Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan: Responses

DeepMind

DISCLAIMER: Please note that the RFI public responses received and posted do not represent the views or opinions of the U.S. Government or any entity within the U.S. Government. We bear no responsibility for the accuracy, legality, or content of the responses and external links included in this document.
DeepMind Response: National Artificial Intelligence Research and Development Strategic Plan

At DeepMind we believe that AI’s extraordinary potential will only be realized if its development and deployment uphold appropriate ethical standards and if it is purposefully directed towards benefitting society.

We welcome the opportunity to respond to the Office of Science and Technology Policy’s Request for Information on the update of the National AI Research and Development Strategic Plan. The past years have seen extraordinary advances in the field, further evidencing the ways AI will impact all parts of society. **We believe the eight Strategies in the Plan reflect a comprehensive vision for AI R&D, and as such remain the right priorities to guide investment. At the same time, clarifying how the United States intends to weave them together into a whole-of-government approach, particularly given the emergence of many new AI-related government efforts in the past two years, remains crucial to continued leadership in this space.** We offer below some considerations to strengthen these strategies, and we look forward to future opportunities to input into OSTP’s work.

About DeepMind

DeepMind\(^1\) is a scientific discovery company, committed to ‘solving intelligence’ to advance science and benefit humanity. This requires a diverse and interdisciplinary team working closely together – from scientists and designers, to engineers and ethicists. AI has the potential to enrich the lives of billions and improve our understanding of the universe. Ultimately we hope that new scientific breakthroughs, driven by innovations in machine learning, can make the crucial difference in helping us prosper in an increasingly complex world, and respond to global challenges such as climate change and tackling diseases.

---

\(^1\) We share these comments on behalf of DeepMind, and not on behalf of Google or any other entity in Alphabet, Inc.
For AI to benefit as many people as possible, it needs to be built and used responsibly. We view responsible AI as an ongoing process of ensuring our research and engineering are informed by the values, needs and expectations of society – with the goal to minimize risks, but also to accelerate and equitably distribute the benefits of AI. In practice, this means: (1) making sure our research addresses major scientific and social challenges; (2) anticipating and mitigating potential risks and harms; and (3) engaging with the wider world, and its complexities, challenges and possibilities.

The evolving AI research landscape

The 2019 update of the Strategy captured priorities in AI development that remain relevant today. We particularly welcomed its emphasis on safe, ethical and trustworthy AI; the recognition of the inherently interdisciplinary nature of AI R&D; and the adoption of a long-term perspective, building preparedness for outcomes that could take decades to manifest. Taking stock of the Strategic Plan is nonetheless pertinent, given developments in the field in recent years and the broader backdrop of tremendous public and private investment in the AI ecosystem.

The growing promise of using AI to advance scientific discovery

AI can deliver transformative benefits by accelerating scientific discovery, helping to address grand challenges like climate change and pandemic preparedness. These benefits are just beginning to be realized, as shown by the launch of a new National Science Foundation–funded institute to harness AI for accelerated discoveries in physics, astronomy and neuroscience last year.

We hope that DeepMind’s deep learning system, AlphaFold, which can predict the 3D structure of a protein based solely on its genetic sequence, will herald a new age of AI-powered scientific breakthroughs. More than 350,000 users from more than 190 countries have already visited the AlphaFold Protein Structure Database, which we established in partnership with the European Molecular Biology Laboratory’s (EMBL) European Bioinformatics Institute (EBI), and which now holds structures for almost one million proteins. We are seeing promising signs of scientists incorporating AlphaFold into their day-to-day work, and socially–beneficial research
emerging in areas like human biology, neglected diseases, and plastic-degrading enzymes.

We are also starting to see how AI can help to tackle climate change, by making existing infrastructure, such as data centers, more efficient, and by enabling better prediction models in areas like weather forecasting. AI's most transformative climate impacts will likely come from accelerating crucial longer-term breakthroughs in areas like fusion and battery design.

**AI community efforts to pioneer responsibly**

AI's potential transformative benefits to society are only possible if it is built and used responsibly, and if negative effects, which have to date disproportionately affected marginalized groups, are sufficiently mitigated. The AI community has been making progress in providing practitioners with resources and tools to identify, analyze and manage potential risks and benefits from their AI research and applications. For example:

- In 2020, the NeurIPS AI research conference introduced a new requirement for authors to produce an impact statement on the potential ethical aspects and societal risks of their work.
- Organizations deploying AI applications can increasingly draw on tools and best practice case studies shared by other organizations,\(^2\) in areas like fairness, privacy, security and explainability\(^3\).
- At DeepMind we invest heavily in safety, security, privacy, ethics and sociotechnical research both to inform our own approach to AI development and governance, and to help foster progress in the broader field. For example, we recently released research that assessed the near-term ethics risks, and longer-term safety risks, posed by large-scale language models – a priority area of focus for AI researchers.

---

\(^2\) Google has made available a range of off-the-shelf and customisable risk analysis and mitigation tools, in areas like fairness, privacy and explainability.

\(^3\) For instance, DeepMind’s recent work on glass dynamics demonstrates how a graph neural network can predict molecules’ future movements based on their current structure. The DeepMind Safety Research Team published on its blog about 'what mechanisms drive agents behavior’. Another example is this paper which seeks to provide explanations for AlphaZero’s chess moves.
Updating the National AI R&D Strategic Plan

Over the past three years, an increasing number of policy instruments have been established to support the AI ecosystem and demonstrate the US government’s commitment to being a world leader in AI. We particularly welcome:

- The increased investment to accelerate AI innovation – including by significantly increasing non-defense AI funding and augmenting R&D in security and robustness – following various recommendations (including from the National Security Commission on AI).
- The establishment of the National AI Office within OSTP and the creation of platforms such as ai.gov to improve the visibility of ongoing AI-related work and to act as a connection point for stakeholders.
- The creation of a National AI Research Resource Task Force to democratize access to research tools and capabilities;
- The US’s support and creation of international fora to improve convergence on AI governance questions, such as the EU-US Trade and Tech Council, and to facilitate the input of international experts, such as the Global Partnership for AI.
- The work initiated by the National Institute of Standards and Technology around trustworthy AI, including the development of a voluntary AI Risk Management Framework (RMF). We’ve welcomed the multistakeholder approach taken by NIST and the multiple opportunities to input as the RMF is being developed.

The National AI R&D Strategic Plan rightly highlights the importance of a whole-of-government approach; the recent multiplication of policy proposals around R&D investments shows coordination is all the more important today. In the next update to the Plan, we recommend prioritizing that a common vision be evident for how the government intends to drive forward its multiple strategies. We hope to see, for instance, how this Strategic Plan will link with OSTP’s development of an “AI bill of rights,” among other initiatives. Considering the complex nature of AI R&D and the overlaps that naturally exist between the eight strategies, the Strategic Plan could also benefit from a better acknowledgement of these links to prevent each strategy from becoming siloed, a challenge which figure 1 of the Strategic Plan tries to capture.
We share below research areas that could be emphasized and more general observations across the eight strategies.

1) R&D areas

*Long-term investments in AI research (Strategy 1)*

The Strategic Plan rightly highlights the long-term perspective needed when investing in AI, particularly in research aiming to achieve general-purpose artificial intelligence. We believe general-purpose learning systems are key to unlocking the long-term potential for AI to benefit society. But as such AI capabilities become more advanced, they raise the possibility of novel safety risks that may manifest on different timelines. While this is also covered in Strategy 4, we recommend that this section place greater emphasis on AI safety research, and connect clearly to the subsection in Strategy 4 that focuses on long-term AI safety and value alignment.

*Effective methods for human–AI collaboration (Strategy 2)*

We strongly agree with the view outlined in the Strategy that "achieving effective interactions between humans and AI systems requires additional R&D to ensure that the system design does not lead to excessive complexity, undertrust, or overtrust." Enabling this careful calibration of the trust one places in an AI system is core to the notion of “trustworthy AI.” While we are seeing promising research in this space, it remains underserved, especially with regard to more advanced AI systems. For example, while researchers in the field of human computer interaction (HCI) have studied the trust placed by humans in narrow applications, such as doctors using ML-based tools for diagnostics, there is comparatively little research on how humans interact with broader or more advanced systems, such as large language models.

The Strategy also seems to focus quite heavily on human–computer–collaboration on a task and on human augmentation, in comparison to studying human perception and meaningful oversight of AI systems in human–AI collaboration settings. A

---

4 In an example of how general learning systems are beginning to be applied for practical applications, DeepMind recently demonstrated how the application of our algorithm, MuZero, helped improve video compression, resulting in a 4% bitrate reduction across a large, diverse set of videos for YouTube.
relevant avenue of research in this direction would focus on how we can design meaningful human control, oversight, and accountability. The Strategy could also look at ways to create human-in-the-loop evaluation pipelines for AI, and how to provide humans with explanations for decisions and outcomes in a way that is judged by a human to be useful.

2) Cross-cutting R&D Foundations

Ethical, legal, and societal implications of AI (Strategy 3)

This Strategy rightly highlights the importance of understanding the ethical, legal, and social implications of AI, as well as developing methods for AI design that align with ethical, legal, and social principles. We need to continue encouraging investments in sociotechnical research, and institutionalizing best practices such as foresight and deliberation. The prompt that the NeurIPS conference introduced for researchers to produce an impact statement on the potential ethical aspects and societal risks of their work is something that could be further replicated in other conferences and processes to award grants.

We also agree that research benefits from multidisciplinary perspectives from computer science, the social and behavioral sciences, ethics, biomedical science, and other fields. Such an interdisciplinary approach will also be particularly important to driving forward much needed research on AI governance. At DeepMind, we have a multidisciplinary leadership group and dedicated internal teams that review research proposals and potential applications of our technology, consult external experts, and develop recommendations to maximize the likelihood and distribution of positive outcomes and minimize the potential for harm. In advance of our AlphaFold release, for instance, this group engaged with leading bioethicists and protein folding researchers to explore potential impacts on the research community, as well as any ways in which harmful actors might use our research. We also sought guidance from experts in areas with potential for beneficial impact, such as neglected diseases, to try and validate and accelerate these opportunities. We’ve also seen other organizations create fora for such discussions, in ways that match their particular research priorities.⁵

⁵ Stanford University for instance created the Ethics and Society Review (ESR) to aid researchers in mitigating negative ethical and societal aspects of their research.
While this is already mentioned throughout the Strategy, we recommend creating a separate subsection on the importance of multi-stakeholder approaches, and on prioritizing diversity and inclusion in discussions on the ethical, legal and societal implications of AI. AI actors (researchers, developers, deployers, and more) need to develop normative thresholds to inform decision-making about what constitutes ‘trustworthy enough’ for an AI system to be deployed, and inclusiveness should be central to such decision-making. Developing normative thresholds requires identifying and engaging with groups that may be most at risk from AI systems. Sociotechnical research can help all sorts of stakeholders — including both parties involved in the design and development of AI systems and end users — better understand how society (including traditionally minority groups) may be affected by AI systems. Participatory approaches with these groups can provide a source of expert insights and lived experience, and a way to empower those who may be most affected by AI systems. There could also be value in mirroring ways in which some governments are creating fora for public engagement: the UK, for instance, has created an AI Council to engage with the broader AI community.

**Safety and security of AI systems (Strategy 4)**

Research into the safety and security of AI systems remains a crucial priority, and we consider the current description of the Strategy still relevant, since challenges such as improving trust and increasing the explainability and transparency of AI need continued research and funding.

The section on ‘long term AI safety and AI alignment’ would benefit from adopting a broader framing: ultimately, the focus of alignment research is to prevent powerful goal-directed systems from pursuing undesired goals. While recursive self-improvement is an important potential factor, it is only one (speculative) way

---

6 By ‘sociotechnical research’, we mean research on the interaction and effects of AI when embedded in a specific social system. For example, DeepMind researchers have applied critical science and decolonial theory to AI to explore risks like algorithmic oppression, dispossession and exploitation, and analysed the potential positive and negative effects of artificial intelligence on queer communities.

7 DeepMind researchers are exploring the potential role of participatory approaches in developing and/or evaluating AI systems. Such approaches are nascent in AI, but are more established in fields like human–computer–interaction (HCI), which we hope can serve as an important source of insights.
that such risks could emerge. We hence recommend this section take a wider lens, including specific longer-term AI safety problems that have been identified as priorities by the research community. An example of such a problem is ‘specification gaming’, where an agent satisfies the literal specification of their objective, but does not achieve the desired outcome. We also recommend considering how to attract more researchers into long-term AI safety. The field is home to a growing number of talented researchers, but their overall number remains limited. High-profile, coordinated activities – such as collective research agendas, dedicated major funds for such research, and/or competitions – could help to further develop the field.

The section on ‘security against attacks’ covers a robust set of priorities on security of AI systems, and the focus should remain on ensuring these systems are developed to detect when they are being attacked, and to withstand attacks (such as data poisoning to evade or manipulate models, model stealing, etc.). Two areas that could be explored further are the verifiability of AI systems and ensuring some level of provenance and traceability of the corresponding models and data of such systems. We recommend discussion on bringing more transparency and openness in deployed AI systems to address emerging risks and enable trust-building.

Shared public datasets and environments for AI training and testing (Strategy 5)

As the 2019 Strategic Plan outlines, it is critical for researchers to have easy access to reliable, clean, as well as findable, accessible, interoperable, and reusable (FAIR) data. We welcome the creation of the National AI Research Resource Task Force and the drafting of a roadmap to expand access to critical resources and educational tools relating to AI. The Strategic Plan could be updated to more directly support the National AI Research Resource Task Force’s mandate.

We also encourage OSTP to consider updating the Strategic Plan to encourage opportunities to make data from countries around the world more available and

---

8 DeepMind has a dedicated team of researchers that identify and work on such problems, in particular specification, robustness and assurance. A recent paper assesses the potential risks from misaligned Language Agents, such as producing language that is deceptive or manipulative.

9 DeepMind collated 60 examples of specification gaming, such as an agent that was trained to carry a ball on its back but instead dropped the ball into a leg joint and wiggled across the floor without the ball dropping.
accessible and, where possible, to use harmonized methods to protect privacy and security.

**Standards and benchmarks (Strategy 6)**

The establishment of effective AI standards and benchmarks is key to trustworthy AI. We welcome the efforts NIST is leading in this space, along with many needed discussions in international fora. At the same time, it is helpful to acknowledge the need for more research and awareness of possible limitations of benchmarks, and the need for new accompanying tools to address this. For instance:

- Current benchmarks disproportionately evaluate accuracy on narrow tasks of interest, and neglect important areas like fairness and explainability, although useful benchmarks are emerging in these areas. Current benchmarks also don’t necessarily test performance in a way that is relevant to real-world use. As highlighted in a paper by Deb Raji, Emily Bender and co–authors at Google, evaluating AI systems based on their aggregate performance on benchmark tasks may provide little indication of their real-world utility. For example, 80% accuracy on Iris classification might be sufficient for the botany world, but to classify a mushroom you need to ingest as poisonous or edible, you would need 99% (or higher) accuracy.

- Benchmarks that are analyzed in isolation may also give a wrong picture of a model, as pointed out by DeepMind researchers in this paper. They find that evaluating a model against toxicity can ignore or even introduce unfair bias, for example against texts about, and dialects of, marginalized groups, demonstrating that benchmarks must consider social harms in concert in order to not fix one problem by aggravating another.

- Most benchmarks are quantitative metrics on task performance, and there is a need for new types of qualitative analysis tools, including feedback from humans–in–the–loop and learnings from other domains, such as auditing practices that evaluate how something is being used in the real world.

**National AI R&D workforce needs (Strategy 7)**

The Strategic Plan mentions the importance of broadening participation among groups traditionally underrepresented in computing and related fields. Diverse
teams are indeed important – not only for the more innovative work that such teams produce, but also because of the diverse values, hopes, and concerns that diverse teams bring into AI design, risk/benefit assessment, mitigation development, and deployment. We believe this is essential and an area where the private sector also has a role to play: the Strategic plan could highlight opportunities to work with the private sector to tackle these challenges and create a strong diverse talent pipeline.

The Strategy could also include mention of the importance of supporting early-career researchers from under-represented groups to pursue postdocs and possibly transition to permanent positions in academia to become role models for the next generation of researchers. This is why we launched the DeepMind Academic Fellowships, in addition to the DeepMind Scholarship, which provides funding and mentoring to graduate students.

Public–private partnerships to accelerate advances in AI (Strategy 8)

The addition of this Strategy in 2019 reflected the growing importance of public–private partnerships in enabling AI R&D. The 2022 update might be the opportunity to create a separate section dedicated to the need for more international projects, across governments but also with the private sector and civil society, to accelerate advances in AI. We’re particularly supportive of initiatives such as the recent US–UK Privacy Enhancing Technologies prize challenge, and encourage more international AI R&D collaborations.