hemes

We want to promote research that ensures AI works for all. Our research themes are designed to reflect the key ethical challenges that exist for us and the wider AI community. We undertake research and collaborations in each of these areas, determined by the urgent challenges ahead.

Privacy, transparency, and fairness

AI systems can use large-scale and sometimes sensitive datasets, such as medical or criminal justice records. This raises important questions about protecting people's privacy and ensuring that they understand how their data is used. Also, the data used for training automated decision-making systems can contain biases, creating systems that might discriminate against certain groups of people.

How do concepts such as consent and ownership relate to using data in AI systems?

What can AI researchers do to detect and minimise the effects of bias?

What policies and tools allow meaningful audits of AI systems and their data?

AI morality and values

AI systems could make societies fairer and more equal. But different groups of people hold different values, meaning it is difficult to agree on universal principles. Likewise, endorsing values held by a majority could lead to discrimination against minorities.

How can we ensure that the values designed for AI systems reflect society?

How do we prevent AI systems from causing discrimination?

How do we integrate inclusive values into AI systems?

Governance and accountability

The creation and use of powerful new technologies requires effective governance and regulation, ensuring they are used safely and with accountability. In the case of AI, new standards or institutions may be needed to oversee its use by individuals, states, and the private sector - both internationally and within national borders.

What kinds of governance makes sense for rapidly developing technologies like AI?

Can existing institutions uphold the rights of everyone affected by AI?

What can we learn from other fields like biotechnology or genetics that might influence how AI is used?

AI and the world's complex challenges

By uncovering patterns in complex datasets and suggesting promising new ideas and strategies, AI technologies may one day help solve some of humanity's most urgent problems. But applying AI technologies to real-world problems takes careful consideration.

Which problems could AI help address?

How can AI research best contribute?

Who should we be working with to help solve problems?

Misuse and unintended consequences

While AI systems have great potential, they also come with risks. For example, they might malfunction or not operate in the ways they were intended. We might also rely on them too heavily in situations that go beyond their abilities or a technology designed to help society might be repurposed in unethical or harmful ways.

How can these risks be monitored across the world?

What structures can be put in place to minimise harm?

How do we ensure that people maintain control of AI systems?

Economic impact: inclusion and equality

Like previous waves of technology, AI could contribute to a huge increase in productivity. However, it could also lead to the widespread displacement of jobs and alter economies in ways that disproportionately affect some sections of the population. This poses important questions about the kinds of societies and economies we want to build.

How can we anticipate the social or economic impacts of AI?

What new opportunities are created?

How do we ensure AI has a net positive effect on the world?

Safety and Ethics

AI can provide extraordinary benefits, but like all technology, it can have negative impacts unless it's built and used responsibly. How can AI benefit society without reinforcing bias or unfairness? How can we build computer systems that invent new ideas, but also reliably behave in ways we want?

Our approach

Our teams working on technical safety, ethics, and public engagement aim to address these questions and more. We help anticipate short and long-term risks, explore ways to prevent these risks from happening, and find ways to address them if they do.

We believe this approach also means ruling out the use of AI technology in certain fields. For example, we've signed public pledges against using our technologies for lethal autonomous weapons, alongside many others from the AI community.

These issues go well beyond any one organisation. Our ethics team works with many brilliant non-profits, academics, and other companies, and creates forums for the public to explore some of the toughest issues. Our safety team also collaborates with other leading research labs, including our colleagues at Google, OpenAI, the Alan Turing Institute, and elsewhere.

It's also important that the people building AI reflect the broader society. We're working with universities on scholarships for people from underrepresented backgrounds, and support community efforts such as Women in Machine Learning and the African Deep Learning Indaba.

We support open research and investigation into the wider impacts of AI.

We created DeepMind Ethics & Society to guide the responsible development and deployment of AI. Our team of ethicists and policy researchers work closely with our AI research team to understand how technical advances will impact society, and find ways to reduce risk.

We also partner with outside experts and the general public to find answers together. We've supported partners including the Royal Society and the RSA to carry out public discussions and citizens' juries around AI ethics, and have given unrestricted financial grants to several universities working on these issues. We also helped co-found the Partnership on AI to bring together academics, charities, and company labs to solve common challenges.