Ada Lovelace Institute

Ethics and accountability in practice

Looking before we leap

Expanding ethical review processes for AI and data science research

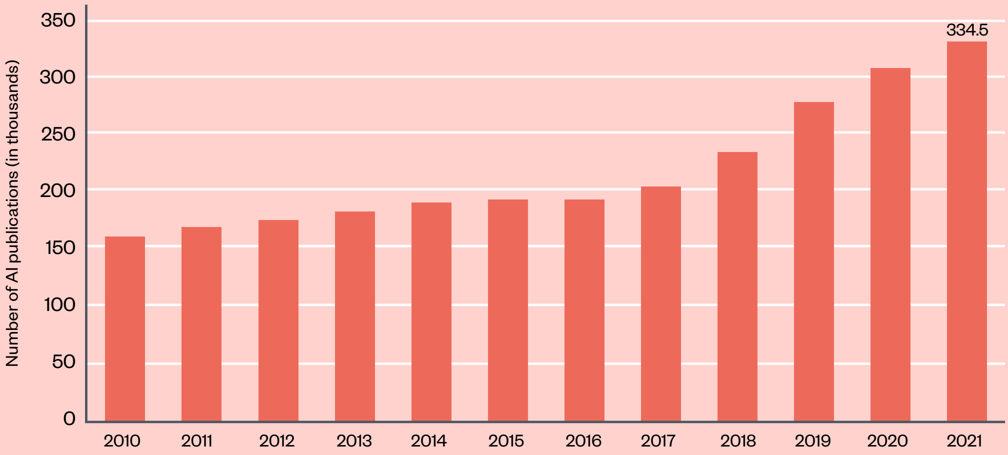
December 2022

Contents

Executive summary

Research in the fields of artificial intelligence (AI) and data science is often quickly turned into products and services that affect the lives of people around the world. Research in these fields is used in the provision of public services like social care, determining which information is amplified on social media, what jobs or insurance people are offered, and even who is deemed a risk to the public by police and security services. There has been a significant increase in the volume of AI and data science research in the last ten years, with these methods now being applied to other scientific domains like history, economics, health sciences and physics.

Figure 1: Number of AI publications in the world 2010-21[1]



Globally, the volume of AI research is increasing year-on-year and currently accounts for more than 4% of all published research.

[1] Source: Zhang, D. et al. (2022). 'The AI Index 2022 Annual Report'. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2205.03468

Since products and services built with AI and data science research can have substantial effects on people's lives, it is essential that this research is conducted safely and responsibly, and with due consideration for the broader societal impacts it may have. However, the traditional research governance mechanisms that are responsible for identifying and mitigating ethical and societal risks often do not address the challenges presented by AI and data science research.

As several prominent researchers have highlighted,[2] inadequately reviewed AI and data science research can create risks that are carried downstream into subsequent products,[3] services and research.[4] Studies have shown these risks can disproportionately impact people from marginalised and minoritised communities, exacerbating racial and societal inequalities.[5] If left unaddressed, unexamined assumptions and unintended consequences (paid forward into deployment as 'ethical debt'[6]) can lead to significant harms to individuals and society. These harms can be challenging to address or mitigate after the fact.

Ethical debt also poses a risk to the longevity of the field of AI: if researchers fail to demonstrate due consideration for the broader societal implications of their work, it may reduce public trust in the field. This could lead to it becoming a domain that future researchers find undesirable to work in – a challenge that has plagued research into nuclear power and the health effects of tobacco.[7]

To address these problems, there have been increasing calls from within the AI and data science research communities for more mechanisms, processes and incentives for researchers to consider the broader

[2] Bender, E.M. (2019). 'Is there research that shouldn't be done? Is there research that shouldn't be encouraged?'. *Medium*. Available at: https://medium.com/@emilymenonbender/is-there-research-that-shouldnt-be-done-is-there-research-that-shouldn-t-be-encouraged-b1bf7d321bb6

[3] Truong, K. (2020). 'This Image of a White Barack Obama Is AI's Racial Bias Problem In a Nutshell'. *Vice*. Available at: https://www.vice.com/en/article/7kpxyy/this-image-of-a-white-barack-obama-is-ais-racial-bias-problem-in-a-nutshell

[4] Small, Z. '600,000 Images Removed from AI Database After Art Project Exposes Racist Bias'. *Hyperallergic*. Available at: https://hyperallergic.com/518822/600000-images-removed-from-ai-database-after-art-project-exposes-racist-bias/

[5] Richardson, R. (2021). 'Racial Segregation and the Data-Driven Society: How Our Failure to Reckon with Root Causes Perpetuates Separate and Unequal Realities'. *Berkeley Technology Law Journal*, 36(3). Available at: https://papers.ssrn.com/abstract=3850317; Buolamwini, J. and Gebru, T. (2018). 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification'. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Conference on Fairness, Accountability and Transparency, PMLR*, pp. 77–91. Available at: https://proceedings.mlr.press/v81/buolamwini18a.html

[6] Petrozzino, C. (2021). 'Who pays for ethical debt in AI?'. *AI and Ethics*, 1(3), pp. 205–208. Available at: https://doi.org/10.1007/s43681-020-00030-3

[7] Abdalla, M. and Abdalla, M. (2021). 'The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity'. *AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. Available at: https://doi.org/10.1145/3461702.3462563

societal impacts of their research.[8]

The current role, scope and function of most academic and corporate RECs are insufficient for the myriad of ethical challenges that AI and data science research can pose

In many corporate and academic research institutions, one of the primary mechanisms for assessing and mitigating ethical risks is the use of Research Ethics Committees (RECs), also known in some regions as Institutional Review Boards (IRBs) or Ethics Review Committees (ERCs). Since the 1960s, these committees have been empowered to review research before it is undertaken and can reject proposals unless changes are made in the proposed research design.

RECs generally consist of members of a specific academic department or corporate institution, who are tasked with evaluating research proposals before the research begins. Their evaluations are based on a combination of normative and legal principles that have developed over time, originally in relation to biomedical human subjects research. A REC's role is to help ensure that researchers justify their decisions for how research is conducted, thereby mitigating the potential harms they may pose.

However, the current role, scope and function of most academic and corporate RECs are insufficient for the myriad of ethical challenges that AI and data science research can pose. For example, the scope of REC reviews is traditionally only on research involving human subjects. This means that the many AI and data science projects that are not considered a form of direct intervention in the body or life of an individual human subject are exempt from many research ethics review processes.[9] In addition, a significant amount of AI and data science research involves the use of publicly available and repurposed datasets, which are considered exempt from ethics review under many current research ethics guidelines.[10]

[8] For example, a recent paper from researchers at Microsoft includes guidance for a structured exercise to identify potential limitations in AI research. See: Smith, J. J. et al. (2022). 'REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research'. *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 587–597. Available at: https://doi.org/10.1145/3531146.3533122

[9] Metcalf, J. and Crawford, K. (2016). 'Where are human subjects in big data research? The emerging ethics divide.' *Big Data & Society*, 3(1). Available at: https://doi.org/10.1177/2053951716650211

[10] Metcalf, J. and Crawford, K. (2016).

> If AI and data science research is to be done safely and responsibly, RECs must be equipped to examine the full spectrum of risks, harms and impacts that can arise in these fields.

In this report, we explore the role that academic and corporate RECs play in evaluating AI and data science research for ethical issues, and also investigate the kinds of common challenges these bodies face.

The report draws on two main sources of evidence: a review of existing literature on RECs and research ethics challenges, and a series of workshops and interviews with members of RECs and researchers who work on AI and data science ethics.

Challenges faced by RECs

Our evaluation of this evidence uncovered **six challenges** that RECs face when addressing AI and data science research:

Challenge 1: Many RECs lack the resources, expertise and training to appropriately address the risks that AI and data science pose.

Many RECs in academic and corporate environments struggle with inadequate resources and training on the variety of issues that AI and data science can raise. The work of RECs is often voluntary and unpaid, meaning that members of RECs may not have the requisite time or training to appropriately review an application in its entirety. Studies suggest that RECs are often viewed by researchers as compliance bodies rather than mechanisms for improving the safety and impact of their research.

Challenge 2: Traditional research ethics principles are not well suited for AI research.

RECs review research using a set of normative and legal principles that are rooted in biomedical, human-subject research practices, which operate under a researcher-subject relationship rather than a researcher-data subject relationship. This distinction has challenged traditional principles of consent, privacy and autonomy in AI research, and created confusion and challenges for RECs trying to apply these principles to novel forms of research.

Challenge 3: Specific principles for AI and data science research are still emerging and are not consistently adopted by RECs.

The last few years have seen an emerging series of AI ethics principles aimed at the development and deployment of AI systems. However, these principles have not been well adapted for AI and data science research practices, signalling a need for institutions to translate these principles into actionable questions and processes for ethics reviews.

Challenge 4: Multi-site or public-private partnerships can exacerbate existing challenges of governance and consistency of decision-making.

An increasing amount of AI research involves multi-site studies and public-private partnerships. This can lead to multiple REC reviews of the same research, which can highlight different standards in ethical review of different institutions and present a barrier to completing timely research.

Challenge 5: RECs struggle to review potential harms and impacts that arise throughout AI and data science research.

REC reviews of AI and data science research are *ex ante* assessments, done before research takes place. However, many of the harms and risks in AI research may only become evident at later stages of the research. Furthermore, many of the types of harms that can arise – such as issues of bias, or wider misuses of AI or data – are challenging for a single committee to predict. This is particularly true with the broader societal impacts of AI research, which require a kind of evaluation and review that RECs currently do not undertake.

Challenge 6: Corporate RECs lack transparency in relation to their processes.

Motivated by a concern to protect their intellectual property and trade secrets, many private-sector RECs for AI research do not make their processes or decisions publicly accessible and use strict non-disclosure agreements to control the involvement of external experts in their decision-making. In some extreme cases, this lack of transparency has raised suspicion of corporate REC processes from external research partners, which can pose a risk to the efficacy of public-private research partnerships.

Recommendations

To address these challenges, we make the following **recommendations**:

For academic and corporate RECs

Recommendation 1: Incorporate broader societal impact statements from researchers.

A key issue this report identifies is the need for RECs to incentivise researchers to engage more reflexively with the broader societal impacts of their research, such as the potential environmental impacts of their research, or how their research could be used to exacerbate racial or societal inequalities.

There have been growing calls within the AI and data science research communities for researchers to incorporate these considerations in various stages of their research. Some researchers have called for changes to the peer review process to require statements of potential broader societal impacts,[11] and some AI/machine learning (ML) conferences have experimented with similar requirements in their conference submission process.[12]

RECs can support these efforts by incentivising researchers to engage in reflexive exercises to consider and document the broader societal impacts of their research. Other actors in the research ecosystem (funders, conference organisers, etc.) can also incentivise researchers to engage in these kinds of reflexive exercises.

[11] Hecht, B. et al. (2021). 'It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process'. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2112.09544

[12] Ashurst, C. et al. (2021). 'AI Ethics Statements – Analysis and lessons learnt from NeurIPS Broader Impact Statements'. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2111.01705

Recommendation 2: RECs should adopt multi-stage ethics review processes of high-risk AI and data science research.

Many of the challenges that AI and data science raise will arise in different stages of research. RECs should experiment with requiring multiple stages of evaluations of research that raises particular ethical concern, such as evaluations at the point of data collection and a separate evaluation at the point of publication.

Recommendation 3: Include interdisciplinary and experiential expertise in REC membership.

Many of the risks that AI and data science research pose cannot be understood without engagement with different forms of experiential and subject-matter expertise. RECs must be interdisciplinary bodies if they are to address the myriad of issues that AI and data science can pose in different domains, and should incorporate the perspectives of individuals who will ultimately be impacted by the research.

For academic/corporate research institutions

Recommendation 4: Create internal training and knowledge-sharing hubs for researchers and REC members, and enable more cross-institutional knowledge sharing.

These hubs can provide opportunities for cross-institutional knowledge-sharing and ensure institutions do not develop standards of practice in silos. They should collect and share information on the kinds of ethical issues and challenges AI and data science research might raise, including case studies of research that raises challenging ethical issues. In addition to our report, we have developed a resource consisting of six case studies that we believe highlight some of the common ethical challenges that RECs might face.[13]

[13] See: Ada Lovelace Institute. (2022). *Looking before we leap: Case studies*. Available at: https://www.adalovelaceinstitute.org/resource/research-ethics-case-studies/

Recommendation 5: Corporate labs must be more transparent about their decision-making and do more to engage with external partners.

Corporate labs face specific challenges when it comes to AI and data science reviews. While many are better resourced and have experimented with broader societal impact thinking, some of these labs have faced criticism for being opaque about their decision-making processes. Many of these labs make consequential decisions about their research without engaging with local, technical or experiential expertise that resides outside their organisation.

For funders, conference organisers and other actors in the research ecosystem

Recommendation 6: Develop standardised principles and guidance for AI and data science research principles.

RECs currently lack standardised principles for evaluating AI and data science research. National research governance bodies like UKRI should work to create a new set of 'Belmont 2.0' principles[14] that offer some standardised approaches, guidance and methods for evaluating AI and data science research. Developing these principles should draw on a wide set of perspectives from different disciplines and communities who are impacted by AI and data science research, including multinational perspectives – particularly from regions that have been historically underrepresented in the development of past research ethics principles.

Recommendation 7: Incentivise a responsible research culture.

AI and data science researchers lack incentives to reflect on and document the societal impacts their research. Different actors in the research ecosystem can encourage ethical behaviour – funders, for example, can create requirements that researchers conduct a broader societal impact statement of their research in order to receive a grant, and conference organisers and journal editors can encourage researchers to include a broader societal impact statement when

[14] Raymond, N. (2019). 'Safeguards for human studies can't cope with big data'. *Nature*, 568(7752), pp. 277–277. Available at: https://doi.org/10.1038/d41586-019-01164-z

submitting research. By creating incentives throughout the research ecosystem, ethical reflection can become more desirable and rewarded.

Recommendation 8: Increase funding and resources for ethical reviews of AI and data science research.

There is an urgent need for institutions and funders to support RECs, including paying for the time of staff and funding external experts to engage in questions of research ethics.

How to read this report

If you are a **member of an academic or corporate Research Ethics Committee (REC)**, we recommend jumping to the challenges section on page 43, to get a sense of some of the common challenges RECs face when grappling with AI and data science research. We also recommend reading pages 72–85, which include a series of recommendations to help structure your REC.

If you are an **administrator or senior leader at a corporate or academic research institution**, we recommend paying particular attention to pages 86–92, where we discuss recommendations for how a wider institution can address these challenges.

If you are a **funder**, **conference organiser**, **journal editor**, or otherwise identify as a member of the AI or data science research communities, we recommend jumping to pages 92–97, where we discuss some of the common challenges and recommendations.

If you are **someone less familiar with how AI or data science research is conducted** but want to understand the history and context of Research Ethics Committees and how they operate, jump to page 17, where we provide some context for how RECs function.

This report is accompanied by a resource consisting of **six case studies,**[15] which provide fictional but representative AI and data research proposals, designed to prompt reflection on the common ethical issues and societal implications raised by different AI and data science research projects. The case studies should not be seen as exclusive to either academic settings or industry and are for use by students, researchers, members of Research Ethics Committees, funders and other actors in the research ecosystem, to support learning about common ethical issues in AI and data science research.

[15] See: Ada Lovelace Institute. (2022). *Looking before we leap: Case studies*. Available at: https://www.adalovelaceinstitute.org/resource/research-ethics-case-studies/

Introduction

The academic fields of AI and data science research have witnessed an explosive growth in the last two decades. According to the Stanford AI Index, between 2015 and 2020, the number of AI publications on open-access publication database arXiv grew from 5,487 to over 34,376 (see also Figure 1, page 3). As of 2019, AI publications represented 3.8% of all peer-reviewed scientific publications, an increase from 1.3% in 2011.[16] The vast majority of research appearing in major AI conferences comes from academic and industry institutions based in the European Union, China and the United States of America.[17] AI and data science techniques are also being applied across a range of other academic disciplines such as history,[18] economics,[19] genomics[20] and biology.[21]

Compared to many other disciplines, AI and data science have a relatively fast research-to-product pipeline and relatively low barriers for use, making these techniques easily adaptable (though not necessarily well suited) to a range of different applications.[22] While these qualities have led AI and data science to be described as 'more important than fire and electricity' by some industry leaders,[23] there have been increased calls from members of the AI research community to require researchers to consider and address 'failures of imagination'[24] of the potential broader societal impacts and risks of their research.

[16] The number of AI journal publications grew by 34.5% from 2019 to 2020, compared to a growth of 19.6% between 2018 and 2019. See: Stanford University. (2021). *Artificial Intelligence Index 2021*, chapter 1. Available at: https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report-_Chapter-1.pdf

[17] Chuvpilo, G. (2020). 'AI Research Rankings 2019: Insights from NeurIPS and ICML, Leading AI Conferences'. *Medium*. Available at: https://medium.com/@chuvpilo/ai-research-rankings-2019-insights-from-neurips-and-icml-leading-ai-conferences-ee6953152c1a

[18] Minsky, C. (2020). 'How AI helps historians solve ancient puzzles'. *Financial Times*. Available at: https://www.ft.com/content/2b72ed2c-907b-11ea-bc44-dbf6756c871a

[19] Zheng, S., Trott, A., Srinivasa, S. et al. (2020). 'The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies'. *Salesforce Research*. Available at: https://blog.einstein.ai/the-ai-economist/

[20] Eraslan, G., Avsec, Ž., Gagneur, J. and Theis, F. J. (2019). 'Deep learning: new computational modelling techniques for genomics'. *Nature Reviews Genetics.* Available at: https://doi.org/10.1038/s41576-019-0122-6

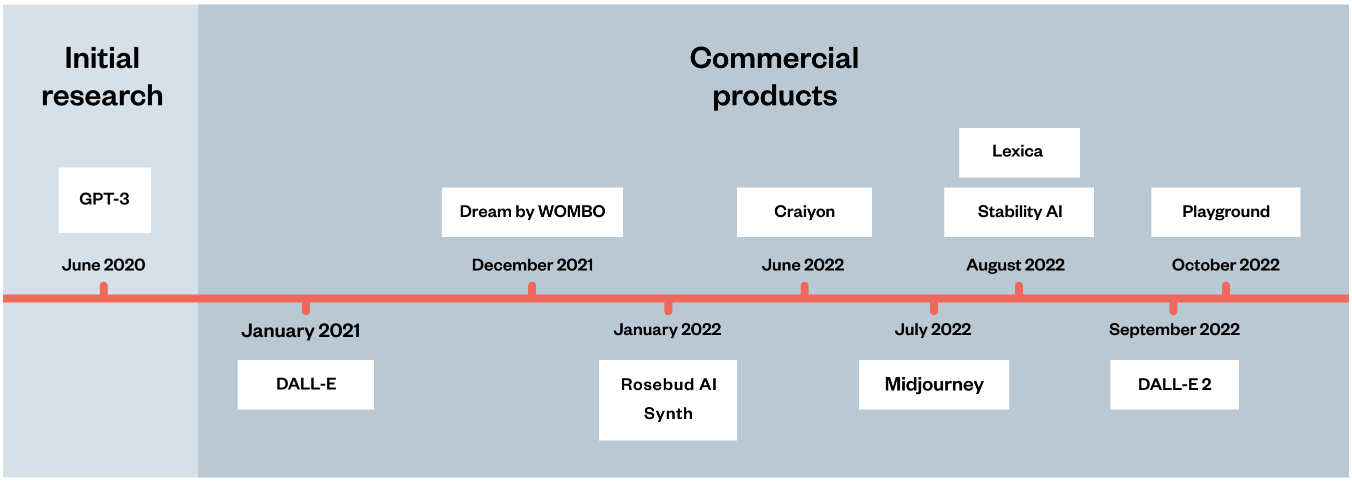
[21] DeepMind. (2020). 'AlphaFold: a solution to a 50-year-old grand challenge in biology'. *DeepMind Blog*. Available at: https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

[22] Boyarskaya, M., Olteanu, A. and Crawford, K. (2020). 'Overcoming Failures of Imagination in AI Infused System Development and Deployment'. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2011.13416

[23] Clifford, C. (2018). 'Google CEO: A.I. is more important than fire or electricity'. *CNBC*. Available at: https://www.cnbc.com/2018/02/01/google-ceo-sundar-pichai-ai-is-more-important-than-fire-electricity.html

[24] Boyarskaya, M., Olteanu, A. and Crawford, K. (2020). 'Overcoming Failures of Imagination in AI Infused System Development and Deployment'. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2011.13416

Figure 2: The research-to-product timeline



This timeline shows how short the research-to-product pipeline for AI can be. It took less than a year from the release of initial research in 2020 and 2021, exploring how to generate images from text inputs, to the first commercial products selling these services.

The sudden growth of AI and data science research has exacerbated challenges for traditional research ethics review processes, and highlighted that they are poorly set up to address questions of broader societal impact of research. Several high-profile instances of controversial AI research passing institutional ethics review include image recognition applications that claim to identify homosexuality,[25] criminality,[26] physiognomy[27] and phrenology.[28] Corporate labs have also experienced high-profile examples of unethical research being approved, including a Microsoft chatbot capable of spreading disinformation,[29] and a Google research paper that contributed to the surveillance of China's Uighur population.[30]

[25] Metcalf, J. (2017). '"The study has been approved by the IRB": Gayface AI, research hype and the pervasive data ethics...' *Medium*. Available at: https://medium.com/pervade-team/the-study-has-been-approved-by-the-irb-gayface-ai-research-hype-and-the-pervasive-data-ethics-ed76171b882c

[26] Coalition for Critical Technology. (2020). 'Abolish the #TechToPrisonPipeline'. *Medium*. Available at: https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16.

[27] Ongweso Jr, E. (2020). 'An AI Paper Published in a Major Journal Dabbles in Phrenology'. *Vice*. Available at: https://www.vice.com/en/article/g5pawq/an-ai-paper-published-in-a-major-journal-dabbles-in-phrenology

[28] Colaner, S. (2020). 'AI Weekly: AI phrenology is racist nonsense, so of course it doesn't work'. *VentureBeat*. Available at: https://venturebeat.com/2020/06/12/ai-weekly-ai-phrenology-is-racist-nonsense-so-of-course-it-doesnt-work/.

[29] Hsu, J. (2019). 'Microsoft's AI Research Draws Controversy Over Possible Disinformation Use'. *IEEE Spectrum*. Available at: https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/microsofts-ai-research-draws-controversy-over-possible-disinformation-use

[30] Harlow, M., Murgia, M. and Shepherd, C. (2019). 'Western AI researchers partnered with Chinese surveillance firms'. *Financial Times*. Available at: https://www.ft.com/content/41be9878-61d9-11e9-b285-3acd5d43599e

This report explores the challenges that public and private-sector RECs face in evaluations of research ethics and broader societal impact issues in AI and data science research

In research institutions, the role of assessing for research ethics issues tends to fall on Research Ethics Committees (RECs), also known in some regions as Institutional Review Boards (IRBs) or Ethics Review Committees (ERCs). Since the 1960s, these committees have been empowered to reject research from being undertaken unless changes are made in the proposed research design.

These committees generally consist of members of a specific academic department or corporate institution, who are responsible for evaluating research proposals before the research begins. Their evaluations combine normative and legal principles, originally linked to biomedical human subjects research, that have developed over time.

Traditionally, RECs only consider research involving human subjects and only consider questions concerning how the research will be conducted. While they are not the only 'line of defence' against unethical practices in research, they are the primary actor responsible for mitigating potential harms to research subjects in many forms of research.

> The increasing prominence of AI and data science research poses an important question: are RECs well placed and adequately set up to address the challenges that AI and data science research pose?

This report explores these challenges that public and private-sector RECs face in evaluations of research ethics and broader societal impact issues in AI and data science research.[31] In doing so, it aims to help institutions that are developing AI research review processes take a holistic and robust approach for identifying and mitigating these risks. It also seeks to provide research institutions and other actors in the research ecosystem – funders, journal editors and conference organisers – with specific recommendations for how they can address these challenges.

[31] This report does not focus on considerations relating to research integrity, though we acknowledge this is an important and related topic.

This report seeks to address four research questions:

1. How are RECs in academia and industry currently structured? What role do they play in the wider research ecosystem?

2. What resources (e.g. moral principles, legal guidance, etc.) are RECs using to guide their reviews of research ethics? What is the scope of these reviews?

3. What are the most pressing or common challenges and concerns that RECs are facing in evaluations of AI and data science research?

4. What changes can be made so that RECs and the wider AI and data science research community can better address these challenges?

To address these questions, this report relied on a review of the literature on RECs, research ethics and broader societal impact questions in AI. The report also draws on a series of workshops with 42 members of public and private AI and data science research institutions in May 2021, along with eight interviews with experts in research ethics and AI issues. More information on our methodology can be found in 'Methodology and limitations' on page 100.

This report begins with an introduction to the history of RECs, how they are commonly structured, and how they commonly operate in corporate and academic environments for AI and data science research. The report then discusses six challenges that RECs face – some of which are longstanding issues, others of which are exacerbated by the rise of AI and data science research. We conclude the paper with a discussion of these findings and eight recommendations for actions that RECs and other actors in the research ecosystem can take to better address the ethical risks of AI and data science research.

Context for Research Ethics Committees and AI research

This section provides a brief history of modern research ethics and Research Ethics Committees (RECs), discusses their scope and function, and highlights some differences between how they operate in corporate and academic environments. It places RECs in the context of other actors in the 'AI research ecosystem', such as organisers of AI and data science conferences, or editors of AI journal publications who set norms of behaviour and incentives within the research community. Three key points to take away from this chapter are:

1. Modern research ethics questions are mostly focused on ethical challenges that arise in research methodology, and exclude consideration of the broader societal impacts of research.

2. Current RECs and research ethics principles stem from biomedical research, which analyses questions of research ethics through a lens of patient-clinician relationships and is not well suited for the more distanced relationship in AI and data science between a researcher and data subject.

3. Academic and corporate RECs in AI research share common aims, but with some important differences. Corporate AI labs tend to have more resources, but may also be less transparent about their processes.

What is a REC, and what is its scope and function?

Every day, RECs review applications to undertake research for potential ethical issues that may arise. Broadly defined, RECs are institutional bodies made up of members of an institution (and, in some instances, independent members outside that institution) who are charged with evaluating applications to undertake research before it begins. They make judgements about the suitability of research, and have the power to approve researchers to go ahead with a project or request that changes are made before research is undertaken. Many academic journals and conferences will not publish or accept research that fails to meet a review by a Research Ethics Committee (though as we will discuss below, not all research requires review).

RECs operate with two purposes in mind:

1. To protect the welfare and interests of prospective and current research participants and minimise risk of harm to them.

2. To promote ethical and societally valuable research.

In meeting these aims, RECs traditionally conduct an *ex ante* evaluation only once, before a research project begins. In understanding what kinds of ethical questions RECs evaluate for, it is also helpful to disentangle three distinct categories of ethical risks in research:[32]

1. Mitigating research process harms (often confusingly called 'research ethics').

2. Research integrity.

3. Broader societal impacts of research (also referred to as Responsible Research and Innovation, or RRI).

The scope of REC evaluations is entirely on questions of mitigating the ethical risks from research methodology, such as how the researcher intends to protect the privacy of a participant, anonymise their data or ensure they have received informed consent.[33] In their evaluations, RECs may look at whether the research poses a serious risk to interests and safety of research subjects, or if the researchers are operating in accordance with local laws governing data protection and intellectual property ownership of any research findings.

[32] For a deeper discussion on these issues, see: Ashurst, C. et al. (2022). 'Disentangling the Components of Ethical Research in Machine Learning'. *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2057–2068. Available at: https://doi.org/10.1145/3531146.3533781

[33] Dove, E. S., Townend, D., Meslin, E. M. et al. (2016). 'Ethics review for international data-intensive research'. *Science*, 351(6280), pp. 1399–1400.

REC evaluations may also probe on whether the researchers have assessed and minimised potential harm to research participants, and seek to balance this against the benefits of the research for society at large.[34] However, there are limitations to the aim of promoting ethical and societally valuable research. There are few frameworks for how RECs can consider the benefit of research for society at large. Additionally, this concept of mitigating methodological risks does not extend to considerations of whether the research poses risks to society at large, or to individuals beyond the subjects of that research.

Three different kinds of ethical risks in research

1. **Mitigating research process (also known as 'research ethics'):** The term research ethics refers to the principles and processes governing how to mitigate the risks to research subjects. Research ethics principles are mostly concerned with the protection, safety and welfare of individual research participants, such as gaining their informed consent to participate in research or anonymising their data to protect their privacy.

2. **Research integrity:**These are principles governing the credibility and integrity of the research, including which whether it is intellectually honest, transparent, robust, and replicable.[35] In most fields, research integrity is evaluated via the peer review process after research is completed.

3. **Broader societal impacts of research:** This refers to the potential positive and negative societal and environmental implications of research, including unintended uses (such as misuse) of research. A similar concept is **Responsible Research and Innovation (RRI)** which refers to steps that researchers can undertake to anticipate and address the potential downstream risks and implications of their research.[36]

RECs, however, often do not evaluate for questions of research integrity, which is concerned with whether research is intellectually honest,

[34] Dove, E. S., Townend, D., Meslin, E. M. et al. (2016).

[35] UKRI. 'Research integrity'. Available at: https://www.ukri.org/what-we-offer/supporting-healthy-research-and-innovation-culture/research-integrity/

[36] Engineering and Physical Sciences Research Council. 'Responsible research and innovation'. *UKRI*. Available at: https://www.ukri.org/councils/epsrc/guidance-for-applicants/what-to-include-in-your-proposal/health-technologies-impact-and-translation-toolkit/research-integrity-in-healthcare-technologies/responsible-research-and-innovation/

transparent, robust and replicable.[37] These can include questions relating to whether data has been fabricated or misrepresented, whether research is reproducible, stating the limitations and assumptions of the research, and disclosing conflicts of interests.[38] The intellectual integrity of researchers is important for ensuring public trust in science, which can be eroded in cases of misconduct.[39]

Some RECs may consider complaints about research integrity issues that arise after research has been published, but these issues are often not considered as part of their ethics reviews. RECs may, however, assess a research applicant's bona fides to determine if they are someone who appears to have integrity (such as if they have any conflicts of interest with the subject of their study). Usually, questions of research integrity are left to other actors in the research ecosystem, such as peer reviewers and whistleblowers who may notify a research institution or the REC of questionable research findings or dishonest behaviour. Other governance mechanisms for addressing research integrity issues include publishing the code or data of the research so that others may attempt to reproduce findings.

Another area of ethical risks that contemporary RECs do not evaluate for (but which we argue they should – see page 72) is the responsibility of researchers to consider the broader societal effects of their research on society.[40] This is referred to as **Responsible Research and Innovation (RRI)**, which moves beyond concerns of research integrity and is: 'an approach that anticipates and assesses potential implications and societal expectations with regard to research and innovation, with the aim to foster the design of inclusive and sustainable research and innovation'.[41]

RRI is concerned with the integration of mechanisms of reflection, anticipation and inclusive deliberation around research and innovation, and relies on individual researchers to incorporate these practices in their research. This includes analysing potential economic, societal or

[37] UKRI. 'Research integrity'. Available at: https://www.ukri.org/what-we-offer/supporting-healthy-research-and-innovation-culture/research-integrity/

[38] Partnership on AI. (2021). *Managing the Risks of AI Research*. Available at: http://partnershiponai.org/wp-content/uploads/2021/08/PAI-Managing-the-Risks-of-AI-Resesarch-Responsible-Publication.pdf

[39] Korenman, S. G., Berk, R., Wenger, N. S. and Lew, V. (1998). 'Evaluation of the research norms of scientists and administrators responsible for academic research integrity'. *Jama*, 279(1), pp. 41–47.

[40] Douglas, H. (2014). 'The moral terrain of science'. *Erkenntnis*, 79(5), pp. 961–979.

[41] European Commission. (2018). *Responsible Research and Innovation, Science and Technology*. Available at: https://data.europa.eu/doi/10.2777/45726

environmental impacts that arise from research and innovation. RRI is a more recent development that emerged separately to RECs, stemming in part from the Ethical Legal and Societal Implications Research (ELSI) programme in the 1990s, which was established to research the broader societal implications of genomics research.[42]

Traditionally, RECs are usually not well equipped to deal with assessing subsequent uses of research, or their impacts on society. RECs often lack the capacity or remit to monitor the downstream uses of research, or to act as an 'observatory' for identifying trends in the use or misuse of research they reviewed at inception. This is compounded by the decentralised and fragmentary nature of RECs, which operate independently of each other and often do not evaluate each other's work.

What principles do RECs rely on to make judgements about research ethics?

In their evaluations, RECs rely on a variety of tools, including laws like the General Data Protection Regulation (GDPR), which cover data protection issues and some discipline-specific norms. At the core of all Research Ethics Committee evaluations, there are a series of moral principles that have evolved over time. These principles largely stem from the biomedical sciences, and have been codified, debated and edited by international bodies like the World Medical Association and World Health Organisation. The biomedical model of research ethics is the foundation for how concepts like autonomy and consent were encoded in law,[43] which often motivate modern discussions about privacy.

Some early modern research ethics codes, like the Nuremberg Principles and the Belmont Report, were developed in response to specific atrocities and scandals involving biomedical research on human subjects. Other codes, like the Declaration of Helsinki, developed out of a field-wide concern to self-regulate before governments stepped in to regulate.[44]

[42] National Human Genome Research Institute. 'Ethical, Legal and Social Implications Research Program'. Available at: https://www.genome.gov/Funded-Programs-Projects/ELSI-Research-Program-ethical-legal-social-implications

[43] Bazzano, L. A. et al. (2021). 'A Modern History of Informed Consent and the Role of Key Information'. *Ochsner Journal*, 21(1), pp. 81–85. Available at: https://doi.org/10.31486/toj.19.0105

[44] Hedgecoe, A. (2017). 'Scandals, Ethics, and Regulatory Change in Biomedical Research'. *Science, Technology, & Human Values*, 42(4), pp. 577–599. Available at: https://journals.sagepub.com/doi/abs/10.1177/0162243916677834

Each code and declaration seeks to address specific ethical issues from a particular regional and historical context. Nonetheless, they are united by two aspects. Firstly, they frame research ethics questions in a way that assumes a clear researcher-subject relationship. Secondly, they all seek to standardise norms of evaluating and mitigating the potential risks caused by research processes, to support REC decisions becoming more consistent between different institutions.

Historical principles governing research ethics

Nuremberg Code: The Nuremberg trials occurred in 1947 and revealed horrific and inhumane medical experimentation by Nazi scientists on human subjects, primarily concentration camp prisoners. Out of concern that these atrocities might further damage public trust in medical professionals and research,[45] the judges in this trial included a set of universal principles for 'permissible medical experiments' in their verdict, which would later become known as the Nuremberg Code.[46] The Code lists ten principles that seek to ensure individual participant rights are protected and outweigh any societal benefit of the research.

Declaration of Helsinki: Established by World Medical Association (WMA), the Helsinki Declaration seeks to articulate universal principles for human subjects research and clinical research practice. The WMA is an international organisation representing physicians from across the globe. The Helsinki Declaration has been updated repeatedly since its first iteration in 1964, with major updates occurring in 1975, 2000 and 2008. It specifies five basic principles for all human subjects research, as well as further principles specific to clinical research.

Belmont Report: This report was written in response to several troubling incidents in the USA, in which patients participating in clinical trials were not adequately informed about the risks involved. These include a 40-year-long experiment by the US Public Health Service and the Tuskegee Institute that sought to study untreated syphilis in Black men. Despite having over 600 participants (399 with syphilis, 201 without), the participants were deceived about the risks and nature of experiment and were not provided with a cure for

[45] Israel, M. (2015). *Research Ethics and Integrity for Social Scientists,* second edition. SAGE Publishing. Available at: https://uk.sagepub.com/en-gb/eur/research-ethics-and-integrity-for-social-scientists/book236950

[46] The Nuremberg Code was in part based on pre-war medical research guidelines from the German Medical Association, which included elements of patient consent to a procedure. These guidelines were disused during the rise of the Nazi Regime in favour of guidelines that contributed to the 'healing of the nation', as defendants at the Nuremberg trial put it. See: Ernst, E. and Weindling, P. J. (1998). 'The Nuremberg Medical Trial: have we learned the lessons?' *Journal of Laboratory and Clinical Medicine*, 131(2), pp. 130–135; and British Medical Journal. (1996). 'Nuremberg'. British Medical Journal, 313(7070). Available at: https://www.bmj.com/content/313/7070

the disease after it had been developed in the 1940s.[47] These developments led to the United States' National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research to publish the Belmont Report in 1979, which listed several principles for research to follow: justice, beneficence and respect for persons.[48]

Council for International Organizations of Medical Sciences Guidelines (CIOMS): CIOMS was formed in 1949 by the World Health Organisations and the United Nations Educational, Scientific and Cultural Organisation (UNESCO), and is made up of a range of biomedical member organisations from across the world. In 2016, it published the *International Ethical Guidelines for Health-Related Research Involving Humans*,[49] which includes specific requirements for research involving vulnerable persons and groups, compensation for research participants, and requirements for researchers and health authorities to engage potential participants and communities in a 'meaningful participatory process' in various stages of research.[50]

Biomedical research ethics principles touch on a wide variety of issues, including autonomy and consent. The Nuremberg Code specified that, for research to proceed, a researcher must have consent given (i) voluntarily by a (ii) competent and (iii) informed subject (iv) with adequate comprehension. At the time, consent was understood as only applicable to healthy, non-patient participants, and thus excluded patients in clinical trials, access to patient information like medical registers and participants (like children or people with a cognitive impairment) who are unable to give consent.

Subsequent research ethics principles have adapted to these scenarios with methods such as legal guardianship, group or community consent,

[47] Center for Disease Control and Prevention. (2021). *The U.S. Public Health Service Syphilis Study at Tuskegee*. Available at: https://www.cdc.gov/tuskegee/timeline.htm

[48] The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report*.

[49] Council for International Organizations of Medical Sciences (CIOMS). (2016). *International Ethical Guidelines for Health-related Research Involving Humans*, Fourth Edition. Available at: https://cioms.ch/wp-content/uploads/2017/01/WEB-CIOMS-EthicalGuidelines.pdf

[50] A more extensive study of the history of research ethics is provided by: Garcia, K. et al. (2022). 'Introducing An Incomplete History of Research Ethics'. *Open Life Sciences*. Available at: https://openlifesci.org/posts/2022/08/08/An-Incomplete-History-Of-Research-Ethics/

and broad or blanket consent.[51] Under the Helsinki Declaration, consent must be given in writing and states that research subjects can give consent only if they have been fully informed of the study's purpose, the methods, risks and benefits involved, and their right to withdraw.[52] In all these conceptions of consent, there is a clearly identifiable research subject, who is in some kind of direct relationship with a researcher.

Another area that biomedical research principles touch on is the risk and benefit of research for research subjects. While the Nuremberg Code was unambiguous about the protection of research subjects, the Helsinki Declaration introduced the concept of benefit from research in proportion to risk.[53] The 1975 document and other subsequent revisions reaffirmed that, 'while the primary purpose of medical research is to generate new knowledge, this goal can never take precedence over the rights and interests of individual research subjects.'[54]

However, Article 21 recommends that research can be conducted if the importance of its objective outweighs the risks to participants, and Article 18 states that a careful assessment of predictable risks to participants must be undertaken in comparison to potential benefits for individuals and communities.[55] The Helsinki Declaration lacks clarity

[501] Hoeyer, K. and Hogle, L. F. (2014). 'Informed consent: The politics of intent and practice in medical research ethics'. *Annual Review of Anthropology*, 43, pp. 347–362; Legal guardianship: The Helsinki Declaration specifies that underrepresented groups should have adequate access to research and to the results of research. However, vulnerable population groups are often excluded from research if they are not able to give informed consent. A legal guardian is usually appointed by a court and can give consent on the participants' behalf, see: Brune C,, Stentzel U., Hoffmann W. and van den Berg, N. (2021). 'Attitudes of legal guardians and legally supervised persons with and without previous research experience towards participation in research projects: A quantitative cross-sectional study'. *PLoS ONE*, 16(9); Group or community consent refers to research that can generate risks and benefits as part of the wider implications beyond the individual research participant. This means that consent processes may need to be supplemented by community engagement activities, see: Molyneux, S. and Bull, S. (2013). 'Consent and Community Engagement in Diverse Research Contexts: Reviewing and Developing Research and Practice: Participants in the Community Engagement and Consent Workshop, Kilifi, Kenya, March 2011'. *Journal of Empirical Research on Human Research Ethics (JERHRE)*, 8(4), pp. 1–18. Available at: https://doi.org/10.1525/jer.2013.8.4.1
Blanket consent refers to a process by which individuals donate their samples without any restrictions. Broad (or 'general') consent refers to a process by which individuals donate their samples for a broad range of future studies, subject to specified restrictions, see: Wendler, D. (2013). 'Broad versus blanket consent for research with human biological samples'. *The Hastings Center report*, 43(5), pp. 3–4. Available at: https://doi.org/10.1002/hast.200

[52] World Medical Association. (2008). *WMA Declaration of Helsinki – ethical principles for medical research involving human subjects.* Available at: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

[53] Ashcroft, R. 'The Declaration of Helsinki' in: Emanuel, E. J., Grady, C. C., Crouch, R. A., Lie, R. K., Miller, F. G. and Wendler, D. D. (eds.). (2008). *The Oxford textbook of clinical research ethics*. Oxford University Press.

[54] World Medical Association. (2008). *WMA Declaration of Helsinki – ethical principles for medical research involving human subjects.* Available at: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

[55] World Medical Association. (2008).

on what constitutes an acceptable, or indeed 'predictable' risk and how the benefits would be assessed, and therefore leaves the challenge of resolving these questions to individual institutions.[56] The CIOMS guidance also suggests RECs should consider the 'social value' of health research in considering a cost/benefit analysis.

The Belmont Report also addressed the trade-off between societal benefit and individual risk, offering specific ethics principles to guide scientific research that include 'respects for persons', 'beneficence' and 'justice'.[57] The principle of 'respect for persons' is broken down into respect for the autonomy of human research subjects and requirements for informed consent. The principle of 'beneficence' requires the use of the best possible research design to maximise benefits and minimise harms, and prohibits any research that is not backed by a favourable risk-benefit ratio (to be determined by a REC). Finally, the principle of 'justice' stipulates that the risks and benefits of research are distributed fairly, research subjects are selected through fair procedures, and to avoid any exploitation of vulnerable populations.

The Nuremberg Code created standardised requirements to identify who bears responsibility for identifying and addressing potential ethical risks of research. For example, the Code stipulates that the research participants have the right to withdraw (Article 9), but places responsibility on the researchers to evaluate and justify any risks in relation to human participation (Article 6), to minimise harm (Articles 4 and 7) and to stop the research if it is likely to cause injury or death to participants (Articles 5 and 10).[58] Similar requirements exist in other biomedical ethical principles like the Helsinki Declaration, which extends responsibility for assessing and mitigating ethical risks to both researchers and RECs.

[56] Millum, J., Wendler, D. and Emanuel, E. J. (2013). 'The 50th anniversary of the Declaration of Helsinki: progress but many remaining challenges'. *Jama*, 310(20), pp. 2143–2144.

[57] The Belmont Report was published by the National Commission for the Protection of Human Subjects in Biomedical and Behavioral Research, which was created for the U.S. Department of Health, Education, and Welfare (DHEW) based on authorisation by the U.S. Congress in 1974. The National Commission had been tasked by the U.S. Congress with the identification of guiding research ethics principles in response to public outrage over the Tuskegee Syphilis Study and other ethically questionable projects that emerged during this time.

[58] The Nuremberg Code failed to deal with several related issues, including how international research trial should be run, questions of care for research subjects after the trial has ended or how to assess the benefit of the research to a host community. See: Annas, G. and Grodin, M. (2008). *The Nazi Doctors and the Nuremberg Code: Human Rights in Human Experimentation*. Oxford University Press

A brief history of RECs in the USA and the UK

RECs are a relatively modern phenomenon in the history of academic research, and their origins stem from early biomedical research initiatives of the 1970s. The 1975 Declaration of Helsinki, an initiative by the World Medical Association (WMA) to articulate universal principles for human subjects research and clinical research practice, declared the ultimate arbiter for making assessments of ethical risk and benefit were specifically appointed, independent Research Ethics Committees who are given the responsibility to assess the risk of harm to research subjects and the management of those risks.

In the USA, the National Research Act of 1974 requires Institutional Review Board (IRB) approval for all human subjects research projects funded by the US Department of Health, Education, and Welfare (DHEW). [59] This was extended in 1991 under the 'Common Rule' so that any research involving human subjects that is funded by the federal government must undergo an ethics review by an IRB. There are certain exceptions for what kinds of research will go before an IRB, including research that involves the analysis of data that is publicly available, privately funded research, and research that involves secondary analysis of existing data (such as the use of existing 'benchmark' datasets that are commonly used in AI research).[60]

In the UK, the first RECs began operating informally around 1966, in the context of clinical research in the National Health Service (NHS), but it was not until 1991 that RECs were formally codified. In the 1980s, the UK expanded the requirement for REC review beyond clinical health research into other disciplines. Academic RECs in the UK began to spring up around this same time, with the majority coming into force after the year 2000.

UK RECs in the healthcare and clinical context are coordinated and regulated by the Health Research Authority, which has passed guidance for how medical healthcare RECs should be structured and operate, including the procedure of submitting an ethics application and the process of ethics review.[61] This guidance allows for greater harmony across different health RECs and better governance for multi-site research projects, but this guidance does not extend to RECs in other academic fields. Some funders such as the UK's Economic and Social Research Council have also released

[59] In 1991, the regulations of the DHEW became a 'common rule' that covered 16 federal agencies.

[60] Office for Human Research Protections. (2009). *Code of Federal Regulations, Part 46: Protection of Human Subjects*. Available at: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html

[61] In 2000, the Central Office for Research Ethics was formed, followed by the establishment of the National Research Ethics Service and later the Health Research Authority (HRA). See: NHS Health Research Authority. (2021). *Research Ethics Committees – Standard Operating Procedures*. Available at: https://www.hra.nhs.uk/about-us/committees-and-services/res-and-recs/research-ethics-committee-standard-operating-procedures/

research ethics guidelines for non-health projects to undergo certain ethics review requirements if the project involves human subjects research (though the definition of human subjects research is contested).[62]

RECs in academia

While RECs broadly seek to protect the welfare and interests of research participants and promote ethical and societally valuable research, there are some important distinctions to draw between the function and role of a REC in academic institutions compared to private-sector AI labs.

Where are RECs located in universities and research institutes?

Academic RECs bear a significant amount of the responsibility for assessing research involving human participants, including the scrutiny of ethics applications from staff and students. Broadly, there are two models of RECs used in academic research institutions:

1. **Centralised:** A single, central REC is responsible for all research ethics applications, including the development of ethics policies and guidance.

2. **Decentralised:** Schools, faculties or departments have their own RECs for reviewing applications, while a central REC maintains and develops ethics policies and guidance.[63]

RECs can be based at the institutional level (such as at universities), or at the regional and federal level. Some RECs may also be run by non-academic institutions, who are charged with reviewing academic research proposals. For example, academic health research in the

[62] There is some guidance for non-health RECs in the UK – the Economic and Social Science Research Council released research ethics guidelines for any project funded by ESRC to undergo certain ethics review requirements if the project involves human subjects research. See: Economic and Social Research Council. (2015). *ESRC Framework for Research Ethics*. UKRI. Available at: https://www.ukri.org/councils/esrc/guidance-for-applicants/research-ethics-guidance/framework-for-research-ethics/

[63] Tinker, A. and Coomber, V. (2005). 'University research ethics committees – A summary of research into their role, remit and conduct'. *Research Ethics*, 1(1), pp. 5–11.

UK may undergo review by RECs run by the National Health Service (NHS), sometimes in addition to review by the academic body's own REC. In practice, this means that publicly funded health research proposals may seek ethics approval from one of the 85 RECs run by the NHS, in addition to non-NHS RECs run by various academic departments.[64]

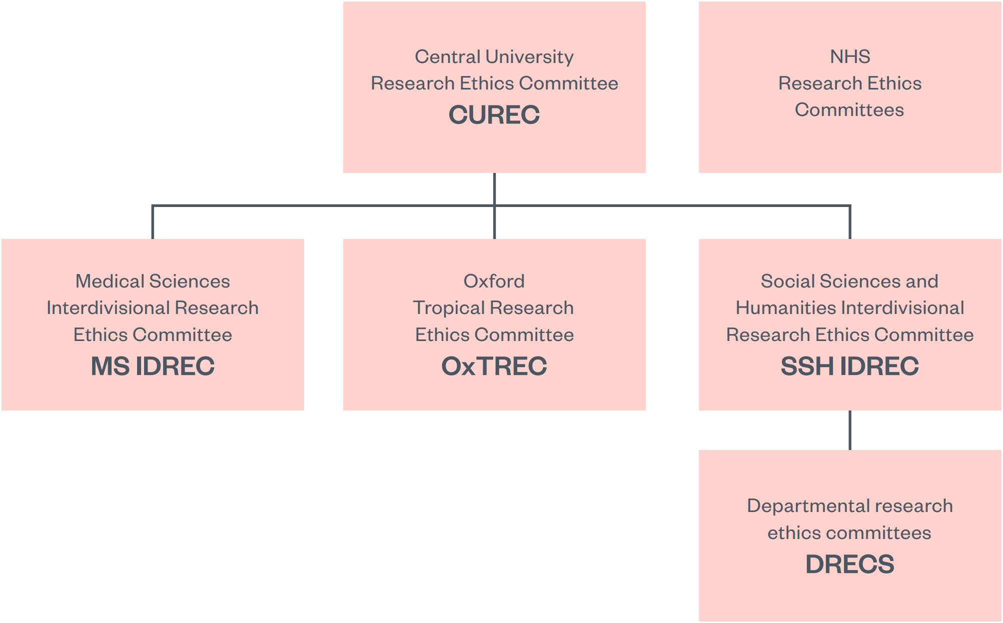
A single, large academic institution, such as the University of Cambridge, may have multiple committees running within it, each with a different composition and potentially assessing different kinds of fields of research. Depending on the level of risk and required expertise, a research project may be reviewed by a local REC, school-level REC or may also be reviewed by a REC at the university level.[65]

For example, Exeter University has a central REC and 11 devolved RECs at college or discipline level. The devolved RECS report to the central REC, which is accountable to the University Council (governing body). Exeter University also implements a 'dual assurance' scheme, with an independent member of the university's governing body providing oversight of the implementation of their ethics policy (see page 106 for more details). The University of Oxford also relies on a cascading system of RECs, which can escalate concerns up the chain if needed, and which may include department and domain-specific guidance for certain research ethics issues.

[64] European Network of Research Ethics Committees. 'Short description of the UK REC system'. Available at: http://www.eurecnet.org/information/uk.html

[65] University of Cambridge. 'Ethical Review'. Available at: https://www.research-integrity.admin.cam.ac.uk/ethical-review

Figure 3: The cascade of RECs at the University of Oxford[66]


Central University Research Ethics Committee
CUREC
NHS Research Ethics Committees
Medical Sciences Interdivisional Research Ethics Committee
MS IDREC
Oxford Tropical Research Ethics Committee
OxTREC
Social Sciences and Humanities Interdivisional Research Ethics Committee
SSH IDREC
Departmental research ethics committees
DRECS

This figure shows how one academic institution's RECs are structured, with a central REC and more specialised committees.

What is the scope and role of academic RECs?

According to a 2004 survey of UK academic REC members, they play four principal roles:[67]

1. Responsibility for ethical issues relating to research involving human participants, including maintaining standards and provision of advice to researchers.

2. Responsibility for ensuring production and maintenance of codes of practice and guidance for how research should be conducted.

[66] University of Oxford. 'Committee information: Structure, membership and operation of University research ethics committees'. Available at: https://researchsupport.admin.ox.ac.uk/governance/ethics/committees

[67] Tinker, A. and Coomber, V. (2005). 'University Research Ethics Committees — A Summary of Research into Their Role, Remit and Conduct'. *SAGE Journals*. Available at: https://doi.org/10.1177/174701610500100103

3. Ethical scrutiny of research applications from staff and, in most cases, students.

4. Reporting and monitoring of instances of unethical behaviour to other institutions or academic departments.

Academic RECs often include a function for intaking and assessing reports of unethical research behaviour, which may lead to disciplinary action against staff or students.

When do ethics reviews take place?

RECs form a gateway through which researchers apply to obtain ethics approval as a prerequisite for further research. At most institutions, researchers will submit their work for ethics approval before conducting the study – typically at the early stages in the research lifecycle, such as at the planning stage or when applying for research grants. This means RECs only consider an **anticipatory assessment** of ethical risks that the proposed method may raise.

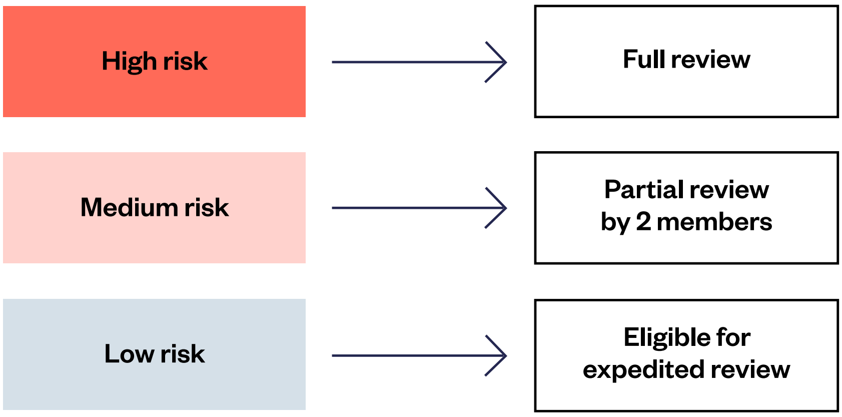
This assessment relies on both 'testimony' from research applicants who document what they believe are the material risks, and a review by REC members themselves who assess the validity of that 'testimony', provide an opinion of what they envision the material risks of the research method might be, and how those risks can be mitigated. There is limited opportunity for revising these assessments once the research is underway, and that usually only occurs if a REC review identifies a risk or threat and asks for additional information. One example of an organisation that takes a different approach is the Alan Turing Institute, which developed a continuous integration approach with reviews taking place at various stages throughout the research life cycle (see page 102 for more details).[68]

The extent of a REC's review will vary depending on whether the project has any clearly identifiable risks to participants, and many RECs apply a triaging process to identify research that may pose particularly significant risks. RECs may use a checklist that asks a researcher

[68] The Turing Way Community et al. Guide for *Ethical Research – Introduction to Research Ethics*. Available at: https://the-turing-way.netlify.app/ethical-research/ethics-intro.html

whether their project involves particularly sensitive forms of data collection or risk, such as research with vulnerable population groups like children, or research that may involve deceiving research participants (such as creating a fake account to study online right-wing communities). If an application raises one of these issues, it must undergo a full research ethics review. In cases where a research application does not involve any of these initial risks, it may undergo an expedited process that involves a review of only some factors of the application such as its data governance practices.[69]

Figure 4: Example of the triaging application intake process for a UK University REC



Examples of high risk

- Participants who are children under the age of 16 years
- Participants in a potentially vulnerable situation
- Recruiting participants by offering financial or other rewards
- Recruiting participants via 'snowballing'
- Participants who do not have the option to be debriefed of the purpose of the project
- Participants who may be identified
- Informed consent will not be obtained
- Participants who take part in the study without their knowledge and consent
- Covert observation in a physical or online environment
- The safety of the researcher may be in question

[69] For an example of a full list of risks and the different processes, see: University of Exeter. (2021). *Research Ethics Policy and Framework: Appendix C – Risk and Proportionate Review checklist*. Available at: https://www.exeter.ac.uk/media/universityofexeter/governanceandcompliance/researchethicsandgovernance/Appendix_C_Risk_and_Proportionate_Review_v1.1_07052021.pdf;
and University of Exeter. (2021). *Research Ethics Policy and Framework*. Available at: https://www.exeter.ac.uk/media/universityofexeter/governanceandcompliance/researchethicsandgovernance/Revised_UoE_Research_Ethics_Framework_v1.1_07052021.pdf.

If projects meet certain risk criteria, they may be subject to a more extensive review by the full committee. Lower-risk projects may be approved by only one or two members of the committee.

During the review, RECs may offer researchers advice to mitigate potential ethical risks. Once approval is granted, no further checks by RECs are required. This means that there is no mechanism for ongoing assessment of emerging risks to participants, communities or society as the research progresses. As the focus is on protecting individual research participants, there is no assessment of potential long-term downstream harms of research.

Composition of academic RECs

The composition of RECs varies between and even within various institutions. In the USA, RECs are required under the 'common rule' to have a minimum of five members with a variety of professional backgrounds, to be made up of people from different ethnic and cultural backgrounds, and to have at least one member who is independent from the institution. In the UK, the Health Research Authority recommends RECs have 18 members, while the Economic and Social Research Council (ESRC) recommends at least seven.[70] RECs operate on a voluntary basis, and there is currently no financial compensation for REC members, nor any other rewards or recognition.

Some RECs are comprised of an interdisciplinary board of people who bring different kinds of expertise to ethical reviews. In theory, this is to provide a more holistic review of research that ensures perspectives from different disciplines and life experiences are factored into a decision. RECs in the clinical context in the UK, for example, must involve both expert members with expertise in the subject area and 'lay members', which refers to people 'who are not registered healthcare professionals and whose primary professional interest is not in clinical research'.[71] Additional expertise can be sourced on an ad hoc

[70] NHS Health Research Authority. (2021). *Governance arrangements for Research Ethics Committees*. Available at: https://www.hra.nhs.uk/planning-and-improving-research/policies-standards-legislation/governance-arrangement-research-ethics-committees/; and Economic and Social Research Council. (2015). *ESRC Framework for Research Ethics*. UKRI. Available at: https://www.ukri.org/councils/esrc/guidance-for-applicants/research-ethics-guidance/framework-for-research-ethics/

[71] NHS Health Research Authority. (2021). *Research Ethics Committee – Standard Operating Procedures*. Available at: https://www.hra.nhs.uk/about-us/committees-and-services/res-and-recs/research-ethics-committee-standard-operating-procedures/

basis.[72] The ESRC also further emphasises that RECs should be multi-disciplinary and include ethnic and gender diversity.[73] According to our expert workshop participants, however, many RECs that are located within a specific department of faculty are often not multi-disciplinary and do not include lay members, although specific expertise might be requested when needed.

The Secure Anonymised Information Linkage databank (SAIL)[74] offers one example of a body that does integrate lay members in their ethics review process. Their review criteria include data governance issues and risks of disclosure, but also whether the project contributes to new knowledge, and whether it serves the public good by improving health, wellbeing and public services (see page 104 for more details).

RECs within the technology industry

In the technology industry, several companies with AI and data science research divisions have launched internal ethics review processes and accompanying RECs, with notable examples being Microsoft Research, Meta Research and Google Brain. In our workshop and interviews with participants, members of corporate RECs we spoke with noted some key facets of their research review processes. It is important, however, to acknowledge that little publicly available information exists on corporate REC practices, including their processes and criteria for research ethics review. This section reflects statements made by workshop and interview participants, and some public reports of research ethics practices in private-sector labs.

Scope

According to our participants, corporate AI research RECs tend to take a broader scope of review than traditional academic RECs. Their reviews may extend beyond research ethics issues and into questions of broader societal impact. Interviews with developers of AI ethics review practices in industry suggested a view that traditional REC models can be too

[72] NHS Health Research Authority. (2021).

[73] Economic and Social Research Council. (2015). *ESRC Framework for Research Ethics*. UKRI. Available at: https://www.ukri.org/councils/esrc/guidance-for-applicants/research-ethics-guidance/framework-for-research-ethics/

[74] See: saildatabank.com

cumbersome and slow for the quick pace of the product development life cycle.

At the same time, *ex ante* review does not provide good oversight on risks that emerge during or after a project. To address this issue, some industry RECs have sought to develop processes that focus beyond protecting individual research subjects and include considerations for the broader downstream effects for population groups or society, as well as recurring review throughout the research/product lifecycle.[75]

Several companies we spoke with have specific RECs that review research involving human subjects. However, as one participant from a corporate REC noted, 'a lot of AI research does not involve human subjects' or their data, and may focus instead on environmental data or other types of non-personal information. This company relied on separate ethics review process for such cases that considers (i) the potential broader impact of the research and (ii) whether the research aligns with public commitments or ethical principles the company has made.

According to a law review article on their research ethics review process, Meta (previously known as Facebook) claims to consider the public contribution of knowledge of research and whether it may generate positive externalities and implications for society.[76] A workshop participant from another corporate REC noted that 'the purpose of [their] research is to have societal impact, so ethical implications of their research are fundamental to them.' These companies also tend to have more resources to undertake ethical reviews than academic labs, and can dedicate more full-time staff positions to training, broader impact mapping and research into the ethical implications of AI.

[75] Moss, E. and Metcalf, J. (2020). *Ethics Owners. A New Model of Organizational Responsibility in Data-Driven Technology Companies.* Data & Society. Available at: https://datasociety.net/library/ethics-owners/

[76] We note this article reflects Facebook's process in 2016, and that this process may have undergone significant changes since that period. See: Jackman, M. and Kanerva, L. (2016). 'Evolving the IRB: building robust review for industry research'. *Washington and Lee Law Review Online*, 72(3), p. 442.

The use of AI-specific ethics principles and red lines

Many corporate companies like Meta, Google and Microsoft have published AI ethics principles that articulate particular considerations for their AI and data science research to consider, as well as 'red line' research areas they will not undertake. For example, in response to employee protests against a US Department of Defense contract, Google stated it will not pursue AI 'weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people'.[77] Similarly, DeepMind and Element AI have signed a pledge against AI research for lethal autonomous weapons alongside over 50 other companies; a pledge that only a handful of academic institutions have made.[78]

According to some participants, articulating these principles can make more salient the specific ethical concerns that researchers at corporate labs should consider with AI and data science research. However, other participants we spoke with noted that, in practice, there is a lack of internal and external transparency around how these principles are applied.

> Many participants from academic institutions we spoke with noted they do not use 'red line' areas of research out of concern that these red lines may infringe on existing principles of academic openness.

Extent of reviews

Traditional REC reviews tend to focus on a single one-off assessment of research risk at the early stages of a project. In contrast, one corporate REC we spoke with described their review as being a continuous process in which a team may engage with the REC at different stages, such as when a team is collecting data prior to publication, and post-publication reviews into whether the outcomes and impacts they were concerned

[77] See: Google AI. 'Artificial Intelligence at Google: Our Principles'. Available at: https://ai.google/principles/.

[78] Future of Life Institute. (2018). *Lethal autonomous weapons pledge.* Available at: https://futureoflife.org/2018/06/05/lethal-autonomous-weapons-pledge/

with came to fruition. This kind of continuous review enables a REC to capture risks as they emerge.

We note that it was unclear whether this practice was common among industry labs or reflected one lab's particular practices. We also note that some academic labs, like the Alan Turing Institute, are implementing similar initiatives to engage researchers at various stages of the research lifecycle.

A related point flagged by some workshop participants was that industry ethics review boards may vary in terms of their power to affect product design or launch decisions. Some may make non-binding recommendations, and others can green light or halt projects, or return a project to a previous development stage with specific recommendations.[79]

Composition of board and external engagement

The corporate REC members we spoke with all described the composition of their boards as being interdisciplinary and reflecting a broad range of teams at the company. One REC, for example, noted that members of engineering, research, legal and operations teams sit on their ethical review committee to provide advice not only on specific projects, but also for entire research programmes. Another researcher we spoke with described how their organisation's ethics review process provides resources for researchers, including a list of 'banned' publicly accessible datasets that have questionable consent and privacy issues but are commonly used by researchers in academia and other parts of industry.

However, none of the corporate RECs we spoke with had lay members or external experts on their boards. This raises a serious concern that perspectives of people impacted by these technologies are not reflected in ethical reviews of their research, and that what constitutes a risk or is considered a high-priority risk is left solely to the discretion of employees of the company.

[79] Moss, E. and Metcalf, J. (2020). *Ethics Owners. A New Model of Organizational Responsibility in Data-Driven Technology Companies.* Data & Society. Available at: https://datasociety.net/library/ethics-owners/

The lack of engagement with external experts or people affected by this research may mean that critical or non-obvious information about what constitutes a risk to some members of society may be missed.

Some participants we spoke with also mentioned that corporate labs experience challenges engaging with external stakeholders and experts to consult on critical issues. Many large companies seek to hire this expertise in-house, bringing in interdisciplinary researchers with social science, economics and other backgrounds. However, engaging external experts can be challenging, given concerns around trade secrets, sharing sensitive data and tipping off rival companies about their work.

Many companies resort to asking participants to sign non-disclosure agreements (NDAs), which are legally binding contracts with severe financial sanctions and legal risks if confidential information is disclosed. These can last in perpetuity, and for many external stakeholders (particularly those from civil society or marginalised groups), signing these agreements can be a daunting risk. However, we did hear from other corporate REC members that they had successfully engaged with external experts in some instances to understand the holistic set of concerns around a research project. In one biomedical-based research project, a corporate REC claimed to have engaged over 25 experts in a range of backgrounds to determine potential risks their work might raise and what mitigations were at their disposal.

Ongoing training

Many corporate RECs we spoke with also place an emphasis on continued skills and training, including providing basic 'ethical training' for staff of all levels. One corporate REC member we spoke with noted several lessons learned from their experience running ethical reviews of AI and data science research:

1. **Executive buy-in and sponsorship:** It is essential to have senior leaders in the organisation backing and supporting this work. Having a senior spokesperson also helped in communicating the importance of ethical consideration throughout the organisation.
2. **Culture:** It can be challenging to create a culture where researchers feel incentivised to talk and think about the ethical implications of their work, particularly in the earliest stages. Having a collaborative company culture in which research is shared openly within the

company, and a transparent process where researchers understand what an ethics review will involve, who is reviewing their work, and what will be expected of them can help address this concern. Training programmes for new and existing staff on the importance of ethical reviews and how to think reflexively helped staff level-set with what is expected of them.

3. **Diverse perspectives:** Engaging diverse perspectives can result in more robust decision-making. This means engaging with external experts who represent interdisciplinary backgrounds, and may include hiring that expertise internally. This can also include experiential diversity, which incorporates perspectives of different lived experiences. It also involves considering one's own positionality and biases, and being reflexive as to how one's own biases and lived experiences can influence consideration for ethical issues.

4. **Early and regular engagement leads to more successful outcomes:** Ethical issues can emerge at different stages of a research project's lifecycle, particularly given quick-paced and shifting political and social dynamics outside the lab. Engaging in ethical reviews at the point of publication can be too late, and the earlier this work is engaged with the better. Regular engagement throughout the project lifecycle is the goal, along with post-mortem reviews of the impacts of research.

5. **Continuous learning:** REC processes need to be continuously updated and improved, and it is essential to seek feedback on what is and isn't working.

Other actors in the research ethics ecosystem

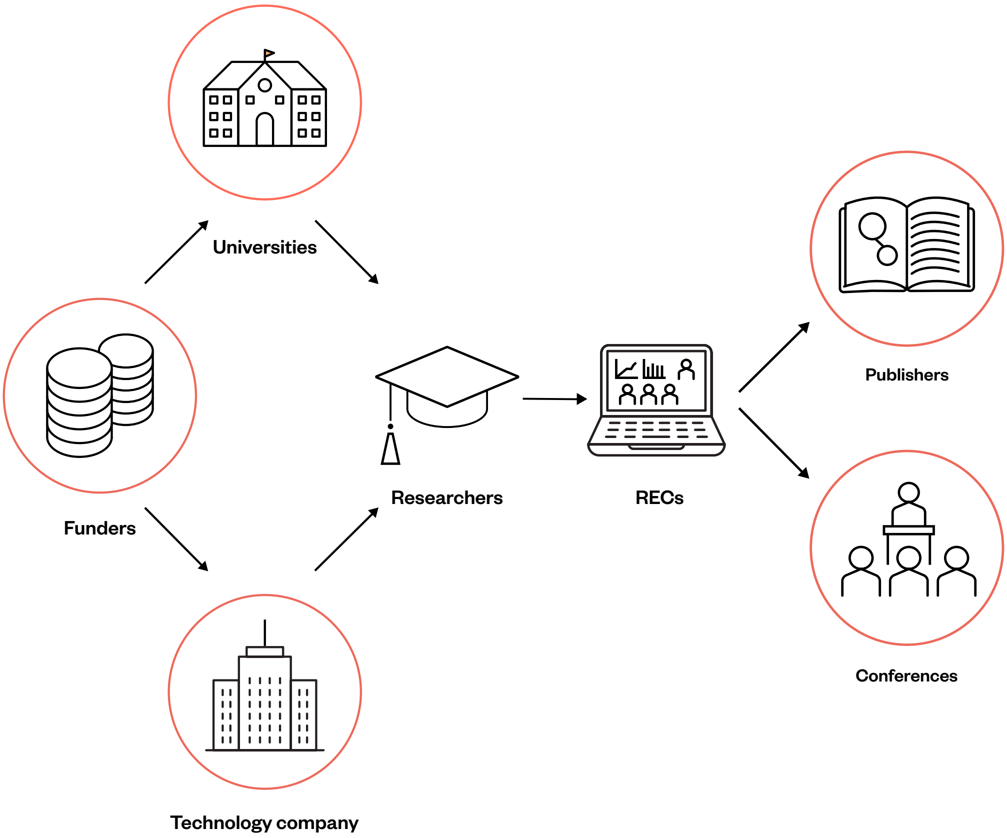
While academic and corporate RECs and researchers share the primary burden for assessing research ethics issues, there are other actors who share this responsibility to varying degrees, including funders, publishers and conference organisers.[80] Along with RECs, these other actors help

[80] Samuel, G., Derrick, G. E., and Van Leeuwen, T. (2019). 'The ethics ecosystem: Personal ethics, network governance and regulating actors governing the use of social media research data.' *Minerva*, 57(3), pp. 317–343. Available at: https://link.springer.com/article/10.1007/s11024-019-09368-3

establish **research culture**, which refers to 'the behaviours, values, expectations, attitudes and norms of research communities'.[81] Research culture influences how research is done, who conducts research and who is rewarded for it.

Creating a healthy research culture is a responsibility shared by research institutions, conference organisers, journal editors, professional associations and other actors in the research ecosystem. This can include creating rewards and incentives for researchers to conduct their work according to a high ethical standard, and to reflect carefully on the broader societal impacts of their work. In this section, we examine in detail only three actors in this complex ecosystem.

Figure 5: Different actors in the research ecosystem



This figure shows some of the different actors that comprise the AI and data science research ecosystem. These actors interact and set incentives for each other. For example, funders can set incentives for

[81] The Royal Society. 'Research Culture'. Available at: https://royalsociety.org/topics-policy/projects/research-culture/

institutions and researchers to follow (such as meeting certain criteria as part of a research application). Similarly, publishers and conferences can set incentives for researchers to follow in order to be published.

Organisers of research conferences can set particular incentives for a healthy research culture. Research conferences are venues where research is rewarded and celebrated, enabling career advancement and growth opportunities. They are also forums where junior and senior researchers from the public and private sectors create professional networks and discuss field-wide benchmarks, milestones and norms of behaviour. As Ada's recent paper with CIFAR on AI and machine learning (ML) conference organisers explores, there are a wide variety of steps that conferences can take to incentivise consideration for research ethics and broader societal impacts.[82]

For example, in 2020, the Conference on Neural Information Processing (NeurIPS) introduced a requirement that submitted papers include a broader societal impact statement of the benefits, limitations and risks of the research.[83] These impact statements were designed to encourage researchers submitting work to the conference to consider the risks their research might raise, and to conduct more interdisciplinary consultation with experts from other domains and engagement with people who may be affected by their research.[84] The introduction of this requirement was hotly contested by some researchers, who were concerned it was an overly burdensome 'tick box' exercise that would become pro-forma over time.[85] In 2021, NeurIPs shifted to adding ethical considerations into a checklist of requirements for submitted papers, rather than requiring a standalone statement for all papers to complete.

[82] Canadian Institute for Advanced Research, Partnership on AI and Ada Lovelace Institute. (2022). *A culture of ethical AI: report*. Available at: https://www.adalovelaceinstitute.org/event/culture-ethical-ai-cifar-pai/

[83] Prunkl, C. E. et al. (2021). 'Institutionalizing ethics in AI through broader impact requirements'. *Nature Machine Intelligence*, 3(2), pp. 104–110. Available at: https://www.nature.com/articles/s42256-021-00298-y

[84] Prunkl et al state that potential negative effects to impact statements are that these could be uninformative, biased, misleading or overly speculative, and therefore lack quality. The statements could lead to trivialising of ethics and governance and the complexity involved in assessing ethical and societal implications. Researchers could develop a negative attitude towards submitting an impact statement, and may find it a burden, confusing or irrelevant. The statements may also create a false sense of security, in cases where positive impacts are overstated or negative impacts understated, which may polarise the research community along political or institutional lines. See: Prunkl, C. E. et al. (2021).

[85] Some authors felt that the requirement of an impact statement is important, but there was uncertainty over who should complete them and how. Other authors also did not feel qualified to address the broader impact of their work. See: Abuhamad, G. and Rheault, C. (2020). 'Like a Researcher Stating Broader Impact For the Very First Time'. *arXiv*. Available at: https://arxiv.org/abs/2011.13032

Editors of academic journals can set incentives for researchers to assess for and mitigate the ethical implications of their work. Having work published in an academic journal is primary goal for most academics, and a pathway for career advancement. Journals often put in place certain requirements for submissions to be accepted. For example, the Committee on Publication Ethics (COPE) has released guidelines on research integrity practices in scholarly publishing, which stipulate that journals should include policies on data sharing, reproducibility and ethical oversight.[86] This includes requirements that studies involving human subjects research must provide self-disclosure that a REC has approved the study.

Some organisations have suggested journal editors could go further towards encouraging researchers to consider questions of broader societal impacts. The Partnership on AI (PAI) published a range of recommendations for responsible publication practice in AI and ML research, which include calls for a change in research culture that normalises the discussion of downstream consequences of AI and ML research.[87]

Specifically for conferences and journals, PAI recommends expanding peer review criteria to include potential downstream consequences by asking submitting researchers to include a broader societal impact statement. Furthermore, PAI recommends establishing a separate review process to evaluate papers based on risk and downstream consequences, a process that may require a unique set of multidisciplinary experts to go beyond the scope of current journal review practices.[88]

Public and private funders (such as research councils) can establish incentives for researchers to engage with questions of research ethics, integrity and broader societal impacts. Funders play a critical role in determining which research proposals will move forward, and what areas of research will be prioritised over others. This presents an opportunity for funders to encourage certain practices, such as requiring that any

[86] Committee on Publication Ethics. (2018). *Principles of Transparency and Best Practices in Scholarly Publishing*. Available at: https://publicationethics.org/files/Principles_of_Transparency_and_Best_Practice_in_Scholarly_Publishingv3_0.pdf

[87] Partnership on AI. (2021). *Managing the Risks of AI Research: Six Recommendations for Responsible Publication*. Available at: https://partnershiponai.org/workstream/publication-norms-for-responsible-ai/

[88] Partnership on AI. (2021).

research that receives funding meets expectations around research integrity, Responsible Research and Innovation and research ethics. For example, Gardner recommends that grant funding and public tendering of AI systems should require a 'Trustworthy AI Statement' from researchers that includes an *ex ante* assessment of how the research will comply with the European HLEG's Trustworthy AI standards.[89]

[89] Gardner, A., Smith, A. L., Steventon, A. et al. (2021). 'Ethical funding for trustworthy AI: proposals to address the responsibilities of funders to ensure that projects adhere to trustworthy AI practice'. *AI and Ethics*. pp.1–15. Available at: https://link.springer.com/article/10.1007/s43681-021-00069-w

Challenges in AI research

In this chapter, we highlight six major challenges that Research Ethics Committees (RECs) face when evaluating AI and data science research, as uncovered during the workshops conducted with members of RECs and researchers in May 2021.

Challenge 1: Many RECs lack the resources, expertise and training to appropriately address the risks that AI and data science pose

Inadequate review requirements

Some workshop participants highlighted that many projects that raise severe privacy and consent issues are not required to undergo research ethics review. For example, some RECs encourage researchers to adopt data minimisation and anonymisation practices and do not require a project to undergo ethics reviews if the data is anonymised after collection. However, research has shown that anonymised data can still be triangulated with other datasets to enable reidentification,[90] raising a privacy risk to data subjects and implications for the consideration of broader impacts.[91] Expert participants noted that it is hard to determine if data collected for a project is anonymous, and that RECs must have the right expertise to fully interrogate whether a research project has adequately addressed these challenges.

As Metcalf and Crawford have noted, data science is usually not considered a form of direct intervention in the body or life of individual human subjects and is, therefore, exempt from many research ethics

[90] Vayena, E., Brownsword, R., Edwards, S. J. et al. (2016). 'Research led by participants: a new social contract for a new kind of research'. *Journal of Medical Ethics*, 42(4), pp. 216–219.

[91] There are three types of disclosure risks and possible reidentification of an individual despite masking or de-identification of data: identity disclosure, attribute disclosure, e.g., when a person is identified to belong to a particular group, or inferential disclosure, e.g., when information about a person can be inferred with released data. See: Xafis, V., Schaefer, G. O., Labude, M. K. et al. (2019). 'An ethics framework for big data in health and research'. *Asian Bioethics Review*, 11(3). Available at: https://doi.org/10.1007/s41649-019-00099-x

review processes.[92] Similar challenges arise with AI research projects that rely on data collected from public sources, such as surveillance cameras or scraped from the public web, which are assumed to pose minimal risk to human subjects. Under most current research ethics guidelines, research projects using publicly available or pre-existing datasets collected and shared by other researchers are also not required to undergo research ethics review.[93]

Some of our workshop participants noted that researchers can view RECs as risk averse and overly concerned with procedural questions and reputation management. This reflects some findings from the literature. Samuel et al found that, while researchers perceive research ethics as procedural and centred on operational governance frameworks, societal ethics are perceived as less formal and more 'fuzzy', noting the absence of standards and regulations governing AI in relation to societal impact.[94]

Expertise and training

Another institutional challenge our workshop participants identified related to the training, composition and expertise of RECs. These concerns are not unique to reviews of AI and data science and reflect long-running concerns with how effectively RECs operate. In the USA, a 2011 study found that university research ethics review processes are perceived by researchers as inefficient, with review outcomes being viewed as inconsistent and often resulting in delays in the research process, particularly for multi-site trials.[95]

Other studies have found that researchers view RECs as overly bureaucratic and risk-averse bodies, and that REC practices and decisions can vary substantially across institutions.[96] These studies have

[92] Metcalf, J. and Crawford, K. (2016). 'Where are human subjects in big data research? The emerging ethics divide'. *Big Data & Society*, 3(1). Available at: https://journals.sagepub.com/doi/full/10.1177/2053951716650211

[93] Metcalf, J. and Crawford, K. (2016).

[94] Samuel, G., Chubb, J. and Derrick, G. (2021). 'Boundaries Between Research Ethics and Ethical Research Use in Artificial Intelligence Health Research'. *Journal of Empirical Research on Human Research Ethics*. Available at: https://journals.sagepub.com/doi/full/10.1177/15562646211002744

[95] Abbott, L. and Grady, C. (2011). 'A systematic review of the empirical literature evaluating IRBs: What we know and what we still need to learn'. *Journal of Empirical Research on Human Research Ethics*, 6(1). Available at: https://doi.org/10.1525/jer.2011.6.1.3

[96] Zywicki, T. J. (2007). 'Institutional review boards as academic bureaucracies: An economic and experiential analysis'. *Northwestern University Law Review*, 101(2), p.861. Available at: https://heinonline.org/HOL/LandingPage?handle=hein.journals/illlr101&div=36&id=&page=

found that that RECs have differing approaches to determining which projects require a full rather than expedited review, and often do not provide a justification or explanation for their assessments of the risk of certain research practices.[97] In some documented cases, researchers have gone so far as to abandon projects due to delays and inefficiencies of research ethics review processes.[98]

There is some evidence these issues are exacerbated in reviews of AI and data science research. Dove et al found systemic inefficiencies and substantive weaknesses in research ethics review processes, including:

- a lack of expertise in understanding the novel challenges emerging from data-intensive research
- a lack of consistency and reasoned decision-making of RECs
- a focus on 'tick-box exercises'
- duplication of ethics reviews
- a lack of communication between RECs in multiple jurisdictions.[99]

One reason for variation in ethics review process outcomes is disagreement among REC members. This can be the case even when working with shared guidelines. For example, in the context of data acquired through social media for research purposes, REC members differ substantially in their assessment of whether consent is required, as well as the risks to research participants. In part, this difference of opinion can be linked to their level of experience in dealing with these issues.[100] Some researchers suggest that reviewers may benefit from more training and support resources on emerging research ethics issues, to ensure a more consistent approach to decision-making.[101]

[97] Abbott, L. and Grady, C. (2011). 'A systematic review of the empirical literature evaluating IRBs: What we know and what we still need to learn'. *Journal of Empirical Research on Human Research Ethics*, 6(1). Available at: https://doi.org/10.1525/jer.2011.6.1.3

[98] Abbott, L. and Grady, C. (2011).

[99] Dove, E. S. and Garattini, C. (2018). 'Expert perspectives on ethics review of international data-intensive research: Working towards mutual recognition'. *Research Ethics*, 14(1), pp. 1–25. Available at: https://journals.sagepub.com/doi/full/10.1177/1747016117711972

[100] Hibbin, R. A., Samuel, G. and Derrick, G. E. (2018). 'From "a fair game" to "a form of covert research": Research ethics committee members' differing notions of consent and potential risk to participants within social media research'. *Journal of Empirical Research on Human Research Ethics*, 13(2). Available at: https://journals.sagepub.com/doi/full/10.1177/1556264617751510

[101] Guillemin, M., Gillam, L., Rosenthal, D. and Bolitho, A. (2012). 'Human research ethics committees: examining their roles and practices'. *Journal of Empirical Research on Human Research Ethics*, 7(3). Available at: https://journals.sagepub.com/doi/abs/10.1525/jer.2012.7.3.38

A significant challenge arises from the lack of training – and, therefore, lack of expertise – of REC members.[102] While this has already been identified as a persistent issue with RECs generally,[103] AI and data science research can be applied to many disciplines. This means that REC members evaluating AI and data science research must have expertise across many fields. However, many RECs in this space frequently lack expertise across both (i) technical methods of AI and data science, and (ii) domain expertise from other relevant disciplines.[104]

Samuel et al found that some RECs that review AI and data science research are concerned with data governance issues, such as data privacy, which is perceived as not requiring AI-specific technical skills.[105] While RECs regularly draw on specialist advice through cross-departmental collaboration, workshop participants questioned whether resources to support examination of ethical issues relating to AI and data science research are made available for RECs.[106] RECs may need to consider which appropriate expertise is required for these reviews and how it will be sourced, for instance, via specialist ad-hoc advice, or the institution of sub-committees.[107]

The need for reviewers with expertise across disciplines, ethical expertise and cross-departmental collaboration is clear. Participants in our workshops questioned whether interdisciplinary expertise is sufficient to review AI and data science research projects, and whether experiential expertise (expertise on the subject matter gained through first-person involvement) is also necessary to provide a more holistic assessment of potential research risks. This could take the form of changing a REC's composition to involve a broader range of stakeholders, such as community representatives or external organisations.

[102] Ferretti, A., Ienca, M., Sheehan, M. et al. (2021). 'Ethics review of big data research: What should stay and what should be reformed?'. *BMC Medical Ethics*, 22(1), pp. 1–13. Available at: https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-021-00616-4

[103] Guillemin, M., Gillam, L., Rosenthal, D. and Bolitho, A. (2012). 'Human research ethics committees: examining their roles and practices'. *Journal of Empirical Research on Human Research Ethics*, 7(3). Available at: https://journals.sagepub.com/doi/abs/10.1525/jer.2012.7.3.38

[104] Yuan, H., Vanea, C., Lucivero, F. and Hallowell, N. (2020). 'Training Ethically Responsible AI Researchers: a Case Study'. *arXiv*. Available at: https://arxiv.org/abs/2011.11393

[105] Samuel, G., Chubb, J. and Derrick, G. (2021). 'Boundaries Between Research Ethics and Ethical Research Use in Artificial Intelligence Health Research'. *Journal of Empirical Research on Human Research Ethics*. Available at: Available at: https://journals.sagepub.com/doi/full/10.1177/15562646211002744

[106] Rawbone, R. (2010). 'Inequality amongst RECs'. *Research Ethics Review*, 6(1), pp. 1–2. Available at: https://journals.sagepub.com/doi/pdf/10.1177/174701611000600101

[107] Hine, C. (2021). 'Evaluating the prospects for university-based ethical governance in artificial intelligence and data-driven innovation'. *Research Ethics*. Available at: https://journals.sagepub.com/doi/full/10.1177/17470161211022790

Resources

A final challenge that RECs face relates to their resourcing and the value given to their work. According to our workshop participants, RECs are generally under-resourced in terms of budget, staffing and rewarding of members. Many RECs rely on voluntary 'pro bono' labour of professors and other staff, with members managing competing commitments and an expanding volume of applications for ethics review.[108] Inadequate resources can result in further delays and have a negative impact on the quality of the reviews. Chadwick shows that RECs rely on the dedication of their members, who prioritise the research subjects, researchers, REC members and the institution ahead of personal gain.[109]

Several of our workshop participants noted reviewers do not have enough time to do a proper ethics review that evaluates the full range of potential ethical issues, or the right range of skills. According to several participants, sitting on a REC is often a 'thankless' task, which can make finding people willing to serve difficult. Those who are willing and have the required expertise risk being overloaded. Reviewing is 'free labour' with little or no recognition, and the question arises how to incentivise REC members. It was discussed that research ethics review should be budgeted appropriately to engage with stakeholders throughout the project lifecycle.

Challenge 2: Traditional research ethics principles are not well suited for AI research

In their evaluations of AI and data science research, RECs have traditionally relied on a set of legally mandated and self-regulatory ethics principles that largely stem from the biomedical sciences. These principles have shaped the way that modern research ethics is understood at research institutions, how RECs are constructed and the traditional scope of their remit.

[108] Page, S. A. and Nyeboer, J. (2017). 'Improving the process of research ethics review'. *Research integrity and peer review*, 2(1), pp. 1–7. Available at: https://researchintegrityjournal.biomedcentral.com/articles/10.1186/s41073-017-0038-7

[109] Chadwick, G. L. and Dunn, C. M. (2000). 'Institutional review boards: changing with the times?'. *Journal of public health management and practice*, 6(6), pp. 19–27. Available at: https://europepmc.org/article/med/18019957

Contemporary RECs draw on a long list of additional resources for AI and data science research in their reviews, including data science-specific guidelines like the Association of Internet Researchers' ethical guidelines,[110] provisions of the EU General Data Protection Regulation (GDPR) to govern data protection issues, and increasingly the emerging field of 'AI ethics' principles. However, the application of these principles raises significant challenges for RECs.

Several of our expert participants noted these guidelines and principles are often not implemented consistently across different countries, scientific disciplines, or across different departments or teams within the same institution.[111] As prominent research guidelines were originally developed in the context of biomedical research (see page 26), questions have been raised about their applicability to other disciplines, such as the social sciences, data science and computer science.[112] For example, some in the research community have questioned the extension of the Belmont principles to research in non-experimental settings due to differences in methodologies, the relationships between researchers and research subjects, different models and expectations of consent and different considerations for what constitutes potential harm and to whom.[113]

We draw attention to four main challenges in the application of traditional bioethics principles to ethics reviews of AI and data science research:

Autonomy, privacy and consent

One example of how biomedical principles can be poorly applied to AI and data science research relates to how they address questions of autonomy and consent. Many of these principles emphasise that 'voluntary consent of the human subject is absolutely essential' and

[110] Association of Internet Researchers. (2020). *Internet Research: Ethical Guidelines 3.0*. Available at: https://aoir.org/reports/ethics3.pdf

[111] Emanuel, E. J., Grady, C. C., Crouch, R. A., Lie, R. K., Miller, F. G. and Wendler, D. D. (eds.). (2008). *The Oxford textbook of clinical research ethics*. Oxford University Press.

[112] Oakes, J. M. (2002). 'Risks and wrongs in social science research: An evaluator's guide to the IRB'. *Evaluation Review*, 26(5), pp. 443–479. Available at: https://journals.sagepub.com/doi/10.1177/019384102236520?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed; and Dyer, S. and Demeritt, D. (2009). 'Un-ethical review? Why it is wrong to apply the medical model of research governance to human geography'. *Progress in Human Geography*, 33(1), pp. 46–64. Available at: https://journals.sagepub.com/doi/10.1177/0309132508090475

[113] Cannella, G. S. and Lincoln, Y. S. (2011). 'Ethics, research regulations, and critical social science'. *The Sage handbook of qualitative research*, 4, pp. 81–90; and Israel, M. (2014). *Research ethics and integrity for social scientists: Beyond regulatory compliance*. SAGE Publishing.

should outweigh considerations for the potential societal benefit of the research.

Workshop participants highlighted consent and privacy issues as one of the most significant challenges RECs are currently facing in reviews of AI and data science research. This included questions about how to implement 'ongoing consent', whereby consent is given at various stages of the research process; whether informed consent may be considered forced consent when research subjects do not really understand the implications of the future use of their data; and whether it is practical to require consent be given more than once when working with large-scale data repositories. A primary concern flagged by workshop participants was whether RECs put too much weight on questions of consent and autonomy at the expense of wider ethical concerns.

Issues of consent largely stem from the ways these fields collect and use personal data,[114] which differs substantially from the traditional clinical experiment format. Part of the issue is the relatively distanced relationship between data scientist and research subject. Here, researchers can rely on data scraped from the web – such as social media posts; or collected via consumer devices – such as fitness trackers or smart speakers.[115] Once collected, many of these datasets can be made publicly accessible as 'benchmark datasets' for other researchers to test and train their models. The Flickr Faces HQ dataset, for example, contains 70,000 images of faces collected from a photo-sharing website and made publicly accessible with a Creative Commons license for other researchers to use.[116]

These collection and sharing practices pose novel risks to the privacy and identifiability of research subjects, and challenge traditional notions of informed consent from participants.[117] Once collected and shared,

[114] The ICO defines personal data as 'information relating to natural persons who can be identified or who are identifiable, directly from the information in question; or who can be indirectly identified from that information in combination with other information.' See: Information Commissioners Office. *Guide to the UK General Data Protection Regulation (UK GDPR) – What is Personal Data?* Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/what-is-personal-data/

[115] Friesen, P., Douglas-Jones, R., Marks, M. et al. (2021). 'Governing AI-Driven Health Research: Are IRBs Up to the Task?' *Ethics & Human Research*, 43(2), pp. 35–42. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/eahr.500085

[116] Karras, T., Laine, S. and Aila, T. (2019). 'A style-based generator architecture for generative adversarial networks'. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410.

[117] Ferretti, A., Ienca, M., Sheehan, M. et al. (2021). 'Ethics review of big data research: What should stay and what should be reformed?'. *BMC Medical Ethics*, 22(1), pp. 1–13. Available at: https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-021-00616-4

datasets may be re-used or re-shared for different purposes than those understood during the original consent process. It is often not feasible for researchers re-using the data to obtain informed consent in relation to the original research. In many cases, informed consent may not have been given in the first place.[118]

Not being able to obtain informed consent does not give the researcher a blank slate, and datasets that are continuously used as a benchmark for technology development risk normalising the avoidance of consent-seeking practices. Some benchmark datasets, such as the longitudinal Pima Indian Diabetes Dataset (PIDD), are tied to a colonial past of oppression and exploitation of indigenous peoples, and its use as a benchmark dataset perpetuates these politics in new forms.[119] The challenges to informed consent can cause significant damage to public trust in institutions and science. One notable example involved a Facebook (now Meta) study in 2014, in which researchers were able to monitor users' emotional states and manipulated their news feed without their consent, showing more negative content to some users.[120] The study led to significant public concern, and raised questions about how Facebook users could give informed consent in instances where they lack control, let alone awareness of the study.

In some instances, AI and data science research may also pose novel privacy risks relating to the kinds of inferences that can be drawn from data. To take one example, researchers at Facebook (now Meta) developed an AI system to identify suicidal intent in user-generated content, which could be shared with law enforcement agencies to conduct wellness checks on identified users.[121] This kind of 'emergent' health data produced through interactions with software platforms or products is not subject to the same requirements or regulatory oversight as data from a mental health professional.[122] This highlights how an AI

[118] Ferretti, A., Ienca, M., Sheehan, M. et al. (2021).

[119] Radin, J. (2017). '"Digital Natives": How Medical and Indigenous Histories Matter for Big Data'. *Osiris*, 32, pp. 43–64. Available at: https://doi.org/10.1086/693853

[120] Kramer, A. D., Guillory, J. E. and Hancock, J. T. (2014). 'Experimental evidence of massive-scale emotional contagion through social networks'. *Proceedings of the National Academy of Sciences*, 111(24), pp. 8788–8790. Available at: https://www.pnas.org/doi/abs/10.1073/pnas.1320040111; and Selinger, E. and Hartzog, W. (2016). 'Facebook's emotional contagion study and the ethical problem of co-opted identity in mediated environments where users lack control'. *Research Ethics*, 12(1), pp. 35–43.

[121] Marks, M. (2020). 'Emergent medical data: Health Information inferred by artificial intelligence'. *UC Irvine Law Review*, 995. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3554118

[122] Marks, M. (2020).

system can infer sensitive health information about an individual based on non-health related data in the public domain, which could pose severe risks for the privacy of vulnerable and marginalised communities.

Questions of consent and privacy point to another tension between principles of research integrity and the ethical obligations towards protecting research participants from harm. In the spirit of making research reproducible, there is a growing acceptance among the AI and data science research community that scientific data should be openly shared, and that open access policies for data and code should be fostered so that other researchers can easily re-use research outputs. At the same time, it is not possible to make data accessible to everyone, as this can lead to harmful misuses of the data by other parties, or uses of that data that are for a purpose the data subject would not be comfortable with. Participants largely agreed, however, that RECs struggle to assess these types of research projects because the existing *ex ante* model of RECs addresses potential risks up front and may not be fit to address the potential emerging risks for data subjects.[123]

Risks to research subjects vs societal benefit

A related topic to consent is the challenge of weighing the societal benefit of research against the risks to the research subjects it poses.

Workshop participants acknowledged how AI and data science research create a different researcher-subject relationship from traditional biomedical research. For example, participants noted that research in a clinical context involves a person who is present and with whom researchers have close and personal interaction. A researcher in these contexts is identifiable to their subject, and vice versa. This relationship often does not exist in AI and data science research, where the ‘subject’ of research may not be readily identifiable or may be someone affected by research rather than someone participating in the research. Some research argues that AI and data science research marks a shift from ‘human subjects’ research to ‘data subjects’ research, in which care and concern for the welfare of participants

[123] Ferretti, A., Ienca, M., Sheehan, M. et al. (2021). ‘Ethics review of big data research: What should stay and what should be reformed?’. BMC Medical Ethics, 22(1), pp. 1–13. Available at: https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-021-00616-4

should be given to those whose data is used.[124]

In many cases, data science and AI research projects rely on data sourced from the web through scraping, a process that challenges traditional notions of informed consent and raises questions about whether researchers are in a position to assess the risk of research to participants.[125] Researchers may not be able to identify the people whose data they are collecting, meaning they often lack a relational dynamic that is essential for understanding the needs, interests and risks of their research subjects. In other cases, AI researchers may use publicly available datasets made available on online repositories like GitHub, and which may be repurposed for reasons that differ from their originally intended basis for collection. Finally, major differences arise with how data is analysed and assessed. Many kinds of AI and data science research rely on the curation of massive volumes of data, a process that many researchers outsource to third-party contract services such as Amazon's MTurk. These processes create further separation between researchers and research subjects, outsourcing important value-laden decisions about the data to third-party workers who are not identifiable, accountable or known to research subjects.

Responsibility for assessing risks and benefit

Another challenge research ethics principles have sought to address is determining who is responsible for assessing and communicating the risk of research to participants.

One criticism has been that biomedical research ethics frameworks do not reflect the 'emergent, dynamic and interactional nature'[126] of fields like the social sciences and humanities.[127] For example, ethnographic or anthropological research methods are open-ended, emergent and need to be responsive to the concerns of research participants throughout the research process. Meanwhile, traditional REC reviews have been solely

[124] Samuel, G., Ahmed, W., Kara, H. et al. (2018). 'Is It Time to Re-Evaluate the Ethics Governance of Social Media Research?'. *Journal of Empirical Research on Human Research Ethics*, 13(4), pp. 452–454. Available at: https://www.jstor.org/stable/26973881

[125] Taylor, J. and Pagliari, C. (2018). 'Mining Social Media Data: How are Research Sponsors and Researchers Addressing the Ethical Challenges?'. *Research Ethics*, 14(2). Available at: https://journals.sagepub.com/doi/10.1177/1747016117738559

[126] Iphofen, R. and Tolich, M. (2018). 'Foundational issues in qualitative research ethics'. *The Sage handbook of qualitative research ethics*, pp. 1–18. Available at: https://methods.sagepub.com/book/the-sage-handbook-of-qualitative-research-ethics-srm/i211.xml

[127] Schrag, Z. M. (2011). 'The case against ethics review in the social sciences'. *Research Ethics*, 7(4), pp. 120–131.

concerned with an up-front risk assessment. In our expert workshops, several participants noted a similar concern within AI and data science research, where risks or benefits cannot be comprehensively assessed in the early stages of research.

Universality of principles

Some biomedical research ethics initiatives have sought to formulate universal principles for research ethics in different jurisdictions, which would help ensure a common standard of review in international research partnerships or multi-site research studies. However, many of these initiatives were created by institutions from predominantly Western countries to respond to Western biomedical research practices, and critics have pointed out that they therefore reflect a deeply Western set of ethics.[128] Other efforts have been undertaken to develop universal principles, including the Emanuel, Wendler and Grady framework, which uses eight principles with associated 'benchmark' questions to help RECs from different regions evaluate potential ethical issues relating to exploitation.[129] While there is some evidence that this model has worked well in REC evaluations for biomedical research in African institutions,[130] it has not yet been widely adopted by RECs in other regions.

Challenge 3: Specific principles for AI and data science research are still emerging and are not consistently adopted by RECs

A more recent phenomenon relevant to the consideration of ethical issues relating to AI and data science has been the proliferation of ethical principles, standards and frameworks for the development and use

[128] Goodyear, M. et al. (2007). 'The Declaration of Helsinki. Mosaic tablet, dynamic document or dinosaur?'. *British Medical Journal*, 335; and Ashcroft, R. E. (2008). 'The declaration of Helsinki'. *The Oxford textbook of clinical research ethics*, pp. 141–148.

[129] Emanuel, E.J., Wendler, D. and Grady, C. (2008) 'An Ethical Framework for Biomedical Research'. *The Oxford Textbook of Clinical Research Ethics*, pp. 123–135.

[130] Tsoka-Gwegweni, J. M. and Wassenaar, D.R. (2014). 'Using the Emanuel et al. Framework to Assess Ethical Issues Raised by a Biomedical Research Ethics Committee in South Africa'. *Journal of Empirical Research on Human Research Ethics*, 9(5), pp. 36–45. Available at: https://journals.sagepub.com/doi/10.1177/1556264614553172?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed

of AI systems.[131, 132, 133, 134] The development of standards for ethical AI systems has been taken up by bodies such as the Institute of Electrical and Electronics Engineers (IEEE) and the International Organization for Standardization (ISO).[135] Some of these efforts have occurred at the international level, such as the OECD or United Nations. A number of principles can be found across this spectrum, including transparency, fairness, privacy and accountability. However, these common principles have variations in how they are defined, understood and scoped, meaning there is no single codified approach to how they should be interpreted.[136]

In developing such frameworks, some have departed from widely adopted guidelines. For example, Floridi and Cowls propose a framework of five overarching principles for AI. This includes the traditional bioethics principles of beneficence, non-maleficence, autonomy and justice, drawn from the Belmont principles, but adds the principle of explicability, which combines questions of intelligibility (how something works) with accountability (who is responsible for the way it works).[137] Others have argued that international human rights frameworks offer a promising basis to develop coherent and universally recognised standards for AI ethics.[138]

Several of our workshop participants mentioned that it is challenging to judge the relevance of existing principles in the context of AI and data science research. During the workshops, a variety of additional

[131] Hagendorff, T. (2020). 'The ethics of AI ethics: An evaluation of guidelines'. *Minds and Machines*, 30(1), pp. 99–120. Available at: https://link.springer.com/article/10.1007/s11023-020-09517-8;

[132] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. and Srikumar, M. (2020). 'Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI'. *Berkman Klein Center Research Publication* No. 2020-1. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482

[133] Gardner, A., Smith, A. L., Steventon, A. et al. (2021). 'Ethical funding for trustworthy AI: proposals to address the responsibilities of funders to ensure that projects adhere to trustworthy AI practice'. *AI and Ethics*, pp. 1–15. Available at: https://link.springer.com/article/10.1007/s43681-021-00069-w

[134] Floridi, L. and Cowls, J. (2019). 'A unified framework of five principles for AI in society'. *Social Science Research Network*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3831321

[135] These include standards initiatives like the IEEE's P7000 series on ethical design of AI systems, which include *P7001 – Standard for Transparency of Autonomous Systems* (2021), *P7003 – Algorithmic Bias Considerations* (2018) and *P7010 – Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems* (2020). *ISO/IEC JTC 1/SC 42 – Artificial Intelligence* takes on a series of related standards around data management, trustworthiness of AI systems and transparency.

[136] Jobin, A., Ienca, M. and Vayena, E. (2019). 'The global landscape of AI ethics guidelines'. *Nature Machine Intelligence*, 1, pp. 389–399. Available at: https://doi.org/10.1038/s42256-019-0088-2

[137] Floridi, L. and Cowls, J. (2019). 'A unified framework of five principles for AI in society'. *Social Science Research Network*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3831321

[138] Yeung, K., Howes, A. and Pogrebna, G. (2019). 'AI governance by human rights-centred design, deliberation and oversight: An end to ethics washing'. *The Oxford Handbook of AI Ethics*. Oxford University Press.

principles were mentioned, for example, 'equality', 'human-centricity', 'transparency' and 'environmental sustainability'. This indicates that there is not yet clear consensus around which principles should guide AI and data science research practices, and that the question of how those principles should be developed (and by which body) is not yet answered. We address this challenge in our recommendations (page 72).

The wide range of available frameworks, principles and guidelines demonstrate the difficulty for researchers and practitioners to select suitable frameworks or principles due to the current inconsistencies and a lack of a commonly accepted framework or principles guiding ethical AI and data science research. As many of our expert participants noted, this has led to confusion among RECs about whether these frameworks or principles should supplement biomedical principles, and how they should apply them to reviews of data science and AI research projects.

Complicating this challenge is the question of whether ethical principles guiding AI and data science research would be useful in practice. In a paper comparing the fields of medical ethics with AI ethics, Mittelstadt argues that AI research and development lacks several essential features for developing coherent research ethics principles and practices. These include the lack of common aims and fiduciary duties, a history of professional norms and bodies to translate principles into practice, and robust legal and professional accountability mechanisms.[139] While medical ethics draws on its practitioners being part of a 'moral community' characterised by common aims, values and training, AI cannot refer to such established norms and practices, given the wide range of disciplines and commercial fields it can be applied to.

The blurring of commercial and societal motives for AI research can cause AI developers to be driven by values such as innovation and novelty, performance or efficiency, rather than ethical aims rooted in biomedicine around concern for their 'patient' or for societal benefit. In some regions, like Canada, professional codes of practice and law around medicine have established fiduciary-like duties between doctors and their patients, which do not exist in the fields of AI and data

[139] Mittelstadt, B. (2019). 'Principles alone cannot guarantee ethical AI'. *Nature Machine Intelligence*, 1(11), pp. 501–507. Available at: https://www.nature.com/articles/s42256-019-0114-4

science.[140] AI does not have a history and professional culture around ethics comparable to the medical field, which has a strong regulating influence on practitioners. Some research has also questioned the aims of AI research, and what kinds of practices are incentivised and encouraged within the research community. A study involving interviews with 53 AI practitioners in India, East and West African countries, and the USA showed that, despite the importance of high-quality data in addressing potential harms, and a proliferation of data ethics principles, practitioners find the implementation of these practices to be one of the most undervalued and 'de-glamorised' aspects of developing AI systems.[141]

Identifying clear principles for AI research ethics is a major challenge. This is particularly the case because so few of the emerging AI ethics principles specifically focus on AI or data science research ethics. Rather, they centre on the ethics of AI system development and use. In 2019, the IEEE published a report entitled *Ethically aligned design: Prioritizing human wellbeing with autonomous and intelligent systems*, which contains a chapter on 'Methods to Guide Ethical Research and Design'.[142] This chapter includes a range of recommendations for academic and corporate research institutions, including that: labs should identify stages in their processes in which ethical considerations, or 'ethics filters', are in place before products are further developed and deployed; and that interdisciplinary ethics training should be a core subject for everyone working in the STEM field, and should be incentivised by funders, conferences and other actors. However, this report stops short of offering clear guidance for RECs and institutions on how they should turn AI ethics principles into clear practical guidelines for conducting and assessing AI research.

Several of our expert participants observed that many AI researchers and RECs currently draw on legal guidance and norms relating to privacy and data protection, which can risk conflating questions of AI ethics into narrower issues of data governance. The rollout of the European General Data Protection Regulation (GDPR) in 2018

[140] Mittelstadt, B. (2019).

[141] Sambasivan, N., Kapania, S., Highfill, H. et al. (2021). '"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI'. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. Available at: https://research.google/pubs/pub49953/

[142] IEEE Standards Association. (2019). *Ethically Aligned Design*, First Edition. Available at: https://ethicsinaction.ieee.org/#ead1e

created a strong incentive for European institutions and institutions working with personal data of Europeans to reinforce existing ethics requirements on how research data is collected, stored and used by researchers. Expert participants noted that data protection questions are common on most REC reviews. As Samuel notes, there is some evidence that AI researchers tend to perceive research ethics as data governance questions, a mindset of thinking that is reinforced by institutional RECs in some of the questions they ask.[143]

There have been some grassroots efforts to standardise research ethics principles and guidance for some forms of data science research, including social media research. The Association of Internet Researchers, for example, has published its third edition of ethical guidelines,[144] which includes suggestions for how to deal with privacy and consent issues posed by scraping online data, how to outline and address questions across different stages of the ethics lifecycle (such as considering issues of bias and in the data analysis stage), and considering issues of potential downstream harms with the use of that data. However, these guidelines are voluntary and are narrowly focused on social media research. It remains unclear whether RECs are consistently enforcing them. As Samuel notes, the lack of established norms and criteria in social media research has caused many researchers to rely on bottom-up, personal 'ethical barometers' that create discrepancies in how ethical research should be conducted.[145]

In summary, there are a wide range of broad AI ethics principles that seek to guide how AI technologies are developed and deployed. The iterative nature of AI research, in which a published model or dataset can be used by downstream developers to create a commercial product with unforeseen consequences, raises a significant challenge for RECs seeking to apply AI and data science research ethics principles. As many of our expert participants noted, AI ethics research principles must touch on both how research is conducted (including what

[143] Samuel, G., Diedericks, H. and Derrick, G. (2021). Population health AI researchers' perceptions of the public portrayal of AI: A pilot study'. *Public Understanding of Science*, 30(2), pp. 196–211. Available at: https://journals.sagepub.com/doi/full/10.1177/0963662520965490

[144] Association of Internet Researchers. (2020). *Internet Research: Ethical Guidelines 3.0*. Available at: https://aoir.org/reports/ethics3.pdf

[145] Samuel, G., Derrick, G. E. and Van Leeuwen, T. (2019). 'The ethics ecosystem: Personal ethics, network governance and regulating actors governing the use of social media research data'. *Minerva*, 57(3), pp. 317–343. Available at: https://link.springer.com/article/10.1007/s11024-019-09368-3

methodological choices are made), and also involve consideration for the wider societal impact of that research and *how* it will be used by downstream developers.

Challenge 4: Multi-site or public-private partnerships can exacerbate existing challenges of governance and consistency of decision-making

RECs face governance and fragmentation challenges in their decision-making. In contrast to clinical research, which is coordinated in the UK by the Health Research Authority (HRA), RECs evaluating AI and data science research are generally not guided by an overarching governing body, and do not have structures to coordinate similar issues between different RECs. Consequently, their processes, decision-making and outcomes can vary substantially.[146]

Expert participants noted this lack of consistent guidance between RECs is exacerbated by research partnerships with international institutions and public-private research partnerships. The specific processes RECs follow can vary between committees, even within the same institution. This can result in different RECs reaching different conclusions on similar types of research. A 2011 survey of research into Institutional Review Board (IRB) decisions found numerous instances where similar research projects received significantly different decisions, with some RECs approving with no restrictions, others requiring substantial restrictions and others rejecting research outright.[147]

This lack of an overarching coordinating body for RECs is especially problematic for international projects that involve researchers working in teams across multiple jurisdictions, often with large datasets that have multiple sources across multiple sites.[148] Most

[146] Vadeboncoeur, C., Townsend, N., Foster, C. ,and Sheehan, M. (2016). 'Variation in university research ethics review: Reflections following an inter-university study in England'. *Research Ethics*, 12(4), pp. 217–233. Available at: https://journals.sagepub.com/doi/full/10.1177/1747016116652650; and Abbott, L. and Grady, C. (2011). 'A systematic review of the empirical literature evaluating IRBs: What we know and what we still need to learn'. *Journal of Empirical Research on Human Research Ethics*, 6(1), pp.3-19. Available at: https://journals.sagepub.com/doi/abs/10.1525/jer.2011.6.1.3

[147] Silberman, G. and Kahn, K. L. (2011). 'Burdens on research imposed by institutional review boards: the state of the evidence and its implications for regulatory reform'. *The Milbank quarterly*, 89(4), pp. 599–627. Available at: https://doi.org/10.1111/j.1468-0009.2011.00644

[148] Dove, E. S. and Garattini, C. (2018). 'Expert perspectives on ethics review of international data-intensive research: Working towards mutual recognition'. *Research Ethics*, 14(1), pp. 1–25. Available at: https://journals.sagepub.com/doi/10.1177/1747016117711972

biomedical research ethics guidelines recommend that multi-site research should be evaluated by RECs located in all respective jurisdictions,[149] on the basis that each institution will reflect the local regulatory requirements for REC review, which they are best prepared to respond to.

Historically, most research in the life sciences was conducted with a few participants at a local research institution.[150] In some regions, requirements for local involvement have developed to provide some accountability for research subjects. Canada, for example, requires social science research involving indigenous populations to meet specific research ethics requirements, including around community engagement and involvement with members of indigenous communities, and around requirements for indigenous communities to own any data.[151]

However, this arrangement does not fit the large-scale, international, data-intensive research of AI and data science, which often relies on the generation, scraping and repurposing of large datasets, often without any awareness of who exactly the data may be from or under what purpose it was collected. The fragmented landscape of different RECs and regulatory environments leads to multiple research ethics applications to different RECs with inconsistent outcomes, which can be highly resource intensive.[152] Workshop participants highlighted how ethics committees face uncertainties in dealing with data sourced and/or processed in heterogeneous jurisdictions, where legal requirements and ethical norms can be very different.

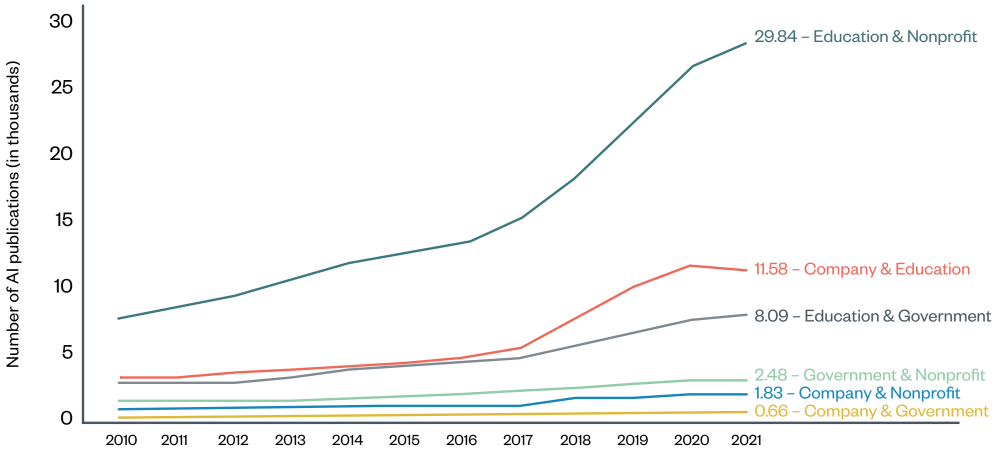
[149] Coleman, C. H., Ardiot, C., Blesson, S.et al . (2015). 'Improving the Quality of Host Country Ethical Oversight of International Research: The Use of a Collaborative 'Pre-Review'Mechanism for a Study of Fexinidazole for Human African Trypanosomiasis'. *Developing World Bioethics*, 15(3), pp. 241–247. Available at: https://onlinelibrary.wiley.com/doi/full/10.1111/dewb.12068

[150] Dove, E. S. and Garattini, C. (2018). 'Expert perspectives on ethics review of international data-intensive research: Working towards mutual recognition'. *Research Ethics*, 14(1), pp. 1–25. Available at: https://journals.sagepub.com/doi/10.1177/1747016117711972

[151] Government of Canada. (2018). *Tri-Council Policy Statement Ethical Conduct for Research Involving Humans, Chapter 9: Research Involving the First Nations, Inuit and Métis Peoples of Canada*. Available at: https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2018.html

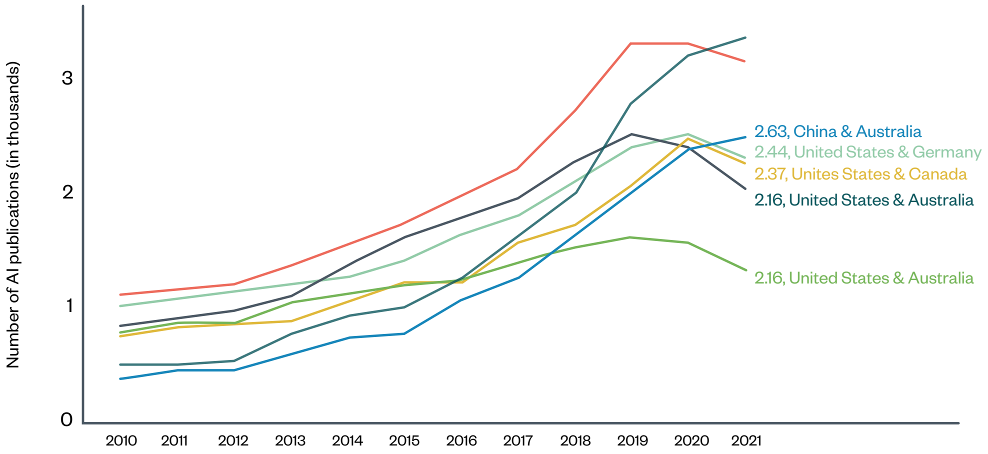
[152] Dove, E. S. and Garattini, C. (2018). 'Expert perspectives on ethics review of international data-intensive research: Working towards mutual recognition'. *Research Ethics*, 14(1), pp. 1–25. Available at: https://journals.sagepub.com/doi/10.1177/1747016117711972

Figure 6: Public-private partnerships in AI research[153]



Cross-sector collaborations in AI publications, 2020–21

Source: Center for Security and Emerging Technology, 2021 | Chart: AI Index Report



Cross-country collaborations in AI publications (excluding US and China), 2020–21

Source: Center for Security and Emerging Technology, 2021 | Chart: AI Index Report

The graphs above show an increasing trend in public-private partnerships in AI research, and in multinational collaborations on AI research. With increasing public-private partnerships and multi-site research, this can increase the challenges for these kinds of research.

[153] Source: Zhang, D. et al. (2022) 'The AI Index 2022 Annual Report'. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2205.03468

Public-private partnerships

Public-private partnerships (PPPs) are common in biomedical research, where partners from the public and private sector share, analyse and use data.[154] The type of collaborations can vary, from project-specific collaborations to long-term strategic alliances between different groups, or large multi-consortia. The data ecosystem is fragmented and complex, as health data is increasingly being shared, linked, re-used or re-purposed in novel ways.[155] Some regulations, such as the General Data Protection Regulation (GDPR) may apply to all research; however, standards, drivers or reputational concerns may differ between actors in the public and private sector. This means that PPPs navigate an equally complex and fragmented landscape of standards, norms and regulations.[156]

As our expert participants noted, public-private partnerships can raise concerns about who derives benefit from the research, who controls the intellectual property of findings, and how data is shared in a responsible and rights-respecting way. The issue of data sharing is particularly problematic when research is used for the purpose of commercial product or service development. For example, wearable devices or apps that track health and fitness data can produce enormous amounts of biomedical 'big data' when combined with other biomedical datasets.[157] While the data generated by these consumer devices can be beneficial for society, through opportunities to advance clinical research in, for instance, chronic illness, consumers of these services may not be aware of these subsequent uses, and their expectations of personal and informational privacy may be violated.[158]

These kinds of violations can have devastating consequences. One can take the recent example of the General Practice Data for Planning and Research (GPDPR), a proposal by England's National Health Service to create a centralised database of pseudonymised patient data that could

[154] Ballantyne, A. and Stewart, C. (2019). 'Big data and public-private partnerships in healthcare and research.' *Asian Bioethics Review*, 11(3), pp. 315–326. Available at: https://link.springer.com/article/10.1007/s41649-019-00100-7

[155] Ballantyne, A. and Stewart, C. (2019).

[156] Ballantyne, A. and Stewart, C. (2019). 'Big data and public-private partnerships in healthcare and research.' *Asian Bioethics Review*, 11(3), pp. 315–326. Available at: https://link.springer.com/article/10.1007/s41649-019-00100-7

[157] Mittelstadt, B. and Floridi, L. (2016). 'The ethics of big data: Current and foreseeable issues in biomedical contexts'. *Science and Engineering Ethics*, 22(2), pp. 303–341. Available at: https://link.springer.com/article/10.1007/s11948-015-9652-2

[158] Mittelstadt, B. (2017). 'Ethics of the health-related internet of things: a narrative review'. *Ethics and Information Technology*, 19, pp. 157–175. Available at: https://doi.org/10.1007/s10676-017-9426-4

be made accessible for researchers and commercial partners.[159] The plan was criticised for failing to alert patients about the use of this data, leading to millions of patients in England opting out of their patient data being accessible for research purposes. As of this publication date, the UK Government has postponed the plan.

Expert participants highlighted that data sharing must be conducted responsibly, aligning with the values and expectations of affected communities, a similar view held by bodies like the UK's Centre for Data Ethics and Innovation.[160] However, what these values and expectations are, and how to avoid making unwarranted assumptions, is less clear. Recent research suggests that participatory approaches to data stewardship may increase legitimacy of and confidence in the use of data that works for people and society.[161]

Challenge 5: RECs struggle to review potential harms and impacts that arise throughout AI and data science research

REC reviews of AI and data science research are *ex ante* assessments done before research takes place. However, many of the harms and risks in AI research may only become evident at later stages of the research. Furthermore, many of the types of harms that can arise – such as issues of bias, or wider misuses of AI or data – are challenging for a single committee to predict. This is particularly true with the broader societal impacts of AI research, which require a kind of evaluation and review that RECs currently do not undertake.

Bias and discrimination

Identifying or predicting potential biases, and consequent discrimination, that can arise in datasets and AI models at various stages of development constitute a significant challenge for the evaluation of AI

[159] Machirori, M. and Patel. R. (2021). 'Turning distrust in data sharing into "engage, deliberate, decide"'. *Ada Lovelace Institute*. Available at: https://www.adalovelaceinstitute.org/blog/distrust-data-sharing-engage-deliberate-decide/

[160] Centre for Data Ethics and Innovation. (2020). *Addressing trust in public sector data use*. UK Government. Available at: https://www.gov.uk/government/publications/cdei-publishes-its-first-report-on-public-sector-data-sharing/addressing-trust-in-public-sector-data-use

[161] Ada Lovelace Institute. (2021). *Participatory data stewardship: A framework for involving people in the use of data*. Available at: https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/

and data science research. Numerous kinds of bias can arise during data collection, model development and deployment, leading to potentially harmful downstream effects.[162] For example, Buolamwini and Gebru demonstrate that many popular facial recognition systems have much poorer performance on darker skin and non-male identities due to sampling biases in the population dataset used to train the model.[163] Similarly, numerous studies have shown predictive algorithms for policing and law enforcement can reproduce societal biases due to choices in their model architecture, design and deployment.[164,165,166] In supervised machine learning, manually annotated datasets can harbour bias through problematic application of gender or race categories.[167,168,169] In unsupervised machine learning, datasets commonly represent different types of historical biases (because data reflect existing sociotechnical bias in the world), which lead to a lack of demographic diversity, aggregation or population.[170] Crawford argues that datasets used for model training purposes are asked to capture a very complex world through taxonomies consisting of discrete classifications, an act that requires non-trivial political, cultural and social choices.[171]

[162] Suresh, H. and Guttag, J. (2021). 'Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle'. *MIT Schwarzman College of Computing*. Available at: https://mit-serc.pubpub.org/pub/potential-sources-of-harm-throughout-the-machine-learning-life-cycle/release/1

[163] Buolamwini, J. and Gebru, T. (2018). 'Gender shades: Intersectional Accuracy Disparities in Commercial Gender Classification.' *Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Conference on Fairness, Accountability and Transparency, PMLR*, pp. 77–91. Available at: https://proceedings.mlr.press/v81/buolamwini18a.html

[164] Asaro, P.M. (2019). *AI Ethics in Predictive Policing: From Models of Threat to an Ethics of Care*. Available at: https://peterasaro.org/writing/AsaroPredicitvePolicingAIEthicsofCare.pdf

[165] O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

[166] Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016). 'Machine Bias'. *ProPublica*. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[167] Keyes, O. (2018). 'The misgendering machines: Trans/HCI implications of automatic gender recognition'. *Proceedings of the ACM on human-computer interaction*, 2(CSCW), pp. 1–22.

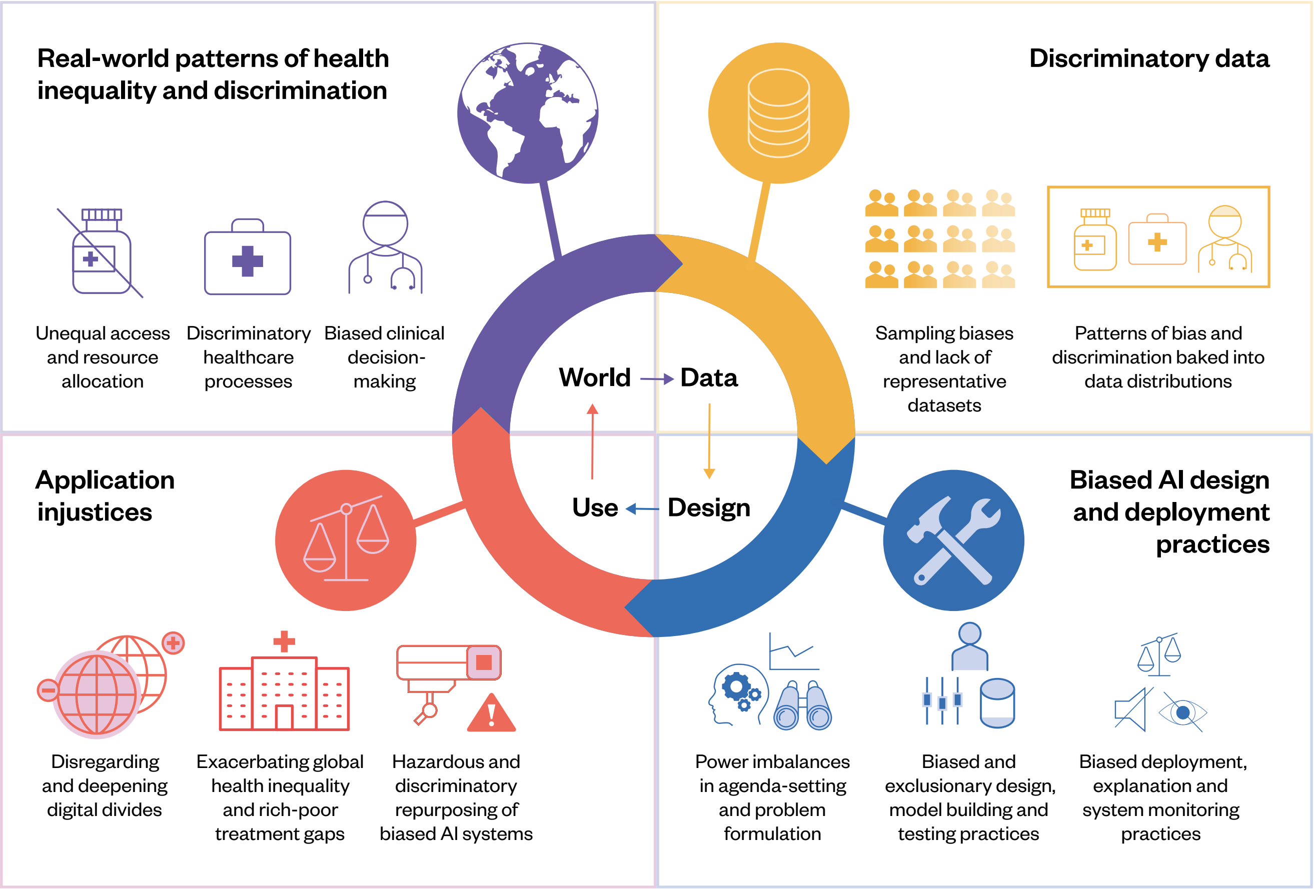
[168] Hamidi, F., Scheuerman, M. K. and Branham, S. M. (2018). 'Gender recognition or gender reductionism? The social implications of embedded gender recognition systems'. CHI '18. *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–13. Available at: https://dl.acm.org/doi/abs/10.1145/3173574.3173582

[169] Scheuerman, M. K., Wade, K., Lustig, C. and Brubaker, J. R. (2020). 'How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis'. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1), pp. 1–35. Available at: https://dl.acm.org/doi/abs/10.1145/3392866

[170] Mehrabi, N., Morstatter, F., Saxena, N. et al. (2021). 'A survey on bias and fairness in machine learning'. *ACM Computing Surveys (CSUR)*, 54(6), pp. 1–35.

[171] Crawford, K. (2021). *The Atlas of AI*. Yale University Press.

Figure 7: How bias can arise in different ways in the AI development lifecycle[172]


Real-world patterns of health inequality and discrimination
Unequal access and resource allocation
Discriminatory healthcare processes
Biased clinical decision-making
Discriminatory data
Sampling biases and lack of representative datasets
Patterns of bias and discrimination baked into data distributions
World → Data
Use ← Design
Application injustices
Disregarding and deepening digital divides
Exacerbating global health inequality and rich-poor treatment gaps
Hazardous and discriminatory repurposing of biased AI systems
Biased AI design and deployment practices
Power imbalances in agenda-setting and problem formulation
Biased and exclusionary design, model building and testing practices
Biased deployment, explanation and system monitoring practices

This figure uses the example of an AI-based healthcare application, to show how bias can arise from patterns in the real world, in the data, in the design of the system, and in its use.

Understanding the ways in which biases can arise in different stages of an AI research project creates a challenge for RECs, which may not have the capacity, time or resources to determine what kinds of biases might arise in a particular project or how they should be evaluated and mitigated. Under current REC guidelines, it may be easier for RECs to challenge researchers on how they can address questions concerning data collection and sampling bias issues, but questions concerning whether research may be used to create biased or discriminatory outcomes at the point of application are outside the scope of most REC reviews.

[172] Source: Leslie, D. et al. (2021). 'Does "AI" stand for augmenting inequality in the era of COVID-19 healthcare?'. *BMJ*, 372. Available at: https://www.bmj.com/content/372/bmj.n304

Data provenance

Workshop participants identified data provenance – how data is originally collected sourced by researchers – as another major challenge for RECs. The issue becomes especially salient when it comes to international and collaborative projects, which draw on complex networks of datasets. Some datasets may constitute 'primary' data – that is, data collected by researchers. Meanwhile, other data may be 'secondary', which includes data that is shared, disseminated or made public by others. With secondary data, the underlying purpose for its collection, its accuracy and biases embedded at the stage of collection may be unclear.

There is a need for RECs to consider not just where data is sourced from but to also probe into what its intended purposes are, how it has been tested for potential biases that may be baked into a project, and other questions about the ethics of its collection. Some participants said that it is not enough to ask whether a dataset received ethical clearance when collected. One practical tool that might address this would be standardisation of dataset documentation practices by research institutions. For example, there is the option to use datasheets, which list critical information about how a dataset was collected, who to contact with questions and what potential ethical issues it may raise.

Labour practices around data labelling

Another issue flagged by our workshop participants related to considerations for the labour conditions and mental and physical wellbeing of data annotators. Data labellers form part of the backbone of AI and data science research, and include people who review, tag and label data to form a dataset, or evaluate the success of a model. These workers are often recruited from services like MTurk. Research and data labeller activism has shown that many face exploitative working conditions and underpayment.[173]

According to some workshop participants, it remains unclear whether data labellers are considered 'human subjects' in their reviews. Their

[173] Irani, L. C. and Silberman, M. S. (2013). 'Amazon Mechanical Turk: Gold Mine or Coal Mine?' *CHI '13: Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 611–620); Available at: https://dl.acm.org/doi/abs/10.1145/2470654.2470742

wellbeing is not routinely considered by RECs. While some institutions maintain MTurk policies, these are often not written from the perspective of workers themselves and may not fully consider the variety of risks that workers face. These can include non-payment of services, or asking workers to undertake too much work in too short of a time.[174] Initiatives like the Partnership on AI's *Responsible Sourcing of Data Enrichment Services* and the Northwestern Institutional Review Board's *Guidelines for Academic Requesters* offer models for how corporate and academic RECs might develop policies.[175]

Societal and downstream impacts

Several experts noted standard RECs practices can fail to assess the broader societal impacts of AI and data science research, leading to traditionally marginalised population groups being disproportionately affected by AI and data science research. Historically, RECs have an anticipatory role, with potential risks assessed and addressed at the initial planning stage of the research. The focus on protecting individual research subjects means that RECs generally do not consider potential broader societal impacts, such as long-term harms to communities.[176]

For example, a study using facial recognition technology to determine sexual orientation of people,[177] or the recognition of Uighur minorities in China,[178] poses serious questions for societal benefit and the impacts on marginalised communities – yet the RECs who reviewed these projects did not consider these kinds of questions. Since the datasets used in these projects consisted of images scraped from the internet and curated, the research did not constitute human subjects research, and therefore passed ethics review.

[174] Massachusetts Institute of Technology – Committee on the Use of Humans as Experimental Subjects. *COUHES Policy for Using Amazon's Mechanical Turk*. Available at: https://couhes.mit.edu/guidelines/couhes-policy-using-amazons-mechanical-turk

[175] Jindal, S. (2021). 'Responsible Sourcing of Data Enrichment Services'. *Partnership on AI*. Available at: https://partnershiponai.org/responsible-sourcing-considerations/; and Northwestern University. *Guidelines for Academic Requesters*. Available at: https://irb.northwestern.edu/docs/guidelinesforacademicrequesters-1.pdf

[176] Friesen, P., Douglas-Jones, R., Marks, M. et al. (2021). 'Governing AI-Driven Health Research: Are IRBs Up to the Task?' *Ethics & Human Research*, 43(2), pp. 35–42. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/eahr.500085

[177] Wang, Y. and Kosinski, M. (2018). 'Deep neural networks are more accurate than humans at detecting sexual orientation from facial images'. *Journal of Personality and Social Psychology*, 114(2), p. 246. Available at: https://psycnet.apa.org/doiLanding?doi=10.1037%2Fpspa0000098

[178] Wang, C., Zhang, Q., Duan, X. and Gan, J. (2018). 'Multi-ethnical Chinese facial characterization and analysis'. *Multimedia Tools and Applications*, 77(23), pp. 30311–30329.

Environmental impacts

The environmental footprint of AI and data science is a further significant impact that our workshop participants highlighted as an area most RECs do not currently review for. Some forms of AI research, such as deep learning and multi-agent learning, can be compute-intensive, raising questions about whether their benefits offset the environmental cost.[179] Similar questions have been raised about large language models (LLMs), such as OpenAI's GPT-3, which rely on intensive computational methods without articulating a clearly defined benefit to society.[180] Our workshop participants noted that RECs could play a role in assessing whether a project's aims justify computationally intensive methods, or whether a researcher is using the most computationally efficient method of training their model (avoiding unnecessary computational spend). However, there is no existing framework for RECs to use to help make these kinds of determinations, and it is unclear whether many REC members would have the right competencies to evaluate such questions.

Considerations of 'legitimate research'

Workshop participants discussed whether RECs are well suited to determine what constitutes 'legitimate research'. For example, some participants raised questions about the intellectual proximity of AI research to discredited forms of pseudoscience like phrenology, citing AI research that is based on flawed assumptions about race and gender – a point raised in empirical research evaluating the use of AI benchmark datasets.[181] AI and data science research regularly involves the categorisation of data subjects into particular groups, which may involve crude assumptions that, nonetheless, can lead to severe population-level consequences. These 'hidden decisions' are often baked into a dataset and, once shared, can remain unchallenged for long periods of time. To give one example, portions of the MIT Tiny Images dataset, first created in 2006, were removed in 2018 after it was discovered to include racist

[179] Strubell, E., Ganesh, A. and McCallum, A. (2019). 'Energy and policy considerations for deep learning in NLP'. *arXiv*. Available at: https://arxiv.org/abs/1906.02243

[180] Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021). 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?' *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pp. 610–623. Available at: https://doi.org/10.1145/3442188.3445922

[181] Denton, E., Hanna, A., Amironesei, R. et al. (2020). 'Bringing the people back in: Contesting benchmark machine learning datasets'. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2007.07399

and sexist categorisations of images of minoritised people and women.[182] This dataset has been used to train a range of subsequent models and may still be in use today, given the ability to download and repost datasets without subsequent documentation explaining their limitations. Several participants noted that RECs are not set up to identify, let alone assess, for these kinds of issues, and may consider defining 'good science' out of their remit.

A lack of incentives for researchers to consider broader societal impacts

Another point of discussion in the workshops was how to incentivise researchers to consider broader societal impact questions. Researchers are usually incentivised and rewarded by producing novel and innovative work, evidenced by publications in relevant scientific journals or conferences. Often, this involves researchers making broad statements about how AI or data science research can have positive implications for society, yet there is little incentive for researchers to consider potentially harmful impacts of their work.

Some of the expert participants pointed out that other actors in the research ecosystem, such as funders, could help to incentivise researchers to reflexively consider and document the potential broader societal impacts of their work. Stanford University's Ethics and Society Review, for example, requires researchers seeking funding from the Stanford Institute for Human-Centered Artificial Intelligence to write an impact statement reflecting on how their proposal might create negative societal impacts for society, how they can mitigate those impacts, and to work with an interdisciplinary faculty panel to ensure those concerns are addressed before funding is received. Participants in this programme overwhelmingly described it as a positive for their research and training experience.[183]

A more ambitious proposal from some workshop participants was to go beyond a risk-mitigation plan and incentivise research that benefits

[182] Birhane, A. and Prabhu, V. U. (2021). 'Large image datasets: A pyrrhic win for computer vision?'. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1537–1547. Available at: https://doi.org/10.48550/arXiv.2006.16923

[183] Jensen, B. (2021). 'A New Approach to Mitigating AI's Negative Impact'. *Institute for Human-Centered Artificial Intelligence*. Available at: https://hai.stanford.edu/news/new-approach-mitigating-ais-negative-impact

society. However, conceptualisations of social, societal or public good are contested, at best – there is no universally agreed on theory of what these are.[184] There are also questions about who is included in 'society,' and whether some benefits for those in a position of power would actively harm other members of society who are disadvantaged.

AI and data science research communities have not yet developed a rigorous method for deeply considering what constitutes public benefit, or a rigorous methodology for assessing the long-term impact of AI and data science interventions. Determining what constitutes the 'public good' or 'public benefit' would, at the very least, require some form of public consultation; even then, it may not be sufficient.[185]

One participant noted it is difficult in some AI and data science research projects to consider these impacts, particularly projects aimed at theory-level problems or small step-change advances in efficiency (for example, research that produces a more efficient and less computationally intensive method for training an image detection model). This dovetails with concerns raised by some in the AI and data science research community that there is too great a focus on creating novel methods for AI research instead of applying research to address applied, real-world problems.[186]

Workshop participants raised a similar concern about AI and data science research that is conducted without any clear rationale for addressing societal problems. Participants used the metaphor of a 'fishing expedition' to describe some types of AI and data science research projects that have no clear aim or objective but sought to explore large datasets to see what they found. As one workshop participant put it, researchers should always be aware that, just because data can be collected, or is already available, it does not mean that it should be collected or used for any purpose.

[184] Green, B. (2019). '"Good" isn't good enough'. *Proceedings of the AI for Social Good workshop at NeurIPS*. Available at: http://ai.ethicsworkshop.org/Library/LibContentAcademic/GoodNotGoodEnough.pdf

[185] For example, the UK National Data Guardian published the results of a public consultation on how health and care data should be used to benefit the public, which may prove a model for the AI and data science research communities to follow. See: National Data Guardian. (2021). *Putting Good Into Practice. A public dialogue on making public benefit assessments when using health and care data.* UK Government. Available at: https://www.gov.uk/government/publications/putting-good-into-practice-a-public-dialogue-on-making-public-benefit-assessments-when-using-health-and-care-data

[186] Kerner, H. (2020). 'Too many AI researchers think real-world problems are not relevant'. *MIT Technology Review*. Available at: https://www.technologyreview.com/2020/08/18/1007196/ai-research-machine-learning-applications-problems-opinion/

Challenge 6: Corporate RECs lack transparency in relation to their processes

Some participants noted that, while corporate lab reviews may be more extensive, they can also be more opaque, and are at risk of being driven by interests beyond research ethics, including whether research poses a reputational risk to the company if published. Moss and Metcalf note how ethics practices in Silicon Valley technology companies are often chiefly concerned with questions of corporate values and legal risk and compliance, and do not systematically address broader issues such as questions around moral, social and racial justice.[187] While corporate ethics reviewers draw on a variety of guidelines and frameworks, they may not address ongoing harms, evaluate these harms outside of the corporate context, or evaluate organisational behaviours and internal incentive structures.[188] It is worth noting that academic RECs have faced a similar criticism. Recent research has documented how academic REC decisions can be driven by a reputational interest to avoid 'embarrassment' of the institution.[189]

Several of our participants highlighted the relative lack of external transparency of corporate REC processes versus academic ones. This lack of transparency can make it challenging for other members of the research community to trust that corporate research review practices are sufficient.

Google, for example, launched a 'sensitive topics' review process in 2020 that asks researchers to run their work through legal, policy and public relations teams if it relates to certain topics like face and sentiment analysis or categorisations of race, gender or political affiliation.[190] According to the policy, 'advances in technology and the growing complexity of our external environment are increasingly leading to situations where seemingly inoffensive projects raise ethical, reputational, regulatory or legal issues.' In at least three reported instances, researchers were told to 'strike a more positive tone' and to remove references to Google products, raising concerns about the

[187] Moss, E. and Metcalf, J. (2020). *Ethics Owners. A New Model of Organizational Responsibility in Data-Driven Technology Companies. Data & Society*. Available at: https://datasociety.net/library/ethics-owners/

[188] Moss, E. and Metcalf, J. (2020).

[189] Hedgecoe, A. (2015). 'Reputational Risk, Academic Freedom and Research Ethics Review'. *British Sociological Association*, 50(3), pp.486–501. Available at: https://journals.sagepub.com/doi/full/10.1177/0038038515590756

[190] Dave, P. and Dastin, J. (2020) 'Google told its scientists to "strike a positive tone" in AI research – documents'. *Reuters*. Available at: https://www.reuters.com/article/us-alphabet-google-research-focus-idUSKBN28X1CB

credibility of findings. In one notable example that became public in 2021, a Google ethical AI researcher was fired from their role after being told that a research paper they had written, which was critical of the use of large language models (a core component in Google's search engine), could not be published under this policy.[191]

[191] Simonite, T. (2021). 'What Really Happened When Google Ousted Timnit Gebru'. *Wired*. Available at: https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/

Recommendations

We conclude this paper with a set of eight recommendations, organised into sections aimed primarily at three stakeholders in the research ethics ecosystem:

1. Academic and corporate Research Ethics Committees (RECs) evaluating AI and data science research.
2. Academic and corporate AI and data science research institutions.
3. Funders, conference organisers, journal editors, and other actors in the wider AI and data science research ecosystems.

For academic and corporate RECs

Recommendation 1: Incorporate broader societal impact statements from researchers

The problem

Broader societal impacts of AI and data science research are not currently considered by RECs. These might include 'dual-use' research (meaning it can be used for both civilian and military purposes), possible harms to society or the environment, and the potential for discrimination against marginalised populations. Instead, RECs focus their reviews on questions of research methodology. Several workshop participants noted that there are few incentives for researchers to reflexively consider questions of societal impact. Workshop participants also noted that institutions do not offer any framework for RECs to follow, or training or guidance for researchers. Broader societal impact statements can ensure researchers reflect on, and document, the full list of potential harms, risks and benefits their work may pose.

Recommendations

Researchers should be required to undertake an evaluation of broader societal impact as part of their ethics evaluation.
This would be an impact statement that included a summary of the positive and negative impacts on society they anticipate from their research. They should include any known limitations or risks for misuse that may arise, such as whether their research findings are premised on assumptions that are particular to a geographic region, or if there is a possibility of using the findings to exacerbate certain forms of societal injustices.

Training should be designed and implemented for researchers to adequately conduct stakeholder and impact assessment evaluations, as a precondition to receive funding or ethics approval.[192]
These exercises should encourage researchers to consider the intended uses of their innovations and reflect on what kinds of unintended uses might arise. The result of these assessments can be included in research ethics documentation that reports on the researchers' reflections on both discursive questions that invite open-ended opinion (such as what the intended use of the research may be) and categorical information that lists objective statistics and data about the project (such as the datasets that will be used, or the methods that will be applied). Some academic institutions are experimenting with this approach for research ethics applications.

Examples of good practice

Recent research from Microsoft provides a structured exercise for how researchers can consider, document and communicate potential broader societal impacts, including who the affected stakeholders are in their work, and what limitations and potential benefits it may have.[193]

Methods for impact assessment of algorithmic systems have emerged from the domains of human rights, environmental studies and data protection law. These methods are not necessarily standardised or

[192] Ferretti, A., Ienca, M., Sheehan, M. et al. (2021). 'Ethics review of big data research: What should stay and what should be reformed?'. *BMC Medical Ethics*, 22(1), pp. 1–13. Available at: https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-021-00616-4

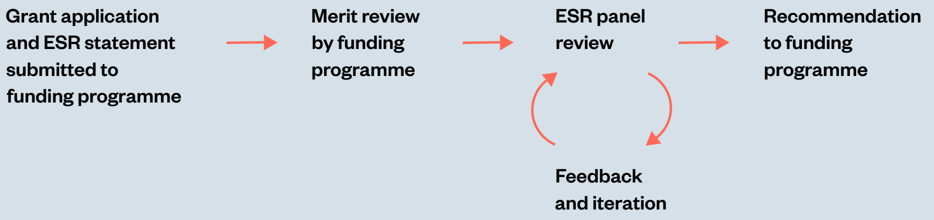
[193] Smith, J. J., Amershi, S., Barocas, S. et al. (2022). 'REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research'. *2022 ACM Conference on Fairness, Accountability, and Transparency (FaccT '22)*. Available at: https://facctconference.org/static/pdfs_2022/facct22-47.pdf

consistent, but they seek to encourage researchers to reflect on the impacts of their work. Some examples include the use of algorithmic impact assessments in healthcare settings,[194] and in public sector uses of algorithmic systems in the Netherlands and Canada.[195]

In 2021, Stanford University tested an Ethics and Society Review board (ESR), which sought to supplement the role of its Institutional Review Board. The ESR requires researchers seeking funding from the Stanford Institute for Human-Centered Artificial Intelligence to consider negative or societal risks from their proposal, develop mitigative measures to assess those risks, and to collaborate with an interdisciplinary faculty panel to ensure concerns are addressed before funds are disbursed.[196] A pilot study of 41 submissions to this panel found that '58% of submitters felt that it had influenced the design of their research project, 100% are willing to continue submitting future projects to the ESR,' and that submitting researchers sought additional training and scaffolding about societal risks and impacts.[197]

Figure 8: How the Stanford University Ethics and Society Review (ESR) works[198]

Grant funding from participating institutions is not released until the researchers successfully complete an iterative process on their proposed project. Conditioning funding on the ESR process helps engage researchers at the formative stages of their research, when projects are still open to change, and ensures broad engagement with the process rather than self-selection of just those who are motivated.


Grant application and ESR statement submitted to funding programme
Merit review by funding programme
ESR panel review
Recommendation to funding programme
Feedback and iteration

[194] Ada Lovelace Institute. (2021). *Algorithmic impact assessment: a case study in healthcare*. Available at: https://www.adalovelaceinstitute.org/project/algorithmic-impact-assessment-healthcare/

[195] Zaken, M. van A. (2022). *Impact Assessment Fundamental Rights and Algorithms.* The Ministry of the Interior and Kingdom Relations. Available at: https://a.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms; Government of Canada. (2021). *Algorithmic Impact Assessment Tool*. Available at: https://www.canada.ca/en/governmentasystem/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

[196] Jensen, B. (2021). 'A New Approach To Mitigating AI's Negative Impact'. *Institute for Human-Centered Artificial Intelligence*. Available at: https://hai.stanford.edu/news/new-approach-mitigating-ais-negative-impact

[185] Bernstein, M. S., Levi, M., Magnus, D. et al. (2021). 'ESR: Ethics and Society Review of Artificial Intelligence Research'. *arXiv*. Available at: https://arxiv.org/abs/2106.11521

[198] Center for Advanced Study in the Behavioral Sciences at Stanford University. 'Ethics & Society Review – Stanford University'. Available at: https://casbs.stanford.edu/ethics-society-review-stanford-university

Understanding the potential impacts of AI and data science research can ensure researchers produce technologies that are fit for purpose and well-suited for the task at hand. The successful development and integration of an AI-powered sepsis diagnostic tool in a hospital in the USA offers an example of how researchers worked with key stakeholders to develop and design a life-changing product. Researchers on this project relied on continuous engagement with stakeholders in the hospital, including nurses, doctors and other staff members, to determine how the system could meet their needs.[199] By understanding these needs, the research team were able to tailor the final product so that it fitted smoothly within the existing practices and procedures of this hospital.

Open questions

There are several open questions on the use of broader societal impact statements. One relates to whether these statements should be a basis for a REC rejecting a research proposal. This was a major point of disagreement among our workshop participants. Some participants pushed back on the idea, out of concern that research institutions should not be in the position to determine what research is appropriate or inappropriate based on potential societal impacts, and that this may cause researchers to view RECs as a policing body for issues that have not occurred. Instead, these participants suggested a softer approach, whereby RECs require researchers to draft a broader societal impact statement but there is not a requirement for RECs to evaluate the substance of those assessments. Other participants noted that these impact assessments would be likely to highlight clear cases where the societal risks are too great, and that RECs should incorporate these considerations into their final decisions.

Another consideration related to whether a broader societal impacts evaluation should involve some aspect of *ex post* reviews of research, in which research institutions monitor the actual impacts of published research. This process would require significant resourcing. While there is no standard method for conducting these kinds of reviews yet,

[199] Sendak, M., Elish, M.C., Gao, M. et al. (2020). '"The Human Body Is a Black Box": Supporting Clinical Decision-Making with Deep Learning.' *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 99–109. Available at: https://doi.org/10.1145/3351095.3372827

some researchers in the health field have called for this kind of *ex post* review conducted by an interdisciplinary committee of academics and stakeholders.[200]

Lastly, some workshop participants questioned whether a more holistic ethics review process could be broken up into parts handled by different sub-committees. For example, could questions of data ethics – how data should be handled, processed and stored, and which datasets are appropriate for researchers to use – have their own dedicated process or sub-committee? This sub-committee would need to adopt clear principles and set expectations with researchers for specific data ethics practices, and could also address the evolving dynamic between researcher and participants.

There was a suggestion that more input from data subjects could help, with a focus on how they can, and whether they should, benefit from the research, and whether this would therefore constitute a different type or segment of ethical review. Participants mentioned the need for researchers to think relationally and understand who the data subject is, the power dynamics at play and to work out the best way of involving research participants in the analysis and dissemination of findings.

Recommendation 2: RECs should adopt multi-stage ethics review processes for AI and data science research

The problem

Ethical and societal risks of AI and data science research can manifest at different stages of research[201] – from early ideation to data collection, to pre-publication. Assessing the ethical and broader societal impacts of AI research can be difficult as the results of data-driven research cannot be known in advance of accessing and processing data or building machine learning (ML) models. Typically, RECs only review research applications once before research beings, and with a narrow focus solely

[200] Samuel, G. and Derrick, D. (2020). 'Defining ethical standards for the application of digital tools to population health research'. *Bulletin of the World Health Organization Supplement*, 98(4), pp. 239–244. Available at: https://pubmed.ncbi.nlm.nih.gov/32284646/

[201] Kawas, S., Yuan, Y., DeWitt, A. et al (2020). 'Another decade of IDC research: Examining and reflecting on values and ethics'. *IDC '20: Proceedings of the Interaction Design and Children Conference*, pp. 205–215. Available at: https://dl.acm.org/doi/abs/10.1145/3392063.3394436

looking at ethical issues pertaining to methodology. This can mean that ethical review processes fail to catch risks that arise in later stages, such as potential environmental or privacy considerations if research is published, particularly for research that is 'high risk' and pertains to protected characteristics or has high potential for societal impact.

Recommendations

RECs should set up multi-stage and continuous ethics reviews, particularly for 'high-risk' AI research

RECs should experiment with requiring multiple stages of evaluations of research that raises particular ethical concern, such as evaluations at the point of data collection and a separate evaluation at the point of publication. Ethics review processes should engage with considerations raised at all stages of the research lifecycle. RECs must move away from being the 'owners' of ethical thinking into being stewards who guide researchers through the review process.

This means challenging the notion of an ethical review being a one-off exercise conducted at the start of a project, and instead shifts the approach of a REC and the ethics review process towards one that embeds ethical reflection throughout a project. This will benefit from more iterative ethics review processes, as well as additional interdisciplinary training for AI and data science researchers.

Several workshop participants suggested that multi-stage ethics review could consist of a combination of formal and informal review processes. Formal review processes could exist at the early and late stages, such as funding or publication, while at other points, the research team could be asked to engage in more informal peer-reviews or discussions with experts or reviewers. In the early stages of the project, milestones could be identified which are defined by the research teams, and in collaboration with RECs. For example, a milestone could be a grant submission, or when changing roles or adding new research partners to the project. Milestones could be used to trigger an interim review. Rather than following a standardised approach, this model allows for flexibility, as the milestones would be different for each project. This could also involve a tiered assessment, which is a standardised assessment based on identified risks a research project poses, which then determines the milestones.

Building on Burr & Leslie,[202] we can speak of four broad stages in an AI or data science research project: design, develop, pre-publication and post-deployment.

At the stage of **designing** a research project, policies and resources should be in place to:

- Ensure new funders and potential partnerships adhere to an ethical framework. Beyond legal due diligence, this is about establishing partnerships on the basis of their values and a project's goals.
- Implement scoping policies that establish whether a particular research project must undertake REC processes. Two ways are suggested in the literature for such policies, and examining each organisation's research and capability will help decide which is most suitable:
  - Sandler et al suggest a consultation process whereby RECs produce either 'an Ethical Issues Profile report or a judgment that there are not substantive ethical issues raised'.[203]
  - The UK Statistics Authority employs an ethics self-assessment tool that determines a project's level of risk.[204]
- Additionally, scoping processes can result in establishing whether a project must undertake data, stakeholder, human rights or other impact assessments that focus on the broader societal impacts of their work (see Recommendation 1). Stanford's Ethical and Societal Review Board offers one model for how institutions can set more 'carrots and sticks' for researchers to reflexively engage in the potential broader impacts of their research by tying the completion of a societal impact statement to their funding proposal.

[202] Burr, C. and Leslie, D. (2021). 'Ethical Assurance: A Practical Approach to the Responsible Design, Development, and Deployment of Data-Driven Technologies'. *Social Science Research Network*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3937983

[203] Sandler, R. and Basel, J. (2019). *Building Data and AI Ethics Committees*, p. 19. Accenture. Available at: https://www.accenture.com/_acnmedia/pdf-107/accenture-ai-and-data-ethics-committee-report-11.pdf

[204] UK Statistics Authority. *Ethics Self-Assessment Tool*. Available at: https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethics-self-assessment-tool/

At the **development** stage of a project, a typical REC evaluation should be undertaken to consider any ethical risks. RECs should provide a point of contact to ensure changes in the project's aims and methods that raise new challenges are subjected to due reflection. This ensures an iterative process that aligns with the practicalities of research. RECs may also experiment with creating specialised sub-committees that address different issues, such as a separate data ethics review board that includes expertise in data ethics and domain-specific expertise, or a health data or social media data review board. It could help evaluate potential impact for people and society; depending on composition, it could also be adept at reviewing the technical aspects of a research project.[205] This idea builds from a hybrid review mechanism that Ferretti et al propose, which merges aspects of the traditional model of RECs with specialised research committees that assess particular parts of a research project.[206]

One question that RECs must turn into practice is to establish which projects must undertake particular REC processes, as it may be too burdensome for all projects to undergo this scrutiny. In some cases, it may be that a REC determines a project should undergo stricter scrutiny if an analysis of its potential impacts on various stakeholders highlights serious ethical issues. Whether or not a project is 'in scope' for a more substantial REC review process might depend on:

- the level of risk it raises
- the training or any certifications its researchers hold
- whether it is reviewed by a relevant partner's REC.

Determining what quantifies a risk is challenging, as not all risks may be evident or within the imagination of a REC. More top-level guidance on risks (see Recommendation 4) and interdisciplinary/experiential membership on RECs (see Recommendation 3) can help ensure that a wider scope of AI risks are identified.

At the stage of **pre-publication** of a research project, RECs should encourage researchers to revisit the ethical and broader societal impact considerations that may have arisen earlier. In light of the research

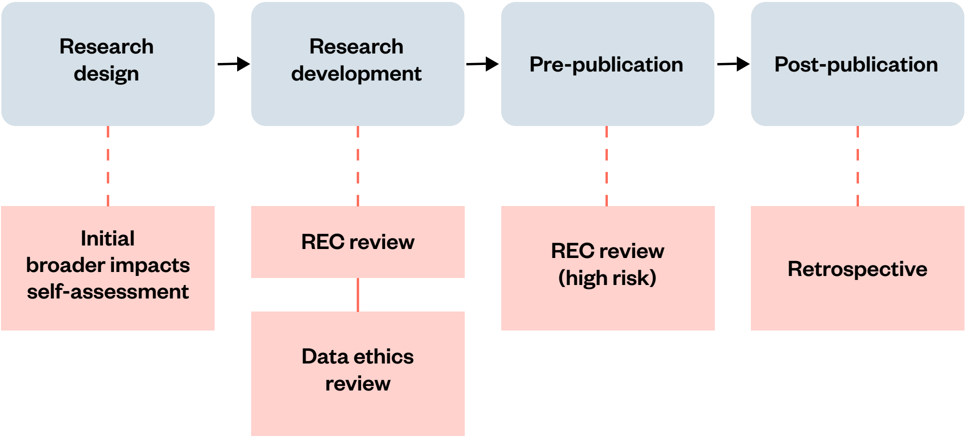
[205] Ferretti, A., Ienca, M., Sheehan, M. et al. (2021). 'Ethics review of big data research: What should stay and what should be reformed?'. *BMC Medical Ethics*, 22(1), pp. 1–13. Available at: https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-021-00616-4

[206] Ferretti, A., Ienca, M., Sheehan, M. et al. (2021).

findings, have these changed at all? Have new risks arisen? At this stage, REC members can act as stewards to help researchers navigate publication requirements, which may include filling in the broader societal impact statements that some AI and ML conferences are beginning to implement. They might also connect researchers with subject-matter experts in particular domains, who can help them understand potential ethical risks with their research. Finally, RECs may be able to provide guidance on how to release research responsibly, including whether to release publicly a dataset or code that may be used to cause harm.

Lastly, RECs and research institutions should experiment with **post-publication** evaluations of the impacts of research. RECs could, for example, take a pool of research submissions that involved significant ethical review and conduct an analysis of how that work was received 2–3 years down the line. Criteria this assessment could look at may include how that work was received by the media or press, who has cited that work subsequently, and whether negative or positive impacts came to fruition.

Figure 9: Example of multi-stage ethics review process



This figure shows what a multi-stage ethics review process could look like. It involves an initial self-assessment for broader impacts issues at the design stage, a REC review (and potential review by a specialised data ethics board at the production stage, another review of high-risk research at pre-publication stage, and a potential post-publication review of the research 2–3 years after it is published.

Examples of good practice

As explored above, there is not yet consensus on how to operationalise a continuous, multi-stage ethics review process, but there is an emerging body of work addressing ethics consideration at different stages in a projects' lifecycle. Building on academic research,[207] the UK's Centre for Data Ethics and Innovation has proposed an 'AI assurance' framework for continuously testing the potential risks of AI systems. This framework involves the use of different mechanisms like audits, testing and evaluation at different stages of an AI product's lifecycle.[208] However, this framework is focused on AI products rather than research, and further work would be needed to adapt this framework for research.

D'Aquin et al propose an ethics-by-design methodology for AI and data science research that takes a broader view of data ethics.[209] Assessment usually happens at the research design/planning stage, and there are no incentives for the researcher to consider ethical issues as they eventually emerge with the progress of research. Instead, considerations for emerging ethical risks should be ongoing.[210] A few academic and corporate research institutions, such as the Alan Turing Institute, have already introduced or are in the process of implementing continuous ethics review processes (see Appendix 2). Further research is required to study how these work in practice.

Open questions

A multi-stage research review process should capture more of the ethical issues that arise in AI research, and enable RECs to evaluate *ex post* impacts of their research. However, continuous, multi-stage reviews require a substantial increase in resources and so are an option only for institutions who are ready to make an investment in ethics practices. These proposals could require multiples of the current time

[207] The concept of 'ethical assurance' is a process-based form of project governance that supports inclusive and participatory ethical deliberation while also remaining grounded in social and technical realities. See: Burr, C. and Leslie, D. (2021). 'Ethical Assurance: A Practical Approach to the Responsible Design, Development, and Deployment of Data-Driven Technologies'. *Social Science Research Network*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3937983

[208] Centre for Data Ethics and Innovation (2022). *The roadmap to an effective AI assurance ecosystem*. UK Government. Available at: https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem

[209] d'Aquin, M., Troullinou, P., O'Connor, N. E. et al. (2018). 'Towards an "Ethics by Design" Methodology for AI research projects'. *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 54–59. Available at: https ://dl.acm.org/doi/abs/10.1145/3278721.3278765

[210] d'Aquin, M., Troullinou, P., O'Connor, N. E. et al. (2018).

commitments of REC members and officers, and therefore require greater compensation for REC members.

The prospect of implementing a multi-stage review process raises further questions of scope, remit and role of ethics reviews. Informal reviews spread over time could see REC members take more of an advisory role than in the compliance-oriented models of the status quo, allowing researchers to informally check in with ethics experts, to discuss emerging issues and the best way to approach them. Dove argues that the role of RECs is to operate as regulatory stewards, who guide researchers through the review process.[211] To do this, RECs should establish communication channels for researchers to get in touch and interact. However, Ferretti et al warn there is a risk that ethics oversight might become inefficient if different committees overlap, or if procedures become confusing and duplicated. It would also be challenging to bring together different ethical values and priorities across a range of stakeholders, so this change needs sustaining over the long term.[212]

Recommendation 3: Include interdisciplinary expertise in REC membership

The problem

The make-up and scope of a REC review came up repeatedly in our workshops and literature reviews, with considerable concern raised about how RECs can accurately capture the wide array of ethical challenges posed by different kinds of AI and data science research. There was wide agreement within our workshop of the importance of ensuring that different fields of expertise have their voices heard in the REC process, and that the make-up of RECs should reflect a diversity of backgrounds.

[211] Dove, E. (2020). *Regulatory Stewardship of Health Research: Navigating Participant Protection and Research Promotion*. Edward Elgar.

[212] Ferretti, A., Ienca, M., Sheehan, M. et al. (2021). 'Ethics review of big data research: What should stay and what should be reformed?'. *BMC Medical Ethics*, 22(1), pp. 1–13. Available at: https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-021-00616-4

Recommendations

RECs must include more interdisciplinary expertise in their membership

In recruiting new members, RECs should draw on members from different research and professional fields that go beyond just computer science, such as the social sciences, humanities, STEM sciences and other fields. By having these different disciplines present, they can each bring a different ethical lens to the challenges that a project may raise. RECs might also consider including members who work in the legal, communications or marketing teams to ensure that the concerns raised speak to a wider audience and respond to broader institutional contexts. Interdisciplinarity involves the development of a common language, a reflective stance towards research, and a critical perspective towards science.[213] If this expertise is not present at an institution, RECs could make greater use of external experts for specific questions that arise from data science research.[214]

RECs must include individuals with different experiential expertise

RECs must also seek to include members who represent different forms of experiential expertise, which includes individuals from historically marginalised groups with perspectives that are often not represented in these settings. This both includes more diverse experiences in discussions about data science and AI research outputs, and ensures that these meet the values of a culturally rich and heterogeneous society.

Crucially, the mere representation of a diversity of viewpoints is not enough to ensure the successful integration of those views into REC decisions. Members must feel empowered to share their concerns and be heard, and careful attention must be paid to the power dynamics that underlie how decisions are made within a REC. Mechanisms for ensuring more transparent and ethical decision-making practices are an area of future research worth pursuing.

[213] d'Aquin, M., Troullinou, P., O'Connor, N. E. et al. (2018). 'Towards an "Ethics by Design" Methodology for AI research projects'. *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 54–59. Available at: https ://dl.acm.org/doi/abs/10.1145/3278721.3278765

[214] Ferretti, A., Ienca, M., Sheehan, M. et al. (2021). 'Ethics review of big data research: What should stay and what should be reformed?'. *BMC Medical Ethics*, 22(1), pp. 1–13. Available at: https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-021-00616-4

In terms of the composition of RECs, Ferretti et al suggest that these should become more diverse and include members of the public and research subjects or communities affected by the research.[215] Besides the public, members from inside an institution should also be selected to achieve a multi-disciplinary composition of the board.

Examples of good practice

One notable example is the SAIL (Secure Anonymised Information Linkage) Databank, a Wales-wide research databank with approximately 30 billion records of individual level population datasets. Requests to access the databank are reviewed by an Information Governance Review Panel which includes representatives from public health agencies, clinicians, and members of the public who may be affected by the uses of this data. More information on SAIL can be found in Appendix 2.

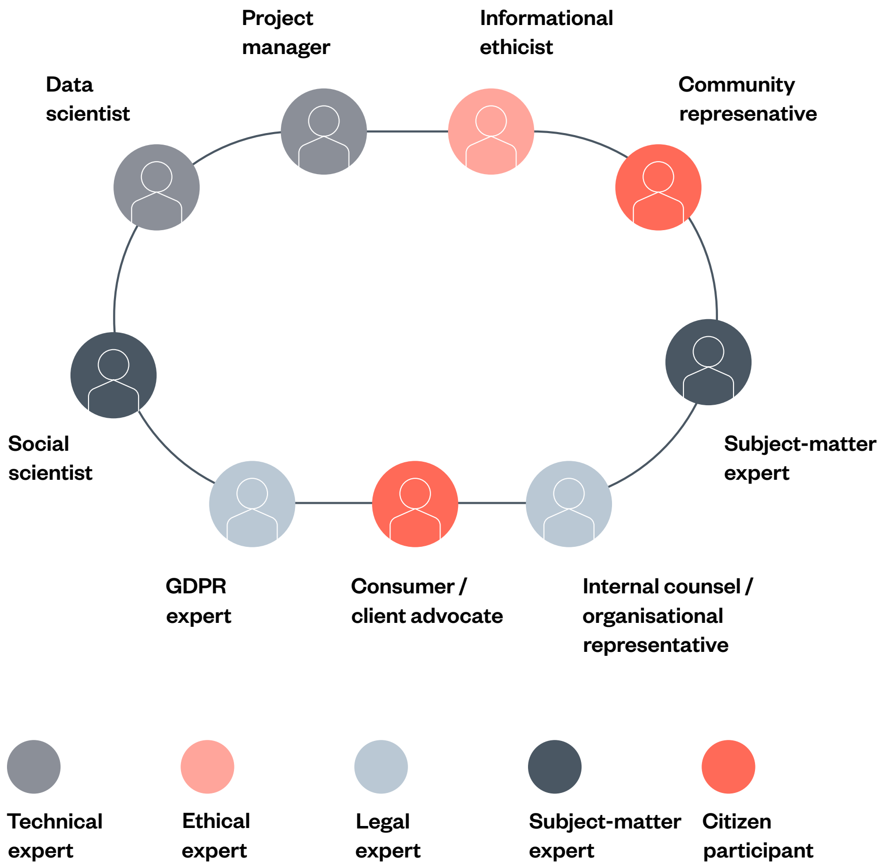
Open questions

Increasing experiential and subject-matter expertise in AI and data science research reviews will hopefully lead to more holistic evaluations of the kinds of risks that may arise, particularly given the wide range of societal applications of AI and data science research. However, expertise from members of the public and external experts must be fairly compensated, and the impact of more diverse representation on these boards should be the subject of future study and evaluation.

[215] Ferretti, A., Ienca, M., Sheehan, M. et al (2021).

Figure 10: The potential make-up of an AI/data science ethics committee[216]

Hypothetical committee on data use



This figure illustrates the possible composition of a data ethics committee. It is meant to exemplify the types of expertise that might be included on the committee, as well as the types of specialists that might provide such expertise. For any given ethics committee, members should have appropriate domain knowledge as well. For example, the consumer advocate, social scientist , subject-matter expert, information ethicist, and internal counsel appropriate for a committee whose domain is health care would be different from one whose domain is financial services.

[216] Source: Sandler, R. and Basl, J. (2019). *Building Data and AI Ethics Committees*, p. 19. Accenture. Available at: https://www.accenture.com/_acnmedia/pdf-107/accenture-ai-and-data-ethics-committee-report-11.pdf

For academic/corporate research institutions

Recommendation 4: Create internal training and knowledge-sharing hubs for researchers and REC members, and encourage more cross-institutional learning

The problem

A recurring concern raised by members of our workshops was a lack of shared resources to help RECs address common ethical issues in their research. This was coupled with a lack of transparency and openness of decision-making in many modern RECs, particularly for some corporate institutions where publication review processes can feel opaque to researchers. When REC processes and decisions are enacted behind closed doors, it becomes challenging to disseminate lessons learned to other institutions and researchers. It also raises a challenge for researchers who may come to view a REC as a 'compliance' body, rather than a resource for seeking advice and guidance. Several workshop participants noted that shared resources and trainings could help REC members, staff and students to better address these issues.

Recommendations

Research institutions should create institutional training and knowledge-sharing hubs

These hubs can serve five core functions:

1. **Pooling shared resources on common AI and data science ethics challenges for students, staff and REC members to use.**

The repository can compile resources, news articles and literature on ethical risks and impacts of AI systems, tagged and searchable by research type, risk or topic. These can prompt reflection on research ethics by providing students and staff with current, real-world examples of these risks in practice.

The hub could also provide a list of 'banned' or problematic datasets that staff or students should not use. This could help address concerns around datasets that are collected without underlying consent from research subjects, and which are commonly used as 'benchmark' datasets. The Duke University MTMC dataset of recorded videos on campus, for example, continues to be used by computer vision researchers in papers, despite being removed by the university due to ethical concerns. Similar efforts to create a list of problematic datasets are underway at some major AI and ML conferences, and some of our workshop participants suggested that some research institutions already maintain lists like this.

2. **Providing hypothetical or actual case studies of previous REC submissions and decisions to give a sense of the kinds of issues others are facing.**

Training hubs could include repositories of previous applications that have been scrutinised and approved by the pertinent REC, which form a body of case studies that can inform both REC policies and individual researchers. Given the fast pace of AI and data science research, RECs can often encounter novel ethical questions. By logging past approved projects and making them available to all REC members, RECs can ensure consistency in their decisions about new projects.

We suggest that logged applications also be made available to the institution's researchers for their own preparation when undertaking the REC process. Making applications available must be done with the permission of the relevant project manager or principal investigator, where necessary. To support the creation of these repositories, we have developed a resource consisting of six hypothetical AI and data science REC submissions that can be used for training purposes.[217]

3. **Listing the institutional policies and guidance developed by the REC, such as policies outlining the research review process, self-assessment tools and societal impact assessments (see Recommendation 1).**

By including a full list of its policies, hubs can foster dialogue between different processes within research institutions. Documentation from across the organisation can be shared and framed in its importance for pursuing thoughtful and responsible research.

In addition to institutional guidelines, we suggest training hubs include

[217] See: Ada Lovelace Institute. (2022). *Looking before we leap: Case studies*. Available at: https://www.adalovelaceinstitute.org/resource/research-ethics-case-studies/

national, international or professional society guidelines that may govern specific kinds of research. For example, researchers seeking to advance healthcare technologies in the UK should ensure compliance with relevant Department of Health and Social Care guidelines, such as their guidelines for good practice for digital and data-driven health technologies.[218]

4. **Providing a repository of external experts in subject-matter domains who researchers and REC members can consult with.**

This would include a curated list of subject-matter experts in specific domains that students, staff and REC members can consult with. This might include contact details for experts in subjects like data protection law or algorithmic bias within or outside of the institution, but may extend to include lived experience experts and civil society organisations who can reflect societal concerns and potential impacts of a technology.

5. **Signposting to other pertinent institutional policies (such as compliance, data privacy, diversity and inclusion).**

By listing policies and resources on data management, sharing, access and privacy, training hubs could ensure researchers have more resources and training on how to properly manage and steward the data they use. Numerous frameworks are readily available online, such as the *FAIR Principles*,[219] promoting findability, accessibility, interoperability and reuse of digital assets; and DCC's compilation of metadata standards for different research fields.[220]

Hubs could also include the institution's policies on data labeller practices (if such policies exist). Several academic institutions have developed policies regarding MTurk workers that cover issues regarding fair pay, communication and acknowledgment.[221, 222] Some resources have even been co-written with input directly from MTurk workers. These

[218] Department of Health and Social Care. (2021). *A guide to good practice for digital and data-driven health technologies.* UK Government. Available at: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology

[219] Go Fair. *Fair principles*. Available at: https://www.go-fair.org/fair-principles/

[220] Digital Curation Centre (DCC). 'List of metadata standards'. Available at: https://www.dcc.ac.uk/guidance/standards/metadata/list

[221] Partnership on AI. (2021). *Responsible Sourcing of Data Enrichment Services*. Available at: https://partnershiponai.org/paper/responsible-sourcing-considerations/

[222] Northwestern University. (2014). *Guidelines for Academic Requesters*. Available at: https://irb.northwestern.edu/docs/guidelinesforacademicrequesters-1.pdf

resources vary from institution to institution, and there is a need for UK Research and Innovation (UKRI) and other national research institutions to codify these requirements into practical guidance for research institutions. One resource we suggest RECs tap into is the know-how and policies of human resources departments. Most large institutions and companies will already have pay and reward schemes in place. Data labellers and annotators must have access to the same protections as other legally defined positions.

The hub can also host or link to forums or similar communication channels that encourage informal peer-to-peer discussions. All staff should be welcomed into such spaces.

Examples of good practice

There are some existing examples of shared databases of AI ethics issues, including the Partnership on AI's *AI Incident Database* and Charlie Pownall's *AI, Algorithmic, and Automation Incidents and Controversies Database*. These databases compile news reports of instances of AI risks and ethics issues and make them searchable by type and function.[223, 224]

The Turing Institute's *Turing Way* offers an excellent example of a research institution's creation of shared resources for training and research ethics issues. For more information on the *Turing Way*, see Appendix 2.

Open questions

One pertinent question is whether these hubs should exist at the institutional or national level. Training hubs could start at the institutional level in the UK, and over time could connect to a shared resource managed by a centralised body like UKRI. It may be easier to start at the institutional level with repositories of relevant documentation, and spaces that foster dialogue among an institution's workforce. An international hub could help RECs coordinate with one another and external stakeholders through international and cross-institutional platforms, and explore the opportunity of inter-institutional review

[223] Partnership on AI. *AI Incidents Database*. Available at: https://partnershiponai.org/workstream/ai-incidents-database/

[224] AIAAIC. *AIAAIC Repository*. Available at: https://www.aiaaic.org/aiaaic-repository

standards and/or ethics review processing. We suggest that training hubs be made publicly accessible and open to other institutions, and that they are regularly reviewed and updated as appropriate.

Recommendation 5: Corporate labs must be more transparent about their decision-making and do more to engage with external partners

The problem

Several of our workshop participants noted that corporate RECs face particular opportunities and challenges in reviews of AI and data science research. Members of corporate RECs and research institutions shared that they are likely to have more resources to undertake ethical reviews than public labs, and several noted that these reviews often come at various stages of a project's lifecycle, including near publication.

However, there are serious concerns around a lack of internal and external transparency in how some corporate RECs make their decisions. Some researchers within these institutions have cited they are unable to assess what kind of work is acceptable or unacceptable, and there are reports of some companies changing research findings for reputational reasons. Some participants claimed that corporate labs can be more risk averse when it comes to seeking external stakeholder feedback, due to privacy and trade secret concerns. Finally, members of corporate RECs are made up of members of that institution, and do not reflect experiential or disciplinary expertise outside of the company. Several interview and workshop participants noted that corporate RECs often do not consult with external experts on research ethics or broader societal impact issues, choosing instead to keep such deliberations in house.

Recommendations

Corporate labs must publicly release their ethical review criteria and process

To address concerns around transparency, corporate RECs should publicly release details on their REC review processes, including what criteria they evaluate for and how decisions are made. This is crucial for public-private research collaborations, which risk the findings of public institutions being censored for private reputational concerns, and for internal researchers to know what ethical considerations they should factor into their research. Corporate RECs should also commit to releasing transparency reports citing how many research studies they have rejected, amended and approved, on what grounds, and some example case studies (even if hypothetical) exploring the reasons why.

Corporate labs should consult with external experts on their research ethics reviews, and ideally include external and experiential experts on members of their ethics review boards

Given their research may have significant impacts on people and society, corporate labs must ensure their research ethics review boards include individuals who sit outside the company and reflect a range of experiential and disciplinary expertise. Not including this expertise will mean that corporate labs lack meaningful evaluations of the risks their research can pose. To complement their board membership, corporate labs should also consult more regularly on ethics issues with external experts to understand the impact of their research on different communities, disciplines and sectors.

Examples of good practice

In a blog post from 2022, the AI research company DeepMind explained how their ethical principles applied to their evaluation of a specific research project relating to the use of AI for protein folding.[225] In this post, DeepMind stated they had engaged with more than 30 experts outside of the organisation to understand what kinds of challenges their research might pose, and how they might release their research responsibly. This offers a model of how private research labs might consult with external expertise, and could be replicated as a standard for DeepMind and other companies' activities.

In our research, we did not identify any corporate AI or data science research lab that has released their policies and criteria for ethical review. We also did not identify any examples of corporate labs that have experiential experts or external experts on their research ethics review boards.

[225] DeepMind. (2022). 'How our principles helped define Alphafolds release'. Available at: https://www.deepmind.com/blog/how-our-principles-helped-define-alphafolds-release

Open questions

Some participants noted that it can be difficult for corporate RECs to be more transparent due to concerns around trade secrets and competition – if a company releases details on its research agenda, competitors may use this information for their own gain. One option suggested by our workshop participants is to engage in questions around research practices and broader societal impacts with external stakeholders at a higher level of abstraction that avoids getting into confidential internal details. Initiatives like the Partnership on AI seek to create a forum where corporate labs can more openly discuss common challenges and seek feedback in semi-private ways. However, corporate labs must engage in these conversations with some level of accountability. Reporting what actions they are taking as a result of those stakeholder engagements is one way to demonstrate how these engagements are leading to meaningful change.

For funders, conference organisers and other actors in the research ecosystem

Recommendation 6: Develop standardised principles and guidance for AI and data science research principles

The problem

A major challenge observed by our workshop participants is that RECs often produce inconsistent decisions, due to a lack of widely accepted frameworks or principles that deal specifically with AI and data science research ethics issues. Institutions who are ready to update their processes and standards are left to take their own risks choosing how to draft new rules. In the literature, a plethora of principles, frameworks and guidance around AI ethics has started to converge around principles like transparency, justice, fairness, non-maleficence, responsibility and privacy.[226] However, there has yet to be a global effort to translate these principles into AI research ethics practices, or to determine how ethical principles should be interpreted or operationalised by research

[226] Jobin, A., Ienca, M. and Vayena, E. (2019). 'The global landscape of AI ethics guidelines'. *Nature*, 1, pp. 389–399. Available at : https://doi.org/10.1038/s42256-019-0088-2

institutions.[227] This requires researchers to consider diverse ethics interpretations and understanding in regions, other than Western societies, which so far have not adequately featured in this debate.

Recommendations

UK policymakers should engage in a multi-stakeholder international effort to develop a 'Belmont 2.0' that translates AI ethics principles into specific guidelines for AI and data science research.

There is a significant need for a centralised body, such as the OECD, Global Partnership on AI or other international body to lead a multinational and inclusive effort to develop more consistent ethical guidance for RECs to use with AI and data science research. The UK must take a lead on this and use its position in these bodies to call for the development of a 'Belmont 2.0' for AI and data science.[228] This effort must involve representatives from all nations and avoid the pitfalls of previous research ethics principle developments that have overly favoured Western conceptions of ethics and principles. This effort should seek to define a minimum global standard of research ethics assessment that is flexible, responsive to and considerate of local circumstances.

By engaging in a multinational effort, UK national research ethics bodies like the UK Research Integrity Office (UKRIO) can develop more consistent guidance for UK academic RECs to address common challenges. This could include standardised trainings on broader societal impact issues, bias and consent challenges, privacy and identifiability issues, and other questions relating to research integrity, research ethics and broader societal impact considerations.

We believe that UKRIO can also help in the effort for standardising RECs by developing common guidance for public-private AI research partnerships, and consistent guidance for academic RECs. A substantial amount of AI research involves public-private partnerships. Common guidance could include specific support for core language around intellectual property concerns and data privacy issues.

[227] Jobin, A., Ienca, M. and Vayena, E. (2019).

[228] Ferretti, A., Ienca, M., Sheehan, M. et al. (2021). 'Ethics review of big data research: What should stay and what should be reformed?'. *BMC Medical Ethics*, 22(1), pp. 1–13. Available at: https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-021-00616-4

Examples of good practice

There are some existing cross-national associations of RECs that jointly draft guidance documents or conduct training programmes. The European Network of Research Ethics Committee (EUREC) is one such example, though others could be created for other regions, or specifically for RECs who evaluate AI and data science research.[229]

In respect to laws and regulations, experts observe a gap in the regulation of AI and data science research. For example, the General Data Protection Regulation (GDPR) does provide some guidance for how European research institutions should collect, handle and use data for research purposes, though our participants noted this guidance has been interpreted by different institutions and researchers in widely different ways, leading to legal uncertainty.[230] While the UK Information Commissioner's Office (ICO) published guidance on AI and data protection,[231] it does not offer specific guidance for AI and data science researchers.

Open questions

It is important to note that standardised principles for AI research are not a silver bullet. Significant challenges will remain in the implementation of these principles. Furthermore, as the history of biomedical research ethics principle development has shown, it will be essential for a body or network of bodies with global legitimacy and authority to steer the development of these principles, and to ensure that they accurately reflect the needs of regions and communities that are traditionally underrepresented in AI and data science research.

[229] Dove, E.S. and Garattini, C. (2018). 'Expert perspectives on ethics review of international data-intensive research: Working towards mutual recognition'. *Research Ethics*, 14(1), pp. 1–25.

[230] Mitrou, L. (2018). 'Data Protection, Artificial Intelligence and Cognitive Services: Is the General Data Protection Regulation (GDPR) "Artificial Intelligence-Proof"?'. *Social Science Research Network*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3386914

[231] Information Commissioner's Office (ICO). *Guidance on AI and data protection*. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/

Recommendation 7: Incentivise a responsible research culture

The problem

RECs form one part of the research ethics ecosystem, a complex matrix of responsibility shared and supported by other actors including funding bodies, conference organisers, journal editors and researchers themselves.[232] In our workshops, one of the many challenges that our participants highlighted was a lack of strong incentives in this research ecosystem to consider ethical issues. In some cases, considering ethical risks may not be rewarded or valued by journals, funders or conference organisers. Considering the ethical issues that AI and data science research can raise, it is essential for these different actors to align their incentives and encourage AI and data science researchers to reflect on and document the societal impacts their research.

Recommendations

Conference organisers, funders, journal editors and other actors in the research ecosystem must incentivise and reward ethical reflection

Different actors in the research ecosystem can encourage a culture of ethical behaviour. Funders, for example, can create requirements that researchers conduct a broader societal impact statement of their research in order to receive a grant, and conference organisers and journal editors can encourage researchers to include a broader societal impact statement when submitting research. Conference organisers and journal editors can put in place similar requirements, and reward papers that exemplify strong ethical consideration. Publishers, for example, could potentially be assigned to evaluate broader societal impact questions in addition to research integrity issues.[233] By creating incentives for ethical reflection throughout the research ecosystem, ethical reflection can become more desirable and rewarded.

[232] Samuel, G., Derrick, G. E. and Van Leeuwen, T. (2019). 'The ethics ecosystem: Personal ethics, network governance and regulating actors governing the use of social media research data'. *Minerva*, 57(3), pp. 317–343. Available at: https://link.springer.com/article/10.1007/s11024-019-09368-3

[233] Ferretti, A., Ienca, M., Sheehan, M. et al. (2021). 'Ethics review of big data research: What should stay and what should be reformed?'. *BMC Medical Ethics*, 22(1), pp. 1–13. Available at: https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-021-00616-4

Examples of good practice

Some AI and data science conference organisers are putting in place measures to incentivise researchers to consider the broader societal impacts of their research. The 2020 NeurIPS conference, one of the largest AI and machine learning conferences in the world, required submissions to include a statement reflecting on broader societal impact, and created guidance for researchers to complete this.[234] The conference had a set of reviewers who specifically evaluated these impact statements. The use of these impact statements led to some controversy, with some researchers suggesting they could led to a chilling effect on particular types of research, and others suggesting difficulties in creating these kinds of impact assessments for more theoretical forms of AI research.[235] As of 2022, the NeurIPs conference has included these statements as part of its checklist of expectations for submission.[236] In a 2022 report, the Ada Lovelace Institute, CIFAR, and the Partnership on AI identified several measures that AI conference organisers could take to incentivise a culture of ethical reflection.[237]

There are also proposals underway for funders to include these considerations. Gardner and colleagues recommend that grant funding and public tendering of AI systems requires a 'Trustworthy AI Statement'.[238]

Open questions

Enabling a stronger culture of ethical reflection and consideration in the AI and data science research ecosystem will require funding and resources. Reviewers of AI and data science research papers for conferences and journals already face a tough task; this work is voluntary and unpaid, and these reviewers often lack clear standards or principles

[234] Ashurst, C., Anderljung, M., Prunkl, C. et al. (2020). 'A Guide to Writing the NeurIPS Impact Statement'. *Centre for the Governance of AI*. Available at: https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832

[235] Castelvecchi, D. (2020). 'Prestigious AI meeting takes steps to improve ethics of research'. *Nature*, 589(7840), pp. 12–13. Available at: https://doi.org/10.1038/d41586-020-03611-8

[236] NeurIPS. (2021). *NeurIPS 2021 Paper Checklist Guidelines*. Available at: https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist

[237] Canadian Institute for Advanced Research, Partnership on AI and Ada Lovelace Institute. (2022). *A culture of ethical AI: report*. Available at: https://www.adalovelaceinstitute.org/event/culture-ethical-ai-cifar-pai/

[238] Gardner, A., Smith, A. L., Steventon, A. et al. (2021). 'Ethical funding for trustworthy AI: proposals to address the responsibilities of funders to ensure that projects adhere to trustworthy AI practice'. *AI and Ethics*, 2. pp.1–15. Available at: https://link.springer.com/article/10.1007/s43681-021-00069-w

to review against. We believe more training and support will be needed to ensure this recommendation can be successfully implemented.

Recommendation 8: Increase funding and resources for ethical reviews of AI and data science research

The problem

RECs face significant operational challenges around compensating their members for their time, providing timely feedback, and maintaining the necessary forms of expertise on their boards. A major challenge is the lack of resources that RECs face, and their reliance on voluntary and unpaid labour from institutional staff.

Recommendations

As part of their R&D strategy, UK policymakers must earmark additional funding for research institutions to provide greater resource, training and support to RECs.

In articulating national research priorities, UK policymakers should mandate an amount of funding towards initiatives that focus on interdisciplinary ethics training and support for Research Ethics Committees. Funding must be made available for continuous, multi-stage research ethics review processes, and rewarding behaviour from organisations including UK Research and Innovation (UKRI) and UK research councils. Future iterations of the UK's National AI Strategy should earmark funding for ethics training and for the work of RECs to expand their scope and remit.

Increasing funding and resources for institutional RECs will enable these essential bodies to undertake their critical work fully and holistically. Increased funding and support will also enable RECs to expand their remit and scope to capture risks and impacts of AI and data science research, which are essential for ensuring AI and data science are viewed as trustworthy disciplines and for mitigating the risks this research can pose. The traditional approach to RECs has treated their labour as voluntary and unpaid. RECs must be properly supported and resourced to meet the challenges that AI and data science pose.

Acknowledgements

This report was authored by:

- Mylene Petermann, Ada Lovelace Institute
- Niccolo Tempini, Senior Lecturer in Data Studies at the University of Exeter's Institute for Data Science and Artificial Intelligence (IDSAI)
- Ismael Kherroubi Garcia, Kairoi
- Kirstie Whitaker, Alan Turing Institute
- Andrew Strait, Ada Lovelace Institute

This project was made possible by the Arts and Humanities Research Council who provided a £100k grant for this work. We are grateful for our reviewers – Will Hawkins, Edward Dove and Gabrielle Samuel. We are also grateful for our workshop participants and interview subjects, who include the following and several others who wished to remain anonymous:

- Alan Blackwell
- Barbara Prainsack
- Brent Mittelstadt
- Cami Rincón
- Claire Salinas
- Conor Houghton
- David Berry
- Dawn Bloxwich
- Deb Raji
- Deborah Kroll
- Edward Dove
- Effy Vayena
- Ellie Power
- Elizabeth Buchanan
- Elvira Perez
- Frances Downey
- Gail Seymour
- Heba Youssef
- Iason Gabriel
- Jade Ouimet
- Josh Cowls

- Katharine Wright
- Kerina Jones
- Kiruthika Jayaramakrishnan
- Lauri Kanerva
- Liesbeth Venema
- Mark Chevilet
- Nicola Stingelin
- Ranjit Singh
- Rebecca Veitch
- Richard Everson
- Rosie Campbell
- Sara Jordan
- Shannon Vallor
- Sophia Batchelor
- Thomas King
- Tristan Henderson
- Will Hawkins

Appendix 1: Methodology and limitations

This report uses the term **data science** to mean the extraction of actionable insights and knowledge from data, which involves preparing data for analysis, performing data analysis using statistical methods leading to the identification of patterns in the data.[239]

This report uses the term **AI research** in its broadest sense, to cover research into software and systems that display intelligent behaviour, which includes subdisciplines like machine learning, reinforcement learning, deep learning and others.[240]

This report relied on a review of the literature on RECs, research ethics and broader societal impact questions in AI, most of which covers challenges in academic RECs. This report also draws on a series of workshops with 42 members of public and private AI and data science research institutions in May 2021, along with eight interviews with experts in research ethics and AI issues. These workshops and interviews provided some additional insight into the ways corporate RECs operate, though we acknowledge that much of this information is challenging to verify given the relative lack of transparency of many corporate institutions in sharing their internal research review processes (one of our recommendations is explicitly aimed at this challenge). We are grateful to our workshop participants and research subjects for their support in this project.

This report contains two key limitations:

1. While we sought to review the literature of ethics review processes in both commercial and academic research institutions, the literature on RECs in industry is scarce and largely reliant on statements

[239] Provost, F. and Fawcett T. (2013). 'Data science and its relationship to big data and data-driven decision making'. *Big Data*, 1(1), pp. 51–59.

[240] We borrow from the definition used by the European Commission's High Level Expert Group on AI. See: European Commission. (2019). *Ethics guidelines for trustworthy AI*. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

and articles published by companies themselves. Their claims are therefore not easily verifiable, and sections relating to industry practice should be read with this in mind.

2. The report exclusively focuses on research ethics review processes at institutions in the UK, Europe and the USA, and our findings are therefore not representative of a broader international context. We encourage future work to focus on how research ethics and broader societal impact reviews are conducted in other regions.

Appendix 2: Examples of ethics review processes

In our workshops, we invited presentations from four UK organisations to share how they currently construct their ethics review processes. We include short descriptions of three of these institutions below:

The Alan Turing Institute

The Alan Turing Institute was established in 2015 as the UK National Institute for Data Science. In 2017, artificial intelligence was added to its remit, on Government recommendation. The Turing Institute was created by five founding universities and the UK Engineering and Physical Sciences Research Council.[241] The Turing Institute has since published *The Turing Way*, a handbook for reproducible, ethical and collaborative data science. The handbook is open source and community-driven.[242]

In 2020, *The Turing Way* expanded to a series of guides that covered reproducible research,[243] project design,[244] communication,[245] collaboration[246] and ethical research.[247] For example, the *Guide for Ethical Research* advises to consider consent in cases where the data is already available, and to understand the terms and conditions under which the data has been made available. The guide also advises to consider further societal consequences. This involves an assessment

[241] The Alan Turing Institute. 'About us'. Available at: https://www.turing.ac.uk/about-us

[242] The Turing Way Community et al. (2019). *The Turing Way: A Handbook for Reproducible Data Science*. Available at: https://the-turing-way.netlify.app/welcome

[243] The Turing Way Community et al. (2020). *Guide for Reproducible Research*. Available at: https://the-turing-way.netlify.app/reproducible-research/reproducible-research.html

[244] The Turing Way Community et al. (2020). *Guide for Project Design*. Available at: https://the-turing-way.netlify.app/project-design/project-design.html

[245] The Turing Way Community et al. (2020). *Guide for Communication*. Available at: https://the-turing-way.netlify.app/communication/communication.html

[246] The Turing Way Community et al. (2020). *Guide for Collaboration*. Available at: https://the-turing-way.netlify.app/collaboration/collaboration.html

[247] The Turing Way Community et al. (2020). *Guide for Ethical Research*. Available at: https://the-turing-way.netlify.app/ethical-research/ethical-research.html

of the societal, environmental and personal risks involved in research, and measures in place to mitigate these risks.

As of writing, the Turing Institute is working on changes to its ethics review processes towards a continuous integration approach based on the model of 'DevOps'. This is a term used in software development that involves a process of continuous integration and feedback loops across the stages of planning, building and coding, deployment and operations. To ensure ethical standards are upheld in a project, this model involves frequent communication and ongoing, real-time collaboration between researchers and Research Ethics Committees. Currently an application to RECs for ethics review is usually submitted after a project is defined, and a funding application has been made. However, the continuous integration approach covers all stages in the research lifecycle, from project design to publication, communication and maintenance. For researchers, this means considering research ethics from the beginning of a research project and fostering a continuous conversation with RECs, for example when defining the project, or so that RECs could offer support when submitting an application for funding. The project documentation would be updated continuously as the project progresses through various stages.

The project would go through several rounds of reviews by RECs, for example, when accessing open data, during data analysis or at the publication stage. This is a rapid, collaborative process where researchers incorporate the comments from the expert reviewers. This model ensures that researchers address ethical issues as they arise throughout the research lifecycle. For example, the ethical considerations of publishing synthetic data cannot be known in advance, therefore, an ongoing ethics review is required.

This model of research ethics review requires a pool of practising researchers as reviewers. There would also need to be decision-makers who are empowered by the institution to reject an ethics application, even if funding is in place. Furthermore, this model requires permanent specialised expert staff who would be able to hold these conversations with researchers, which also requires additional resources.

SAIL Databank

The Secure Anonymised Information Linkage (SAIL) Databank[248] is a platform for robust secure storage and use of anonymised person-based data for research to improve health, wellbeing and services in Wales. The data held in this repository can be linked together to address research questions, subject to safeguards and approvals. The databank contains over 30 billion records from individual-level population datasets from about 400 data providers, used by approximately 1,200 data users. The data is primarily sourced in Wales, but also England.

The data is securely stored, and access is tightly controlled through a robust and proportionate 'privacy by design' methodology, which is regulated by a team of specialists and overseen by an independent Information Governance Review Panel (IGRP). The core datasets come from Welsh organisations, and include hospital inpatient and outpatient data. With the Core Restricted Datasets, the provider reserves the right to review every proposed use of the data, while approval for the Core Datasets is devolved to the IGRP.

The data provider divides the data into two parts. The demographic data goes to a trusted third party (an NHS organisation), which matches the data against a register of the population of Wales and assigns each person represented a unique anonymous code. The content data is sent directly to SAIL. The two parts can be brought together to create de-identified copies of the data, which are then subjected to further controls and presented to researchers in anonymised form.

The 'privacy by design' methodology is enacted in practice by a suite of physical, technical and procedural controls. This is guided by the 'five safes' model, for example, 'safe projects', 'safe people' (through research accreditation) or 'safe data' (through encryption, anonymisation or control before information can be accessed).

In practice, if a researcher wishes to work with some of the data, they submit a proposal and SAIL reviews feasibility and scoping. The researcher is assigned to an analyst who has extensive knowledge of the available datasets and who advises on which datasets they need to

[248] See: https://saildatabank.com/

request data from, and which variables will help the researcher answer the questions. After this process, the researcher makes an application to SAIL, which goes to the IGRP. The application can be approved, rejected or recommendations for amendments made. The IGRP is comprised of representatives from organisations including Public Health Wales, Welsh government, Digital Health and Care Wales and the British Medical Association (BMA), and members of the public.

The criteria for review include, for example, an assessment of whether the research contributes to new knowledge, whether it improves health, wellbeing and public services, whether there is a risk that the output may be disclosive of individuals or small groups, and whether measures are in place to mitigate the risks of disclosure. In addition, public engagement and involvement ensures that a public voice is present in terms of considering potential societal impact, and who also provide a public perspective on research.

Researchers must complete a recognised safe researcher training programme and abide by the data access agreement. The data is then provided through a virtual environment, which allows the researchers to carry out the data analysis and request results. However, researchers cannot transfer data out of the environment. Instead, researchers must propose to SAIL which results they would like to transfer for publication or presentation, and these are then checked by someone at SAIL to ensure that they do not contain any disclosive elements.

Previously, the main data types were health data, but more recently, SAIL deals increasingly with administrative data, e.g. the UK Census, and with emerging data types, which may require multiple approval processes, and which can be a problem in terms of coordination. For example, data access that falls under the Digital Economy Act must have approval from the Research Accreditation Panel, and there is an expectation that each project will have undergone formal research ethical review, in addition to the IGRP.

University of Exeter

The University of Exeter has a central University Ethics Committee (UEC) and 11 devolved RECs at college or discipline level. The devolved RECS report to the UEC, which is accountable to the University Council (the governing body).[249] Exeter University also has a dual assurance scheme, with an independent member of the governing body also providing oversight.

The work of RECs is based on a single research ethics framework[250] which was first developed in 2013. This sets common standards and requirements, which also allows for flexibility to adapt to local circumstances. The framework underwent further substantial revision in 2019/20, which was a collaborative process with researchers from all disciplines with the aim to make it as reflective as possible of all discipline requirements while meeting common standards. Exeter also provides guidance and training on research ethics and as well as taught content for undergraduate and postgraduate students.

The REC operating principles[251] include:

- independence (mitigating conflicts of interest and ensuring sufficient impartial scrutiny; enhancing lay membership of committees)
- competence (ensuring that membership of committees/selection of reviewers is informed by relevant expertise and that decision-making is consistent, coherent, and well-informed; cross-referral of projects)
- facilitation (recognising the role of RECs in facilitating good research and support for researchers; ethical review processes recognised as valuable by researchers)
- transparency and accountability (REC decisions and advice to be open to scrutiny with responsibilities discharged consistently).

Some of the challenges include the lack of specialist knowledge, especially on emerging issues, such as AI and data science, new methods, or interdisciplinary research. Another challenge is information

[249] University of Exeter. (2021). *Ethics Policy*. Available at: https://www.exeter.ac.uk/media/universityofexeter/governanceandcompliance/researchethicsandgovernance/Ethics_Policy_Revised_November_2020.pdf

[250] University of Exeter. (2021). *Research Ethics Policy and Framework*. Available at: https://www.exeter.ac.uk/media/universityofexeter/governanceandcompliance/researchethicsandgovernance/Revised_UoE_Research_Ethics_Framework_v1.1_07052021.pdf

[251] University of Exeter (2021).

governance, e.g. ensuring that researchers have access to research data, as well as appropriate options for research data management and secure storage. Furthermore, ensuring transparency and clarity for research participants is important, e.g. active, or ongoing consent, where relevant. Secondary data use reviews include a risk-adapted or proportionate approach.

In terms of data sharing, researchers must have the appropriate permissions in place and understand the requirements of those. There are concerns about the potential misuse of data and research outputs, and researchers are encouraged to reflect on the potential implications or uses of their research, and to consider the principles of Responsible Research and Innovation (RRI) with the support of RECs. The potential risks with data sharing and international collaborations means that it is important to ensure that there is informed decision-making around these issues.

Due to the potentially significant risks of AI and data science research, Exeter University currently focuses on the Trusted Research Guidance issued by the Centre for Protection of National Infrastructure. Export Control compliance plays a role as well, but there is a greater need for awareness and training.

The University of Exeter has scope in the existing research ethics framework for setting up a specialist data science and AI ethics reference group (advisory group), which requires further work, e.g. how to balance the conflict between having a very specialist group of researchers reviewing the research, while also maintaining a certain level of independence. This would require more specialist training for RECs and researchers.

Furthermore, the University is currently evaluating how to review international and multi-site research, and how to streamline the process of ethics review as much as possible to avoid potential duplication in research ethics applications. This also requires capacity building with research partners.

Finally, improving the ability for reporting, auditing and monitoring plays a significant role, especially as the University recently implemented a new single, online research ethics application and review system.

About the Ada Lovelace Institute

The Ada Lovelace Institute was established by the Nuffield Foundation in early 2018, in collaboration with the Alan Turing Institute, the Royal Society, the British Academy, the Royal Statistical Society, the Wellcome Trust, Luminate, techUK and the Nuffield Council on Bioethics.

The mission of the Ada Lovelace Institute is to ensure that data and AI work for people and society. We believe that a world where data and AI work for people and society is a world in which the opportunities, benefits and privileges generated by data and AI are justly and equitably distributed and experienced.

We recognise the power asymmetries that exist in ethical and legal debates around the development of data-driven technologies, and will represent people in those conversations. We focus not on the types of technologies we want to build, but on the types of societies we want to build.

Through research, policy and practice, we aim to ensure that the transformative power of data and AI is used and harnessed in ways that maximise social wellbeing and put technology at the service of humanity.

We are funded by the Nuffield Foundation, an independent charitable trust with a mission to advance social well-being. The Foundation funds research that informs social policy, primarily in education, welfare and justice. It also provides opportunities for young people to develop skills and confidence in STEM and research. In addition to the Ada Lovelace Institute, the Foundation is also the founder and co-funder of the Nuffield Council on Bioethics and the Nuffield Family Justice Observatory.

Find out more:

Website: Adalovelaceinstitute.org
Twitter: @AdaLovelaceInst
Email: hello@adalovelaceinstitute.org

About the Alan Turing Institute

The Alan Turing Institute is the UK's national institute for data science and artificial intelligence.

The Institute is named in honour of Alan Turing, whose pioneering work in theoretical and applied mathematics, engineering and computing is considered to have laid the foundations for modern-day data science and artificial intelligence. The Institute's goals are to undertake world-class research in data science and artificial intelligence, apply its research to real-world problems, drive economic impact and societal good, lead the training of a new generation of scientists, and shape the public conversation around data and algorithms.

Find out more:

Website: turing.ac.uk
Twitter: @Turinginst

About the University of Exeter Institute for Data Science and Artificial Intelligence (IDSAI)

The University of Exeter founded the Institute for Data Science and Artificial Intelligence (AI) in 2018 to provide the focus for the University's data analytic capabilities, to develop innovative approaches to the use of data science and artificial intelligence and to create transformative impact. The University leads on data science and AI research relating to all aspects of modern society, covering the entire spectrum from interrogation and analysis of real world data to interpretation, visualisation, governance and communication. The University is a Partner of the Alan Turing Institute, the National Centre for Data Science and AI.

This report was co-authored by Dr Niccolo Tempini from the University of Exeter, a member of the Institute of Data Science and AI (where he co-leads the Data Ethics, Governance and Openness research theme) and of Egenis, the Centre for the Study of Life Sciences. Dr Tempini is also a Turing Fellow of the Alan Turing Institute. He has served in the Turing's own research ethics committee since Spring 2019.

Find out more:

Website: www.exeter.ac.uk

Preferred citation: Ada Lovelace Institute. (2022). *Looking before we leap: Ethical review processes for AI and data science research*. Available at: https://www.adalovelaceinstitute.org/report/looking-before-we-leap/

ISBN: 978-1-7397950-6-1