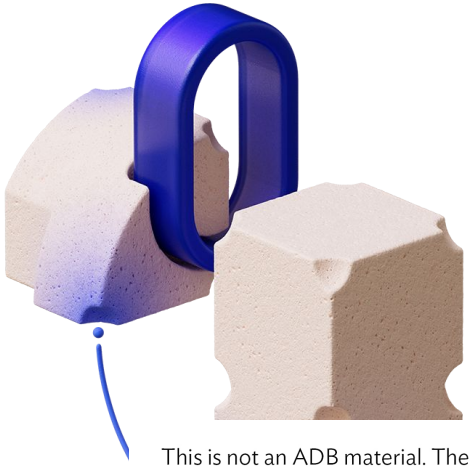
Google DeepMind



Navigating the Ethical Use and Governance of AI

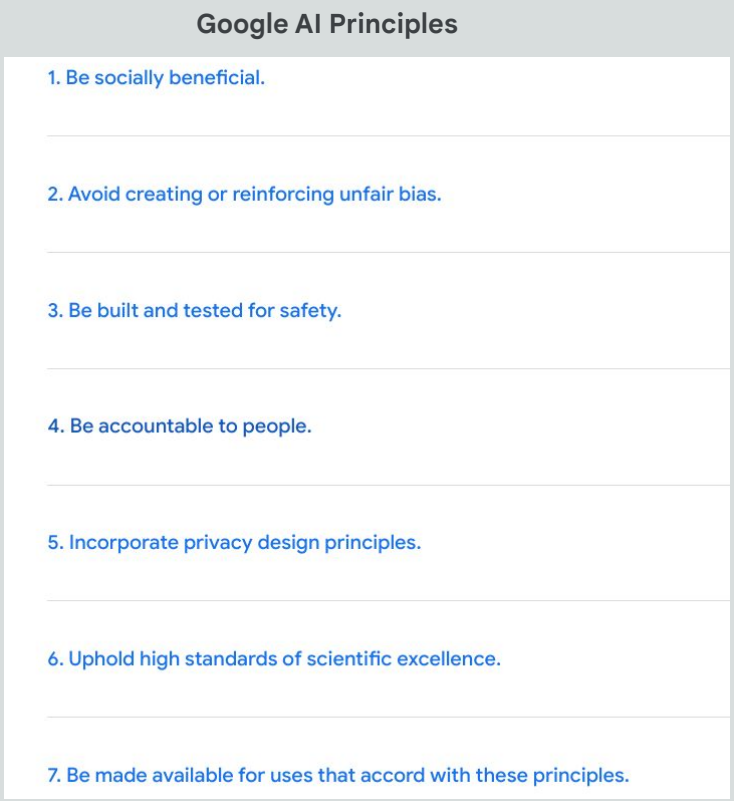
Alexandra Belias
Head of Product Policy & Partnerships
Google DeepMind

Our approach to pioneering responsibly

How we think about responsibility:

- **Governance**
- **Research**
- **Impact**

Responsible governance



- Guided by our **AI Principles,** we work to anticipate and evaluate our systems against a broad spectrum of AI-related risks, taking a holistic approach to responsibility and safety.
- To empower teams to pioneer responsibly and safeguard against harm, the **Responsibility and Safety Council (RSC),** evaluates Google DeepMind's research, projects and collaborations against our AI Principles, advising and partnering with research and product teams on our highest impact work.
- Our **AGI Safety Council,** led by our Co-Founder and Chief AGI Scientist Shane Legg, works closely with the RSC, to safeguard our processes, systems and research against extreme risks that could arise from powerful AGI systems in the future.
- We've also signed **public commitments** to ensure safe, secure and trustworthy AI, statements urging mitigation of AI risks to society, and pledges against using our technologies for lethal autonomous weapons.

Responsible research

- Effective **foresight** serves to anticipate and evaluate the risks that emerging technology pose, while supporting the application of mitigation approaches to manage those risks effectively.
- Some of our recent **research and papers**:
  - Ethical and social risks of harm from LMs.
  - Ethical implications of advanced AI assistants.
  - Frontier Safety Framework.
- Building pioneering research and **technical solutions such as SynthID**, our tool for watermarking and identifying our AI-generated content.

DeepMind

Ethical and social risks of harm from Language Models

Laura Weidinger[1], John Mellor[1], Maribeth Rauh[1], Conor Griffin[1], Jonathan Uesato[1], Po-Sen Huang[1], Myra Cheng[1,2], Mia Glaese[1], Borja Balle[1], Atoosa Kasirzadeh[1,3], Zac Kenton[1], Sasha Brown[1], Will Hawkins[1], Tom Stepleton[1], Courtney Biles[1], Abeba Birhane[1,4], Julia Haas[1], Laura Rimell[1], Lisa Anne Hendricks[1], William Isaac[1], Sean Legassick[1], Geoffrey Irving[1] and Iason Gabriel[1]

[1]DeepMind, [2]California Institute of Technology, [3]University of Toronto, [4]University College Dublin

Abstract

This paper aims to help structure the risk landscape associated with large-scale Language Models (LMs). In order to foster advances in responsible innovation, an in-depth understanding of the potential risks posed by these models is needed. A wide range of established and anticipated risks are analysed in detail, drawing on multidisciplinary literature from computer science, linguistics, and social sciences.

The paper outlines six specific risk areas: I. Discrimination, Exclusion and Toxicity, II. Information Hazards, III. Misinformation Harms, IV. Malicious Uses, V. Human-Computer Interaction Harms, VI. Automation, Access, and Environmental Harms.

The first risk area discusses fairness and toxicity risks in large-scale language models. This includes four distinct risks: LMs can create unfair discrimination and representational and material harm by perpetuating stereotypes and social biases, i.e. harmful associations of specific traits with social identities. Social norms and categories can exclude or marginalise those who exist outside them. Where a LM perpetuates such norms - e.g. that people called “Max” are “male”, or that “families” always consist of a father, mother and child - such narrow category use can deny or burden identities who differ. Toxic language can incite hate or violence or cause offense. Finally, a LM that performs more poorly for some social groups than others can create harm for disadvantaged groups, for example where such models underpin technologies that affect these groups. These risks stem in large part from choosing training corpora that include harmful language and overrepresent some social identities.

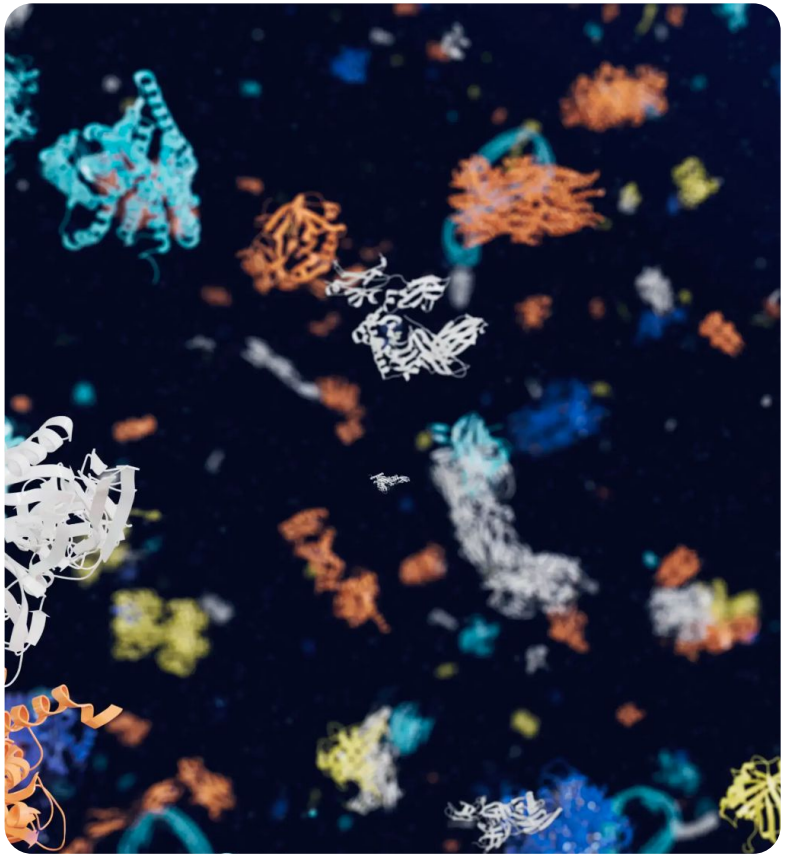
The second risk area includes risks from private data leaks or from LMs correctly inferring private or other

Responsible Impact

Responsibility and safety issues go well beyond any one organization.

We work with many brilliant non-profits, academics, and other companies to apply AI to solve problems that underpin global challenges, while proactively mitigating risks.

AlphaFold folded all 200M proteins known to science



All predictions are freely available in the **AlphaFold Protein Structure Database**

1.8M+ users

190+ countries

20,000+ citations

~1B years of research years

AlphaFold is accelerating progress on a range of important problems

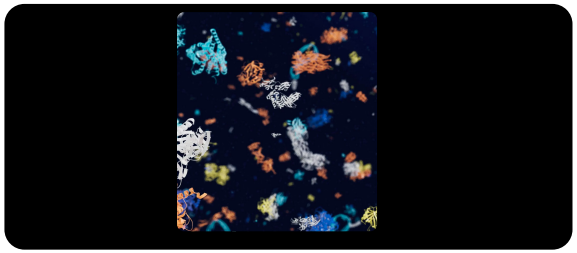
See more case studies at **unfolded.deepmind.com**

Plastic pollution

Designing plastic eating enzymes
McGeehan et al. (Centre for Enzyme Innovation)

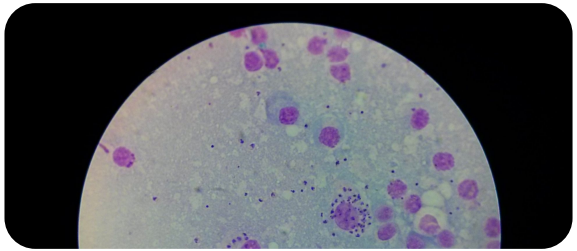
Antibiotic resistance

Blocking antibiotic resistance mechanisms
Sousa & Mitchell (Colorado)



Structural biology

Determined structure of nuclear pore complex
Fontana et al., Mosalaganti, et al. (Science)



Drug discovery & neglected diseases

Accelerating drug discovery
Perry (DNDi) & Kapeller (ROME Therapeutics)

Malaria vaccine

Designing a more effective malaria vaccine
Higgins (Oxford)

Drug delivery

Molecular syringe for protein delivery
Zhang (Broad Institute)

Thank you.

Google DeepMind