DeepMind

Request for Information:
Artificial Intelligence Risk Management Framework

86 FR 40810
Docket # 210726-0151
September 2021

At DeepMind we believe that AI's extraordinary potential will only be realised if its development and deployment upholds appropriate ethical standards and is purposefully directed towards benefitting society.

We welcome the opportunity to respond to the National Institute of Standards and Technology's (NIST) request for information on an Artificial Intelligence (AI) Risk Management Framework (RMF) and look forward to future opportunities to input on the framework.

About DeepMind[1]

DeepMind 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 tough challenges such as climate change and tackling diseases.

With our deep learning system AlphaFold, for instance, we brought together experts from the fields of structural biology, physics, and machine learning to apply cutting-edge techniques to predict the 3D structure of a protein, based solely on its genetic sequence and showed how artificial intelligence research can drive and accelerate new scientific discoveries. **We are excited for resources such as the AlphaFold Protein Structure Database**, which we recently released in partnership with the European Molecular Biology Laboratory's (EMBL) European Bioinformatics Institute (EBI), to herald a new age of AI-powered scientific breakthroughs.

For AI to benefit as many people as possible**, it needs to be built and used responsibly.** For DeepMind, responsibility means aligning our research with society. We view responsible AI as an ongoing process of ensuring our research and engineering are informed by the values,

[1] We share these comments on behalf of DeepMind, and not on behalf of Google or any other entity in Alphabet, Inc.

needs and expectations of society – with the goal to minimise 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.

Responses to NIST Requests for Feedback

We support the attributes NIST has identified as being crucial to trustworthy AI. We endorse standards being formed through a multistakeholder process, and welcome NIST soliciting a broad variety of views and input from diverse parties. It is also valuable to recognise and build on existing frameworks and standards that are relevant to address these areas, including ones led by NIST on related topics (such as privacy and cybersecurity) and work by others on AI (including organisations like the OECD and standards organisations such as ISO). Over time, it is crucial to achieve global coherence and interoperability between AI governance frameworks.

We offer some considerations below that respond to specific questions in the NIST RFI, grounded in DeepMind's experience as a research-focused organisation. (Please note our response is not a holistic overview of how DeepMind approaches risk management.)

1. Key challenges in managing AI-related risk (NIST Q1, Q8)

A. Encouraging ethical risk assessment early, including at the research design stage

The NIST RFM should **encourage appropriately early and appropriately broad deliberation** on potential risks - i.e. not just in the run up to deploying AI-based technologies, products and services, but also in earlier research efforts. Such early deliberation is crucial, as **specific risks can emerge from how AI research efforts are designed** - for example, bias issues may emerge from how a research problem is framed or how datasets are annotated - while unique mitigation opportunities may also be available at the research stage.

To encourage such deliberation, we aim to foster a culture at DeepMind where we routinely **share and interrogate each other's research** and **think critically about** potential long-term impacts.[2] We have a **multidisciplinary leadership group**, and dedicated working groups, that review research proposals and applications of our technology, consult external

[2] For example, we provide bespoke AI and ethics training to every new DeepMind employee.

experts, and develop recommendations to maximise the likelihood of positive outcomes, and minimise the potential for harm – for example, by advising on dataset use.[3]

B. Designing technical benchmarks and normative thresholds for evaluating 'trustworthiness'

The NIST RMF should encourage the development of practicable **technical benchmarks** and methods to evaluate **trustworthiness characteristics,** such as safety, fairness and explainability, which are lacking today.

Just as software engineering has a set of benchmarks and methods to ensure security and reliability, **DeepMind's research helps to inform** novel benchmarks, datasets and methods for AI safety characteristics, such as specification, robustness and assurance, as well as other trustworthy AI characteristics, such as fairness.[4] We aim to **share our findings widely so they can be applied by others developing AI systems** - e.g. via our dedicated AI safety research blog, and via best practice sharing organisations like the Partnership for AI - a coalition of civil society, academic, and industry organisations that we helped to found.

No AI system will ever be completely free of risk of harm, and these risks also need to be considered against the potential benefits of AI systems. To do so, AI actors need to develop **normative thresholds** to inform decision-making about what constitutes 'trustworthy enough' for an AI system to be deployed. **Inclusiveness** should be central to such decision-making. As such, we believe in the importance of diverse teams - 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.[5]

Developing normative thresholds also requires **identifying and engaging with groups that may be most at risk** from AI systems. Sociotechnical research can help AI actors better understand how traditionally minoritised groups may be affected by AI systems, including both end-users and stakeholders involved in the design and development of AI systems.[6]

---

[3] For example, DeepMind's Raia Hadsell was co-chair of the 2020 NeurIPS conference, which introduced a new requirement for authors to produce an impact statement on the potential ethical aspects and societal risks of their work. Our ethicists also helped review NeurIPs papers that were flagged for ethical review.
[4] DeepMind has also published on methods to evaluate explainability, verification and uncertainty evaluation
[5] DeepMind approaches to improving opportunities for AI researchers from underrepresented groups include our diversity scholarship programme and pilot fellowship programmes.
[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.

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.[7]

C. Using foresight to identify, assess and mitigate longer-term risks

Beyond assessing risks from current AI systems, it is important to **incentivise deliberation over complex longer-term risks** from advancing AI capabilities, particularly given the inherent uncertainty that such risks entail, which can make them difficult to interpret and analyse in the near-term. For example, with rapid progress in the capabilities of natural language systems in mind, DeepMind researchers recently assessed potential **safety and alignment risks from future Language Agents**, such as 'gaming' misspecified objectives and producing deceptive and manipulative language.

D. Maintaining flexibility and adaptability to enable innovation

Given the rapid pace of AI development, an overarching challenge for any risk framework will be to strike a balance: **offering enough meaningful, concrete guidance while remaining flexible and adaptable over time**. We believe the advancing state of AI will help us make progress on challenges that are currently very difficult – including many of the challenges outlined in NIST's framework. On the other hand, as mentioned above, we may need to consider shifting risks in the longer term. The framework should address risks while continuing to encourage excellence and innovation in the AI ecosystem.

2. A workable approach to identifying, defining and managing 'AI trustworthiness' (NIST Q2, Q3)

As mentioned above, we are generally supportive of the list of trustworthy AI principles and characteristics put forward by NIST. To make the RMF as actionable as possible, NIST could consider **removing the distinction between 'characteristics'** (e.g. explainability, interpretability, reliability, privacy, robustness, safety), **and 'principles'** (e.g. fairness, transparency, and accountability), as this is not a consistent distinction in the AI community and could cause unnecessary confusion. The following considerations could also make the framework more actionable.

A. Build on definitions with tangible examples and case studies

Many characteristics of AI trustworthiness - such as 'fairness' and 'explainability' - **do not have a specific definition that enjoys widespread agreement**, and personal or institutional preferences sometimes play a role. For example, some AI practitioners prefer to use more

[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.

neutral terms like 'fairness', while others prefer terms like 'discrimination' to bring focus on potential harmful outcomes - even if discussing the same underlying risk.[8]

Definitions can also **fail to capture what effective risk mitigation** looks like, as this often varies significantly, depending on the context in question. For example, the main goal of 'interpretability/explainability' mitigations may be to provide more *mechanistic* understanding to researchers working to improve or test an AI system, or the goal may be to provide explanations that are perceived as *useful* by users with diverse needs. NIST's RMF should not only define trustworthy AI characteristics, but also clarify **how different characteristics relate to each other and provide tangible case studies** for what effective mitigation looks like, across different contexts.

B. Acknowledge trade-offs between different AI trustworthiness characteristics

Some of the trustworthiness characteristics outlined by NIST will inevitably come into conflict, while efforts to mitigate one risk can exacerbate another. Organisations need to be **aware of these trade-offs and deliberate over them appropriately in their decision-making**. For example, retaining audit logs of AI systems may have some merits, for example with respect to improving explainability, but they can also create privacy risks, for example if they contain personal information.[9,10]

C. Incentivise technical and sociotechnical research

Despite their importance, technical and sociotechnical research in trustworthy characteristics such as robustness, reliability, safety, bias, explainability and privacy remain significantly under-resourced. This is a crucial gap, as **the characteristics of AI trustworthiness in the NIST RFI are also open research problems.** For example, there are promising research approaches to better explain the predictions or behaviour of complex AI models via conversational AI agents, as well as approaches to evaluate how 'useful' these explanations are to humans in different contexts. We hope that NIST's RMF will incentivise investment and support for such technical and sociotechnical research.

---

[8] As the NIST RFI notes, there is no objective standard for the ethical values that AI systems should adhere to. DeepMind has carried out research to explore the question of AI value alignment, including how to align AI systems with the plurality of values endorsed by people across the world.

[9] Similarly, upcoming DeepMind research found that efforts to mitigate 'toxicity' in the outputs of language models can have negative consequences for texts about, and dialects of, marginalised groups. See: "Challenges in Detoxifying Language Models" to appear in Findings of EMNLP 2021.

[10] Conversely, while some characteristics of AI trustworthiness can conflict, there are also mitigation methods that can be used across multiple trustworthy characteristics. For example, DeepMind research used differential 'privacy' techniques to improve the adversarial 'robustness' of models.

DeepMind

3 Existing frameworks, methodologies, and tools for AI risk management (NIST Q5, Q7)

A. Build on existing international approaches to AI risk management and trustworthiness

In our recent response to the EU Artificial Intelligence Act proposal, we highlighted the need for **harmonised global governance for AI** and we believe NIST's work could be a central part of this effort. For example, we are very encouraged by the creation of global fora such as the US-EU Trade and Tech Council and hope it will lead to more interoperability between the EU and the US on AI safety and risk assessment standards.

NIST's Risk Management Framework (RMF) can support harmonised global governance by **clearly building on leading international frameworks for AI risk management and trustworthy AI** and clarifying any differences in the approach taken. Of particular note are the Organisation for Economic Co-operation (OECD)'s AI principles and their broader efforts to provide shared definitions and alignment on key AI concepts and challenges. The ongoing work of the International Organisation for Standardization (ISO)'s SC42 Committee, including on AI Risk Management is also highly-relevant. On a national level, the UK's Centre for Data Ethics and Innovation (CDEI)'s is developing an AI Assurance Roadmap which will explore and provide helpful guidance on related concepts and challenges.[11]

B. Some considerations from DeepMind practices, processes and frameworks

We aim to incorporate the considerations outlined above into our **own approach to ethical risk assessment.** We have a set of underlying ethical principles - such as a commitment to scientific excellence and designing systems that are accountable to people - that are overseen by a multidisciplinary leadership group and which are in line with Google's AI Principles. We also have a dedicated team that works with DeepMind's researchers, engineers, and external stakeholders, to analyse the potential ethical risks of our research, including downstream risks, and to identify and develop mitigations. They also analyse the potential benefits of our research and ways to accelerate and equitably distribute these benefits.

Across the organisation, we use various frameworks and tools to support this work. At a high level, we typically break down analysis of a proposed AI research effort into a number of overarching steps which include, among others, considering socially beneficial uses; direct and indirect ethical risks; and potential mitigations.[12] At each step, we draw on frameworks

[11] When it comes to defining AI trustworthy characteristics, there is also scope to draw on well-established human rights laws and documents, for example with respect to anti-discrimination.
[12] For further detail, please see this DeepMind/UCL lecture on Responsible Innovation.

and tools to **make the analysis more tractable and comprehensive,** including by helping with prioritisation of risks/mitigations**.** For example, to aid deliberation on potential safety risks from AI systems, DeepMind researchers designed an **AI Safety Framework** that groups such risks into three categories, which has since been used by external actors.[13][14]

[13] The three categories are: *Specification*: ensuring that an AI system's behaviour aligns with the operator's true intentions. *Robustness*: ensuring that an AI system continues to operate within safe limits upon encountering risk, unpredictability and volatility in real-world settings. *Assurance*: ensuring that we can understand and control AI systems during operation, enabling them to be continuously monitored and adjusted where required. The Centre for Security and Emerging Technologies (CSET) recently drew on the framework to outline 'hypothetical but realistic' scenarios in which AI systems can accidentally cause harm.

[14] DeepMind researchers have also considered ways to identify and categorise risks and mitigations for other trustworthiness characteristics, such as fairness, bias, and discrimination.