the literacy and/or legal knowledge to understand and are therefore excluded. AI could help address this challenge by supporting farmers to understand and complete their applications.

"While some risks will be evident from the capabilities of the models themselves, many more will result from the way those models interact with their environments and society at large."

Centre for Security and Emerging Technology (CSET), Skating to Where the Puck is Going: Anticipating and Managing Risks from Frontier AI Systems Roundtable Report

Safety risks associated with emergent capabilities are inherently challenging to manage

At the CSET roundtable, we proposed **five ways** in which existing AI systems are currently being augmented that give rise to concern - namely multimodality, tool use, deeper reasoning and planning, larger and more capable memory and increased interaction between these systems and users.[4]

Of these, **tool use** was considered by some to be the most concerning near term capability, in part because of "the wide array of potential actions it enables, as well as the potentially high stakes and unpredictable outcomes of those actions."[6]

AI systems could also be used to enable adversarial attacks. The potential for AI to gain advanced cognition skills (e.g. long-term planning or error correction) and to develop situational awareness (e.g. an awareness of its own testing, development or deployment) were both explored. Biosecurity risks could result from an AI system being given knowledge about biological production, and in the most extreme scenario, an AI system could develop novel synthetic weapons.

Each of these capabilities has advantages, including the potential to make AI systems more useful, transparent or beneficial. But downsides and risks were identified in relation to each. It is important to be vigilant in monitoring these augmentations, including in the context of the environments in which they operate. As the CSET roundtable report points out, "model capabilities that may seem concerning may in fact be harmless — for instance, if a model produces instructions on how to create a chemical weapon, but the necessary reagents are strictly controlled. On the other hand, ways in which a model may seem too limited to cause harm may be misleading — for instance, if a model's context window is too short to develop and carry out a mass spear-phishing attack in one go, but tool use and memory allow the model to call external programming libraries, save files to refer back to later."[7]

DEFINITION: MULTIMODALITY

A multimodal AI system is one that is capable of receiving multiple types of input (such as text, images, audio, or video) or generating multiple types of outputs.

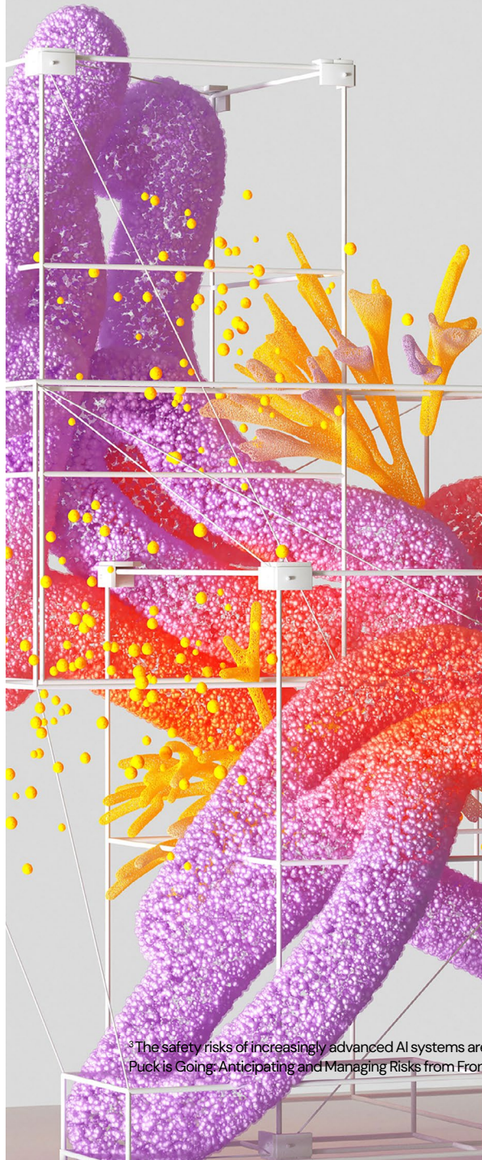
DEFINITION: TOOL USE

Tool use refers to the capability of AI systems to interact with a broader environment outside of the AI itself relatively autonomously through a set of tools, such as internet plug-ins. [...] Providing an AI system with a user-interface control would allow it to take actions on sites across the web, not simply retrieve and generate text information. This is significant because soon, frontier AI systems will output not just static language, images, audio, or video, but will likely have the capability to interact with the open internet or user data and applications.[5]

[4] These five points are drawn from material presented by Matthew Botvinick (Google DeepMind) at the Centre for Security and Emerging Technology roundtable.
[5] Defined using quoted material from: Centre for Security and Emerging Technology, 'Skating to Where the Puck is Going: Anticipating and Managing Risks from Frontier AI Systems'.
[6] Centre for Security and Emerging Technology, 'Skating to Where the Puck is Going: Anticipating and Managing Risks from Frontier AI Systems'.
[7] Centre for Security and Emerging Technology, 'Skating to Where the Puck is Going: Anticipating and Managing Risks from Frontier AI Systems'.

"By necessity, the concerns being raised are speculative, since they relate to the development of novel capabilities that have only been observed in primitive forms. However, waiting to take action until it is definitively proved that AI systems do have the capabilities under contemplation would be irresponsible, given the potential severity of the harm that could result."

IAS, Comment of the AI Policy and Governance Working Group on the NTIA AI Accountability Policy Request for Comment



TYPOLOGY OF RISKS[3]

The work of Laura Weidinger et al. captures the ethical and social risks of harm across a broader spectrum of modalities than are considered in this work.

- Discrimination, exclusion and toxicity
- Malicious uses
- Misinformation harms
- Information hazards
- Human-computer interaction harms
- Automation, access and environmental harms

[3] The safety risks of increasingly advanced AI systems are explored in detail in the report from the Centre for Security and Emerging Technology, 'Skating to Where the Puck is Going: Anticipating and Managing Risks from Frontier AI Systems'.

AI accountability and oversight infrastructure is nascent

The **lack of established testing, accountability and oversight mechanisms** was identified as a risk, albeit one that is being addressed. For policymakers, **achieving alignment on the right AI accountability structures**, including evaluation, access and disclosure processes, is essential.

The appropriate alarm structures for when a dangerous capability is emerging was also discussed. **Time is a key variable here**. Decision-makers should disambiguate between two separate timelines: 1) when the capability first emerges and 2) when it actually causes significant harm. This distinction is critical when allocating responsibility because policy responses to AI harms will be more dependent on the second timeline (when actual harm is expected to occur) than the first (when the capability has been discovered but has not yet led to harm). Without sufficient oversight and attention to this emergency timeline, policymakers leave societies vulnerable to harm.

Participants strongly encouraged policymakers to monitor the development of AI systems that may be able to recreate themselves without human oversight (also known as autonomous replication). While model evaluations or 'dangerous capability evaluations' were discussed as an important way to interrogate AI systems, a warning emerged from the roundtables not to rely too heavily on their results. Evaluations are usually designed to investigate one particular risk or element, meaning that worrying model competencies outside of that scope might be missed.

The fast tempo of AI progress means that even a **shared language for grappling with policy challenges is still emerging**. Certain key terms have been ushered into use before they've been fully defined. For example, there's currently no agreed definition for the term 'frontier models,' while 'data equity' can be understood in a variety of ways. Even apparently self-explanatory terms, like 'education' were questioned during the roundtables. For example, what does it mean to be educated and does this change as society changes?

> **"The use of AI undoubtedly poses an array of complex challenges, but policymakers should not be dissuaded from taking action to address emerging concerns by supposed tensions between innovation and safety, the evolving nature of the field, or the relatively nascent mechanisms for accountability."**

IAS, Comment of the AI Policy and Governance Working Group on the NTIA AI Accountability Policy Request for Comment

PATHWAYS AND IDEAS

Delivering the benefits of AI responsibly and equitably

While risks were naturally prevalent, the roundtables also surfaced many ideas for seizing and sharing the benefits of AI. Many of the individual roundtable reports include detailed, context-specific ideas. The following themes recurred throughout the discussions and serve as examples of practical insights for further exploration and action.

Equitable data is a prerequisite for equitable AI

Data is crucial to the training, testing and deployment of AI models. While more data is being generated than ever before, the fact that the global data landscape is highly unequal has implications for the impact AI will have on society when deployed. The governance of data - its collection, quality, robustness, representativeness and readiness - must be recognised as fundamental to the effective governance of AI.

Participants emphasised the need for all stakeholders to understand the complexities surrounding the equitable use and collection of data, including concepts of data access and ownership, data hygiene, data quality and robustness, and bias, fairness and representation in datasets. To meet the opportunities and needs of societies as AI advances, **frameworks for governing data and enabling transparency** and access may need more than updating - they may need to be reimagined at a fundamental level.

Specifically, opportunities were identified to:

- Develop **standards and institutions** to organise, document and share data effectively.
- **Incentivise transparency and reporting requirements** to build public trust.
- **Invest in AI literacy and skills** to develop a new generation of data practitioners – especially in the Global South – and support subsidies for lower-income communities to enter the field of AI.
- **Incentive talent** into the less glamorous 'service layer' of data architecture and management.

> **"The AI industry has an opportunity to bring individuals from these communities (and those who may have multiple or intersectional identities) together to acknowledge the various aspects and needs of the human experience. By creating a more inclusive design process, AI tools can become more inclusive and resonant for the users they serve."**

Claypool Consulting, Using AI to Improve Employment Outcomes for People with Disabilities Roundtable Report

> “The changing landscape necessitates identifying and developing the complementary skills and values essential for effective interaction with AI.”

Claypool Consulting, Using AI to Improve Employment Outcomes for People with Disabilities Roundtable Report

AI-related expertise and skills are needed across all sectors of society

The need to support, develop and evenly distribute AI-related expertise and skills featured prominently in the roundtables. **AI literacy will be an essential factor in realising the potential of AI**, mitigating the risks associated with its use and building public confidence in the technology. One prevailing sentiment is that the centre of gravity of AI expertise currently rests too much in the private sector and that it’s critical to expand it.

AI expertise and skills are not limited to coding and computer science. Increasingly, the development and responsible deployment of AI systems will require skills in social science, business, the humanities and data analysis, as well as the ability to **collaborate across disciplines**. What and how much people need to understand about AI will differ across sectors and roles, but shared definitions and language will be needed to facilitate collaboration and build equitable systems. For example:

- **Governments** need sufficient technical expertise to understand and balance the risks and opportunities posed by AI in order to design policies and interventions that are fit for purpose.
- “An **ecosystem for AI in science** requires skills in the curation of data, servicing AI models, applying models to scientific questions and using the outputs.”
- **Civil society organisations** have access to data that could be used to understand and tackle societal challenges, but may lack the expertise and capacity to turn these into datasets for AI/ML systems.

The widely recognised need to **expand the diversity of the global talent pool**, including by supporting universities and companies in the Global South to attract and retain local talent, was also highlighted throughout the conversations. It’s important to attract a range of people with a diversity of backgrounds to work in AI, including in the building of datasets, and to embed within communities to facilitate equitable data collection and use.

Broader participation in the development of AI systems is crucial, albeit hard to achieve at scale

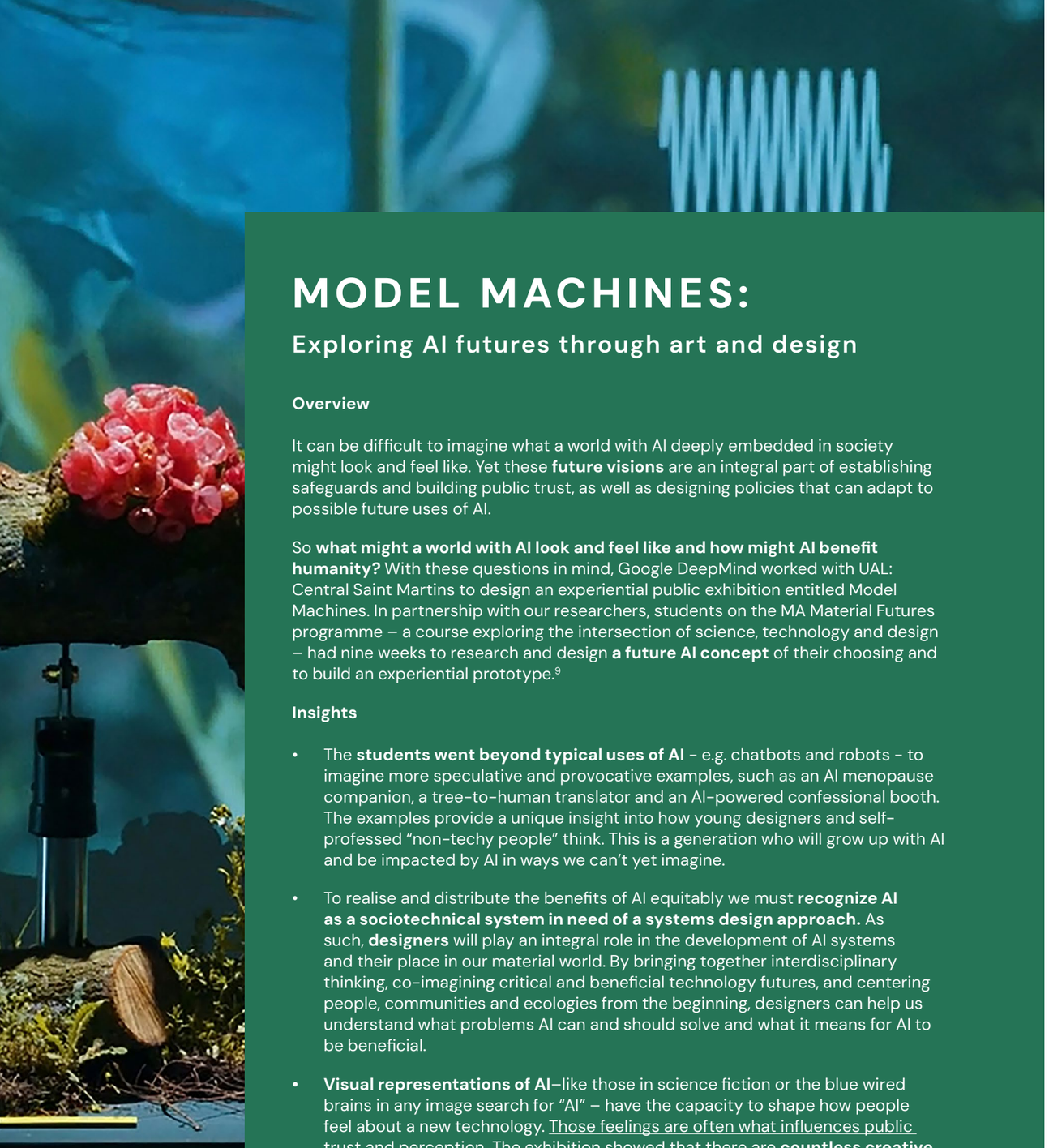
To understand how AI might deliver benefits across sectors, it’s essential to look beyond AI companies and government and engage with a wide range of stakeholders across the private sector, civil society and academia. One participant quoted the disability rights motto, **“Nothing about us, without us.”**

The challenge is **how to do this effectively at scale** in a way that ensures people are recognised and compensated for their contributions. One roundtable participant reflected that an essential question is: “How might we create the right market to scale solutions at lower costs for participating in AI?”[8] Throughout the roundtables, several design principles emerged for guiding inclusive engagement:

- **Communities who will be most impacted by the development and deployment of AI must have a voice in the conversation**. For example, if an AI tool for supporting students in classrooms is rolled out, students, parents and teachers should all be consulted as part of the development process as they each have unique perspectives to share in regards to learning outcomes. Institutions that collect data must educate the people whose data they’re collecting on how those services will help them.
- It will be increasingly important to **include thinkers and designers** who are skilled in imagining and analysing future implications to mitigate short-term thinking.
- Policies that **support the adoption of a ‘sandbox’ approach** will encourage swift and thoughtful experimentation of the use of AI in different sectors.

The challenges and opportunities of participatory AI have been experienced and discussed among civil society and community organisations for several years and explored in scholarship including the 2022 paper, ‘Power to the People? Opportunities and Challenges for Participatory AI’ by Birhane et al.

[8] Gro Intelligence., ‘AI and the Future of Food Security’



MODEL MACHINES:

Exploring AI futures through art and design

Overview

It can be difficult to imagine what a world with AI deeply embedded in society might look and feel like. Yet these **future visions** are an integral part of establishing safeguards and building public trust, as well as designing policies that can adapt to possible future uses of AI.

So **what might a world with AI look and feel like and how might AI benefit humanity?** With these questions in mind, Google DeepMind worked with UAL: Central Saint Martins to design an experiential public exhibition entitled Model Machines. In partnership with our researchers, students on the MA Material Futures programme – a course exploring the intersection of science, technology and design – had nine weeks to research and design **a future AI concept** of their choosing and to build an experiential prototype.[9]

Insights

- The **students went beyond typical uses of AI** - e.g. chatbots and robots - to imagine more speculative and provocative examples, such as an AI menopause companion, a tree-to-human translator and an AI-powered confessional booth. The examples provide a unique insight into how young designers and self-professed "non-techy people" think. This is a generation who will grow up with AI and be impacted by AI in ways we can't yet imagine.
- To realise and distribute the benefits of AI equitably we must **recognize AI as a sociotechnical system in need of a systems design approach.** As such, **designers** will play an integral role in the development of AI systems and their place in our material world. By bringing together interdisciplinary thinking, co-imagining critical and beneficial technology futures, and centering people, communities and ecologies from the beginning, designers can help us understand what problems AI can and should solve and what it means for AI to be beneficial.
- **Visual representations of AI**–like those in science fiction or the blue wired brains in any image search for "AI" – have the capacity to shape how people feel about a new technology. Those feelings are often what influences public trust and perception. The exhibition showed that there are **countless creative uses for AI, most of which we could never imagine on our own** and that aren't currently represented in the public imagination.
- The exhibition was a **creative way to engage industry leaders** and the general public in a conversation around how the world might prepare for Artificial General Intelligence (AGI). It provided a new way for leaders to engage with the topic and explore their own personal views of the technology.

[9] Photo by: Maël Hénaff, Project by: Scarlett Mercer and Yu Watanabe.

Building public trust in AI is essential for delivering benefits at scale

Delivering the benefits of AI to society will require public trust in AI systems. In addition to the risks and challenges described above, it's evident that **trustworthiness is not inherent to AI systems and tools**. AI developers, civil society and policymakers each have a pivotal role to play in interrogating these systems, developing accountability mechanisms and earning public trust in those that are deployed. Specifically, this might include **promoting greater transparency and public awareness of the technology, setting and enforcing clear guardrails and accountability measures for industry and showing how governments can work together** even when their policies differ.

At several of the roundtables, participants from various national governments explicitly asked for ideas and frameworks to address the governance challenges and economic and societal opportunities of AI. While there were no definitive answers, participants made valuable suggestions:

- There is scope to lay out a **clear vision** of how AI could be used, providing the public with a sense of where we are now, where we are headed, how AI could be a tool to help us get there and how risks are being managed. During the CSET roundtable, participants noted that "most debates about the future of AI are anchored in current technologies — such as today's LLM-based chatbots — but lack a clear sense of which tools or capabilities might bridge the gap between the present and the future."[10]
- **Information asymmetries must be addressed through accountability and transparency measures.** As the AIPGWG shared, "designers and deployers of AI must demonstrate that their products are safe and effective — and therefore merit the public's trust — through iterative accountability mechanisms that span the full development and deployment lifecycle and address risks related to both highly specialised and more general purpose AI systems."
- There was a perceived opportunity to corral efforts and energy into the development of **best practices and standards** in multiple areas, from the responsible collection and governance of data, to effective red-teaming and evaluation. Rather than aiming for homogeneity of systems and standards, **interoperability should be the priority**.
- **Evidence-based stories** about the benefits of AI in society are needed to build public trust. There is significant hope that AI could help solve some of the world's greatest challenges, for example around climate and public health, but participants shared the need to know and hear about a) the positive impacts already being realised and b) what can realistically be expected for the future. There is a role for all stakeholders to play in commissioning evidenced research, including scientometrics and economic and social impact analyses of how AI is being used in various fields and its impacts.

DEFINING FRONTIER MODELS

A common definition for frontier models and an understanding of the associated risks are still being established. Absent a precise definition, the 'AI research frontier' or 'frontier models' may be thought of as referring to AI models that are comparable to or slightly beyond the current cutting edge.[11]

[10] Centre for Security and Emerging Technology., 'Skating to Where the Puck is Going: Anticipating and Managing Risks from Frontier AI Systems'.

[11] Centre for Security and Emerging Technology., 'Skating to Where the Puck is Going: Anticipating and Managing Risks from Frontier AI Systems'.

Agile governance is needed for long term ecosystem alignment and accountability

Alignment on governance and AI accountability measures, including evaluation, access and disclosure processes, was called for by participants from across sectors and groups. The need for **accountability measures to be dynamic and iterative**, to respond to emergent risks and opportunities, was emphasised strongly. To this end, it will be essential to understand the policy levers at different stages of the AI development lifecycle, as shown in the table overleaf from the CSET roundtable. It is equally important to consider the roles that industry, academia, civil society and the public sector can play at each of these stages. Levers include:

- **Iterative and sociotechnical accountability mechanisms:** As highlighted by the AIPGWG, "responsibility for accountability in the design and deployment of AI systems and tools must begin with technology developers, but industry, academia, civil society, and the public sector each have a key role to play in the development of an effective AI accountability system."

- **Incentive structures to optimise for safety and benefits.** AI labs can be incentivised to test and evaluate their frontier AI systems and report dangerous capabilities to oversight bodies. Likewise, incentives can promote the testing of education tools in a sandbox before they are used in schools. In the case of AI labs, as discussed in the CSET roundtable, incentives could be explored via procurement requirements, establishing industry certifications for frontier AI systems and loosening liability in exchange for transparency.

- **Keeping pace with developments** in a fast-moving field will continue to be a challenge. Regular updates of mechanisms, models, evaluations and audits will be needed to anticipate and meet new risks. In this regard, there may be lessons to be taken from the field of cybersecurity and the mechanisms that the sector has in place globally to keep pace with the development of new capabilities.

- **Advocating for an interoperable governance framework** would allow countries, regions, and regulatory bodies to work effectively together. As recommended by the AIPGWG, "Policy interoperability also enables jurisdictions to set their own policy priorities – in line with local needs and the specific context relevant for determining thresholds for fairness, responsibility, and safety – while still aligning with a globally recognized set of core commitments and accountability and safety mechanisms."

- **International Institutions may have an important role to play** in enabling effective global governance and ensuring advanced AI systems benefit humanity, as highlighted in the paper on 'International Institutions for Advanced AI' that was presented at the Brookings Institution roundtable. When it comes to the possible establishment of a new institution for governing AI, there is a range of governance functions that could be performed at an international level to address key governance challenges, ranging from supporting access to frontier AI systems to setting international safety standards.

- **Greater access to and distribution of infrastructure,** such as compute, data centres and cloud infrastructure, will help ensure that power is not overly concentrated within industry, and that more businesses and communities can use and benefit from AI-enabled technologies. An example of this is the UK's Future of Compute Review.

- **Listening as well as sharing expertise.** While demonstrating the capabilities, opportunities and risks of emergent systems will continue to be a core responsibility of AI companies and labs, the obligation to listen to and collaborate with outside experts and affected communities should be prioritised too.

Table 1.
Policy levers are categorized by goal along various stages of the AI development lifecycle.

	Model development	Initial deployment and proliferation	Deployment in narrow contexts	Broader societal impacts
Visibility and understanding	Pre-training disclosure	Pre-deployment disclosure	Incident sharing	Measuring and forecasting societal impacts
Defining best practices	Risk assessment guidelines, evaluations, and standards for developers	Deployment decisions informed by risk assessments	Sector-specific guidelines, e.g. assurance requirements for AI in high-stakes contexts	Guidelines for ongoing monitoring and risk assessment
Incentives and enforcement	Public funding for safety research Licensing and/ or liability for development	Licensing and/ or liability for development Export controls Open source restrictions	Domain-specific regulation	

Source: Centre for Emerging Technology and Security

WHAT WE LEARNED

Lessons and limitations

The process of participating in the roundtables yielded insights into the importance of engaging with civil society and building a thriving AI ecosystem across stakeholder groups. These insights will inform the next phase of our work and our hope is that they may be useful for others too as we continue to explore together how to make public engagement impactful.

Working with organisations with specific expertise in each topic enriched the discussions

Working with civil society and social impact organisation to convene each conversation meant that we had a **more diverse mix of participants and a richer discussion than we could have achieved alone**, or with any single organisation. Working with expert groups also helped us to grapple with the broad scope of the topics we wanted to discuss. Some topics, like data and safety, proved so fundamental and expansive that knowing where to focus required collaboration. Agendas were designed to ensure conversations would not be dominated by a single perspective or a narrow set of risks.

Pre-reading was influential in how conversations were framed

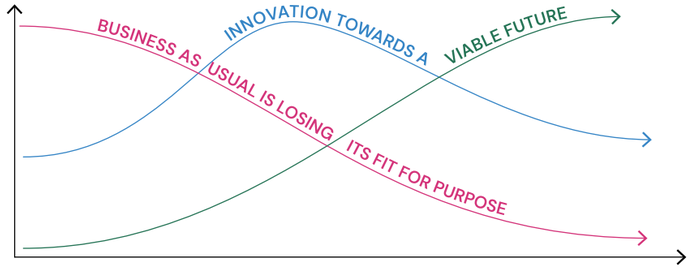
The process of compiling the right pre-reading for participants was instructive and also required close collaboration with the convening parties. Without wanting to overly influence the conversation's direction, it was important to create shared context and establish a baseline understanding of the questions under discussion and the goals of the roundtables. Thoughtfully composed pre-reads were important to achieving this balance.

A range of organisations had advantages and trade-offs

While working with different organisations for each roundtable meant that we could delve deeper into the topics we wanted to explore with a broader range of people, a single organisation may have lent more consistency to the roundtables. We chose not to be overly prescriptive in how the conversations and their outputs were structured and two conversations diverged even from the roundtable format[12]. That approach was effective, but a single organisation may have increased cohesion and allowed parallels and thematic insights to emerge more readily.

Imagining the future productively is hard

Future scenarios are challenging to engage with, especially when the trajectory and ideal destination are unclear and the technology is evolving at pace. In some cases, it was easier to reach alignment on existing issues than on how best to handle possible future scenarios. In others, discussion focused on potential future governance mechanisms without bridging the gap between these ideas and the current state of play. Established frameworks for thinking about the future, like the Three Horizons model[13], were drawn on in some of the roundtables to productive effect.



The Three Horizons (3H) model is an adapted framework that helps us conceptualise long-term social change – in the case of our roundtable series, related to education and learning.

Recency bias means that generative AI can dominate conversations

Generative AI dominated many of the conversations because it is so prevalent in the public imagination at present. Yet AI capabilities extend far beyond it and many debates have much broader relevance. **Communication and education by AI developers is a prerequisite for thoughtful civil society engagement** because not all sector-specific experts will also be AI experts.

Connections were forged and ideas generated between groups

Participants who met through the roundtables identified areas of shared interest and forged connections, leading them to pursue ideas together that we hope will yield concrete and beneficial outputs.

[12] One was a workshop we contributed to with Brookings Institute on global governance, and the other was an exhibition we sponsored with UAL: Central Saint Martins (CSM) to design an experiential exhibition imagining AI Futures.

[13] Leaders Quest, 'Three Horizons Model': https://leadersquest.org/three-horizons-introduction/#:~:text=It%20charts%20Horizon%201%2C%20the,our%20desired%20future%20a%20reality. and McKinsey & Co., 'Three Horizons Model'. "https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/enduring-ideas-the-three-horizons-of-growth"

CONCLUSION

If the development and deployment of AI systems and frontier models are steered responsibly and in positive directions, the collective benefits for society will be huge.

Hearing the perspectives of diverse practitioners and communities proved extremely valuable - sometimes in surprising ways - as this report has aimed to show. Policy debates can easily be dominated by the priorities and perspectives of those with the strongest voices and biggest platforms. But clearly, **the hopes and fears of different communities relating to AI are not homogenous**. In some parts of the world, fears about labour market displacement are more pressing than safety risks. Among populations that tend to be seen - justifiably - as at risk of exclusion from the benefits of AI in Global North-centric debates, there is actually a great deal of hope and excitement about AI. And while representation is rightly a key focus of most debates about data, there are also valid reasons why certain groups and individuals may want to exercise their right **not** to have their data used. These are just a few examples of perspectives that provide food for thought and demonstrate that questioning assumptions is a crucial foundational step in inclusive policy making.

The **pace of development in AI technologies means that policy discovery and development needs to evolve quickly**, too. This set of roundtables was the beginning of a process, with the next phase to be informed by the insights generated and the nuanced conversations sparked by these exploratory conversations. We are committed to continuing to work with civil society and community-led organisations to ensure they have a voice in fast-moving conversations that will impact them and the people they represent.

Some of the most important outcomes from these roundtables will stem from the connections they facilitated. A number of participants are now exploring collaborations, which we look forward to seeing develop. **Creating alignment and a shared understanding of certain key terms and processes** will help strengthen the collective capacity of the ecosystem to operationalise and honour requirements like the White House Commitments. To this end, we plan to convene further conversations, as well as targeted working groups, to work towards establishing the definitions and standards we need to enable progress towards safe and beneficial AI. We further plan to develop and support initiatives that address some of the key opportunities and challenges that emerged from these conversations, in partnership with civil society, the academic community, policymakers and our industry peers.

We hope that by sharing this report publicly, we can spark continued conversations and catalyse collective action towards inclusive policies that support equitable AI.

Authors and Acknowledgements

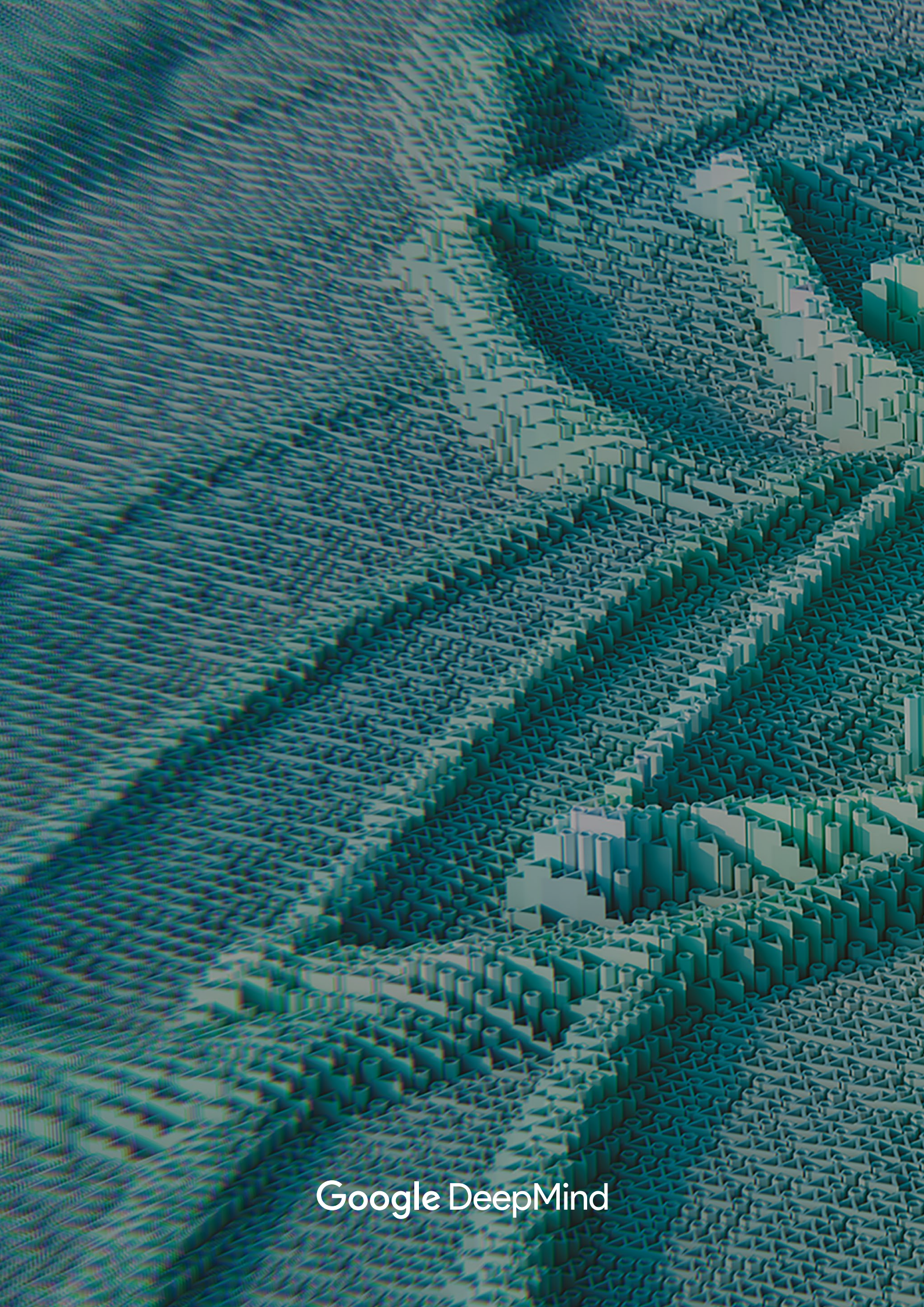
Authors

Eimear Nolan is Policy & Public Engagement Manager at Google DeepMind

Rachel Foley is Policy & Public Engagement Manager at Google DeepMind

Acknowledgements

We would like to acknowledge Lucy Lim, Lewis Ho and Dorothy Chou from Google DeepMind and Jo Sparber, Melissa Hinkley, Aidan Peppin and Malcolm Glenn for their editorial support. Thank you to Studio La Plage for their design support. Finally, we would like to thank all of the organisations and roundtable participants for their contributions to the discussions and individual roundtable reports that this summary has drawn from.


Google DeepMind