Monograph:
Woods, H.B. and Newman-Griffis, D. orcid.org/0000-0002-0473-4226 (2024) GRAIL: Codesigning responsible uses of AI in research funding and evaluation. Working Paper. RoRI Working Paper (13). Research on Research Institute

https://doi.org/10.6084/m9.figshare.27291459.v1

RoRI RESEARCH ON RESEARCH INSTITUTE

RoRI Working Paper No. 13

GRAIL: Codesigning responsible uses of AI in research funding and evaluation

Oslo workshop scoping paper

Helen Buckley Woods, Denis Newman-Griffis

October 2024

About the authors

Helen Buckley Woods is Senior Research Fellow in Metascience at RoRI and University College London. https://researchonresearch.org/team/helen-buckley-woods/, h.woods@ucl.ac.uk, https://orcid.org/0000-0002-0653-9803, @HelenBWoods.

Denis Newman Griffis is Research Fellow at RoRI and a Lecturer in Data Science at The University of Sheffield. https://www.sheffield.ac.uk/is/people/academic/denis-newman-griffis, d.r.newman-griffis@sheffield.ac.uk, https://orcid.org/0000-0002-0473-4226, @drgriffis

Acknowledgements

This scoping paper forms part of the GRAIL project of the Research on Research Institute (RoRI).

RoRI's second phase (2023–2027) is funded by an international consortium of partners, including: Australian Research Council (ARC); Canadian Institutes of Health Research (CIHR); Digital Science; Dutch Research Council (NWO); Gordon and Betty Moore Foundation *[Grant number GBMF12312; DOI 10.37807/GBMF12312];* King Baudouin Foundation; La Caixa Foundation; Leiden University; Luxembourg National Research Fund (FNR); Michael Smith Health Research BC; National Research Foundation of South Africa; Novo Nordisk Foundation *[Grant number NNF23SA0083996]*; Research England (part of UK Research and Innovation); Social Sciences and Humanities Research Council of Canada (SSHRC); Swiss National Science Foundation (SNSF); University College London (UCL); Volkswagen Foundation; and Wellcome Trust *[Grant number 228086/Z/23/Z]*.

Sincere thanks to all our partners for their engagement and support. We would also like to record our gratitude to members of the project Steering Group for advice and guidance at every stage: Jon Holm (RCN, chair), Tobias Philipp (SNSF), Anke Reinhardt (DFG), Freddy Navas Torres (ARC), Alexandra Apavaloae (SSHRC), Marieke van Duin (NWO), Carla Carbonell Cortés (LCF), Nicholas Hooper (UKRI), Justin Boylan-Toomey (Wellcome), Stephen Curry (RoRI), and former members Katrin Milzow (SNSF) and Gustav Petersson (SRC). We also thank all members of the wider GRAIL Working Group for their engagement and contributions to project discussions.

Responsibility for the content of RoRI outputs lies with the authors and RoRI CIC. Any views expressed do not necessarily reflect those of our partners. RoRI is committed to open research as an important enabler of our mission, as set out in our Open Research Policy. Any errors or omissions remain our own.

RESEARCH ON RESEARCH INSTITUTE

http://researchonresearch.org

Contents

1. AI in a changing research ecosystem

The current rapid pace of change in artificial intelligence (AI) technologies and their proliferation throughout daily life seems poised to profoundly transform research, with some asking if AI advances signal the 'end of science.'[1] AI and machine learning (ML) have been successfully used for decades as powerful tools in research,[2] but recent advances in the accessibility of AI tools and the availability of vast amounts of scientific data have accelerated the pace of change. AI technologies have become instrumental to major ongoing advances in fields from medicine[3] and

[1] Garisto, D. Don't Panic AI isn't coming to end scientific exploration. *Scientific American.* https://www.scientificamerican.com/article/dont-panic-ai-isnt-coming-to-end-scientific-exploration

[2] The Royal Society, (2024). *Science in the age of AI.* https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai-report.pdf

[3] Johnson, K et al. (2021). Precision Medicine, AI, and the Future of Personalized Health Care. *Clin Transl Sci*, 14:86-93. doi:10.1111/cts.12884

biology[4] to astrophysics,[5] and in 2024, two Nobel prizes were awarded in physics and chemistry for AI-driven research.[6,7]

For research funders, as stewards of the research ecosystem, AI and ML present unique pressures. AI is regarded by many as a "general purpose technology" with the capacity to boost productivity and transform working practices across entire economies, and with particular opportunities in knowledge-focused sectors such as research. AI approaches use knowledge about people and the world to guide analysis; ML is used to draw on real-world data as a source of this knowledge. AI and ML thus present significant opportunities for enhancing the knowledge work of research practice and research funding, and pose equally significant dilemmas and uncertainties around the changing nature of scientific knowledge, risks to reliability and validity of research, and the shape of good scientific practice in an "AI everywhere" world.[8, 9]

These topics are the focus of intense debate in many sectors, including higher education and research. The rise of consumer-grade generative AI technologies such as ChatGPT,[10] Gemini, Dall-E, and many others has also accelerated broader societal debates around the regulation of AI technologies, how best to deploy them, and the ethics of their use, raising profound questions ranging from the effects of AI on human interaction and creativity[11,12] to how diverse publics will

---

[4] Harvard Medical School, (2024). How Machine Learning Is Propelling Structural Biology. https://hms.harvard.edu/news/how-machine-learning-propelling-structural-biology

[5] Agarwal, A., & Nemade, A. (2023). AI-Enabled Black Hole Detection and Deflection: A New Frontier in Astrophysics. *Int. J. Res. Eng. Technol.*, *10*(09), 164-168. https://www.irjet.net/archives/V10/i9/IRJET-V10I924.pdf

[6] The Nobel Prize. (2024). https://www.nobelprize.org/prizes/chemistry/2024/press-release/

[7] The Nobel Prize. (2024). https://www.nobelprize.org/prizes/physics/2024/press-release/

[8] Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT and generative AI mean for science. *Nature*, *614*(7947), 214-216. https://www.nature.com/articles/d41586-023-00340-6

[9] Advance HE, (2024). Generative AI in Practice. https://www.advance-he.ac.uk/membership/collaborative-development-fund-2023-24/generative-ai-research-practice

[10] Hetler, A. (2024). What is ChatGPT? *TechTarget* https://www.techtarget.com/whatis/definition/ChatGPT

[11] Muldoon, J. (2024). Sex machina: in the wild west world of human-AI relationships, the lonely and vulnerable are most at risk. *The Conversation*. https://theconversation.com/sex-machina-in-the-wild-west-world-of-human-ai-relationships-the-lonely-and-vulnerable-are-most-at-risk-239783

[12] Krol, C. (2023). Nick Cave calls ChatGPT and AI songwriting "a grotesque mockery of what it is to be human" *New Musical Express*. https://www.nme.com/news/music/nick-cave-calls-chatgpt-and-ai-songwriting-a-grotesque-mockery-of-what-it-is-to-be-human-3381620

benefit or may be harmed by the use of AI in public services.[13] As generative AI has become more commonplace and augmented the other AI methods already in use in the research sector, specific calls have been made to build these technologies into research management & evaluation, including the practice of research funding.[14]

1.1. The GRAIL project

In 2023, RoRI launched the *Getting Responsible about AI and machine Learning in research funding and evaluation (GRAIL)* project[15] in partnership with an international consortium of research funders. GRAIL is filling the need for new research evidence on effective strategies for responsible and successful use of AI/ML in application contexts like research funding, and for practical guidelines and resources to guide funders in adopting best practices for designing, using, and evaluating AI/ML tools in their unique contexts.

GRAIL builds on a workshop series convened by RoRI and the Research Council of Norway in January 2021 to discuss the opportunities and challenges for AI/ML in the research funding setting.[16] This workshop and the subsequent discussions it engendered identified a clear need for broader and more in-depth work to develop guidance on best practices for responsibly integrating AI/ML technologies into decision-making processes in the research funding context. The GRAIL project aims to address these needs and to inform good practice and understanding for users of AI/ML within funding organisations, as well as academic and industry audiences designing AI/ML tools for practical impact.

[13] Shah, H. (2024). Tony Blair is wrong – AI will not magically solve our public services. Too many people might be left out of its revolution. *The New Statesman.* https://www.newstatesman.com/comment/2024/10/tony-blair-is-wrong-artificial-intelligence-ai-publ

[14] Nording, L. (2023). How research managers are using AI to get ahead. *Nature* https://www.nature.com/articles/d41586-023-04160-6

[15] Research on Research Institute. (2024). GRAIL Project. https://researchonresearch.org/project/grail/

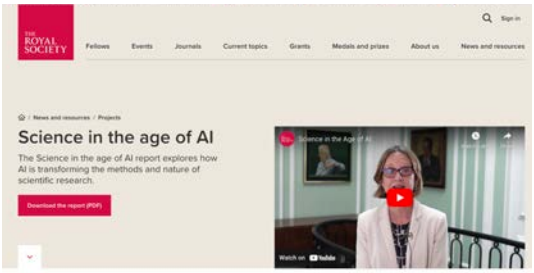
[16] Holm, J., Waltman, L., Newman-Griffis, D., & Wilsdon, J. (2022). Good practice in the use of machine learning & AI by research funding organisations: insights from a workshop series. Research on Research Institute. Report. https://doi.org/10.6084/m9.figshare.21710015.v1

1.2. Purpose of this scoping paper

This scoping paper has been prepared for the RoRI / Research Council of Norway GRAIL workshop in October 2024, as part of the ongoing GRAIL project continuing until the summer of 2025. The workshop will create a space for GRAIL partners and other research funders to hear more about the developments and discussions in the project, reflect on emerging research findings, and look ahead to next steps on the shape of AI in the research ecosystem.

At the workshop, we hope that everyone will feel free to share their experiences, ask questions and make new connections. We will gather notes from the sessions, reflect on the ideas generated and feed this into the next part of the project. This will include an expanded version of this scoping paper featuring summaries of emerging practice and deeper reflections on sector-wide opportunities and challenges for AI in research funding. This will also inform the GRAIL Research Funder's Handbook of AI [working title] to be released Summer 2025.

1.3. Use of AI in science and research practice

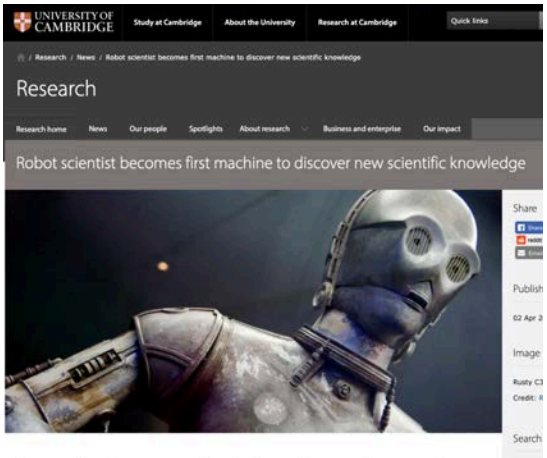


In their 2024 report 'Science in the age of AI,[17] led by Professor Alison Noble, The Royal Society investigated the use of AI technologies in research and the changes to research practice; including research skills, methodologies, and ethics. The report drew on the knowledge and experience of an expert working group, and activities with over a hundred scientists (such as workshops and interviews), to discover how AI was being used by the scientific community, and also learn about challenges, limitations and risks of using AI.

[17] The Royal Society. (2024). *Science in the age of AI.* https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai-report.pdf

> '...ensure we do not undermine progress...with black box systems which cannot be interrogated'
>
> Professor Alison Noble FRS[18]

In an analysis commissioned as part of the report, the use of AI was found to be widespread, with examples across all fields of science, technology, engineering and medicine. The physical sciences and medicine appeared to be most active in the use of AI applications. For example, using AI to extract information from large data streams and identify patterns within these stream[19] such as from the Large Hadron Collider. In health sciences AI was found to be used in numerous ways, as a public health tool, to support clinical decision making, and to improve training. These and other developments point to there being a change in scientific methods, with AI expediting processes and providing new ways to produce knowledge. The report identified the following key methodological changes in science: the use of 'deep learning' (such as processing substantial data sets and recognising patterns in data), bringing unstructured data together and making sense of it, large scale data simulations (for example simulating how molecules interact at an atomic level), the use of large language models (LLM) and natural language processing (NLP) to speed up text based tasks such as academic writing and reviewing literature, AI assisting in developing software code, and the automation of tasks including the use of 'robot scientists.'[20] Linked to these specific observations of how scientific methods are changing, the increasing dominance of big data research, the increasing importance of computing power and the integration of AI and Human intelligence and skills are cited as key overarching changes.



[18] The Royal Society. (2024). *Science in the age of AI.* (trailer video) https://youtu.be/o7Y9Jhz0WYA

[19] CERN. (n.d.). The Large Hadron Collider. https://home.cern/science/accelerators/large-hadron-collider

[20] University of Cambridge. (2009). Robot scientist becomes first machine to discover new scientific knowledge. https://www.cam.ac.uk/research/news/robot-scientist-becomes-first-machine-to-discover-new-scientific-knowledge

The report also indicates several challenges and limitations to the use of AI tools in science. For example, the nebulous nature of deep learning methods presents a challenge to the reproduction of studies and replication of results. There are also limitations to the use of AI technologies in writing tasks, for example in communicating values and recognising and articulating the complexity of thought inherent in scientific knowledge. In addition, the use of AI in science raises philosophical questions about the diminution of creativity in researchers' roles and how creative work is protected, for example the role of intellectual property in the use of data to train LLMs. In addition, will an increase in automation decrease methodological skills such as study design and development of hypotheses?

These ideas raise questions about how scientific work will be organised in the future, and whether future generations of researchers will be de-skilled and perhaps de-incentivised to take up a research career in the first place, if numerous tasks are automated and there is less outlet for creativity. The report also highlights the challenge to open science presented by the use of proprietary tools. That is, the lack of transparency in how these tools work and what data is used to train them, which then renders research results less trustworthy, as a key part of the process to reach them is opaque and unexplainable. This means results aren't replicable, and have reduced reliability, if the underpinning data is not known. However, evidence presented in the report suggests that having a fully open data model may be subject to illegal and immoral activity, so regulation is required. These, and other arguments against the use of AI in research have also been usefully summarised here.[21]

As part of a wider study[22] The International Science Council also presented a synthesis of evidence on the topic of AI in science, addressing the question 'What are the critical issues for the integration of artificial intelligence in science systems?'. The study used a systematic process to identify evidence using a high-precision keyword search and bibliometric techniques. In total, 317 documents were included, published between 2018 and 2023. The research team identified

[21] Miller, A. (2024). 'The top arguments against artificial intelligence in science'. ReHack, https://rehack.com/ai/arguments-against-artificial-intelligence/
[22] International Science Council. (2024). Preparing National Research Ecosystems for AI: strategies and progress in 2024. https://council.science/publications/ai-science-systems

45 core issues, which were then clustered into three key themes using the OECD's framework for technology governance (see below). An example from each theme is highlighted.

1. 'Research and development agenda setting, technology assessment, foresight and science advice.' For example, the effect on grant-giving. There is a danger that applicants that have access to AI have an unfair advantage. AI could become 'an inappropriate deciding factor' in grant allocation. There are also implications for evaluation panels, which are usually put together on disciplinary lines, whilst AI typically produces interdisciplinary results. Practically resolving this mismatch would be challenging.

2. 'Public engagement, science communication and public accountability.' For example, scientific integrity in research. The use of AI highlights strongly held values in science, and tensions between these longstanding values and use of AI can be seen in the binaries of 'openness vs. rigour; privacy and confidentiality vs. open science; massive data vs. high quality data; or explainability vs. "black box" results.' In addition, and as already reported, questions arise about the reliability and explainability of results, conflict of interest, accountability and ethical use. Ethics boards will be required to identify possible harms to human participants in AI based research. Furthermore, who is responsible for poor research practice if the fault lies with an AI?

3. 'Regulation, standards, private sector governance and self-regulation.' For example, law, regulation and policy. Where does an AI become liable for its work, rather than the person that created it? In addition, should AI generated products or outputs be protected by copyright law? If patenting is used instead, this may limit public access to AI created knowledge. Works mined for data are also protected under copyright rules. For example the EU protects data extracted from scientific databases for research purposes.

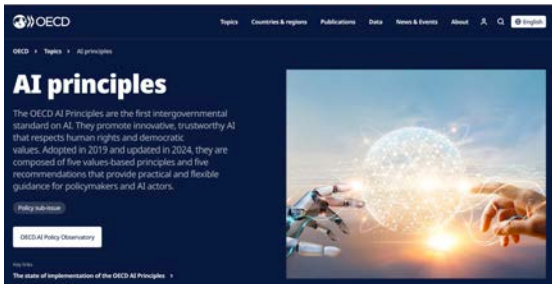
The literature review concludes with a reminder that although the key themes of their review reflect the overall impact of AI on research practice; the use and regulation of AI is often highly influenced by geographical context. For example, particular countries' aspirations for AI in growing their economy.

In this brief section, we have cited evidence to the widespread and increasing use of AI in research, its current uses, and potential challenges and limitations (which are becoming increasingly well charted). In the next section, we touch on the development of regulation and guidance for the use of AI, particularly in research systems.

1.4. AI regulation and guidance: a developing area

> 'Our central insight must always be that AI can be a great servant but would be a terrible master.'
>
> Kevin Brennan MP All-Party Parliamentary Group on Music Chair[23]

As with any fast paced developing technology, regulators are challenged to keep up with the speed at which new products are developed and deployed. This is particularly so, given the legacy of earlier technological developments, and the associated model prevalent in Silicon Valley of 'ask for forgiveness, not permission.'[24] This situation can be witnessed across employment sectors such as in the regulation of copyright and IP in creative industries.[25] In response to these challenges, as well as particular industries responding to the effects of AI, other pan-industry responses have been developed such as the EU AI Act[26], with regulations also emerging in the USA, UK and China amongst others.



The first intergovernmental set of principles on the use of AI were produced by the OECD[27] and adopted in 2019. Recently updated, they aim to

[23] https://www.ukmusic.org/wp-content/uploads/2024/04/APPG-AI-Report-Low-res.pdf

[24] https://www.telecoms.com/mobile-devices/silicon-valley-s-ask-for-forgiveness-not-permission-attitude-is-wearing-thin

[25] The All-Party Parliamentary Group on Music in association with UK Music. (2023) Artificial Intelligence and the Music Industry – Master or Servant? https://www.ukmusic.org/wp-content/uploads/2024/04/APPG-AI-Report-Low-res.pdf

[26] Madiega, T. (2021). Artificial intelligence act. *European Parliament: European Parliamentary Research Service*.https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[27] OECD AI principles (2024). https://www.oecd.org/en/topics/sub-issues/ai-principles.html

promote ‘innovative, trustworthy AI that respects human rights and democratic values’ and provide practical guidance for AI use. The principles provide a foundation for policy formulation and global collaboration. The five principles foster sustainable growth, protect human rights, promote transparency and explainability of AI use, systems should be safe and secure, and all stakeholders should be accountable, according to their role, to ensure effective systems, which are fit for purpose. In addition, risks such as harmful bias, safety, security, and privacy should be protected against, in addition to protection against risks to intellectual property and labour rights. Moreover, the guidance includes five recommendations for policy makers: long term investment in AI research & development (both public and private), to encourage the creation and use of trustworthy AI governments should support the development of appropriate digital ecosystems, these will include, for example, systems to support the ethical and legal use of data.

The broader environment of AI regulation and guidance continues to actively develop around the world. The 2023 EU AI Act imposed some of the first enforceable regulation on AI technologies and providers across European economies.[28] As the AI Act comes into force in 2024, it is spurring development of pan-sector Codes of Practice to guide development and implementation of AI technologies across a variety of European contexts.[29] UNESCO’s Recommendations on the Ethics of AI,[30] adopted in 2021, are helping to set global good practice in ethical use of AI. Emerging AI regulation in countries such as the US,[31] the UK,[32] and China[33] are representative of the rapidly evolving regulatory environment and development of good practice in individual states around the world.

[28] European Parliament, (2023). EU AI Act: first regulation on artificial intelligence. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[29] European Commission, (2024). Commission launches consultation on the Code of Practice for general-purpose Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/news/commission-launches-consultation-code-practice-general-purpose-artificial-intelligence

[30] UNESCO, (2023). Recommendation on the Ethics of Artificial Intelligence. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

[31] The White House, (2024). Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[32] UK Government, (2024). A pro-innovation approach to AI regulation: government response. https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response

[33] MIT Technology Review, (2024). Four things to know about China’s new AI rules in 2024. https://www.technologyreview.com/2024/01/17/1086704/china-ai-regulation-changes-2024/

1.5. Research funder responses to AI

In their recent **scoping review on 'artificial intelligence for research funding organisations'** (available as a preprint),[34] a team of researchers from the National Institute for Health and Care Research (NIHR) in the UK, explored the potential benefits and challenges that AI presents to funders. The study identified academic type literature (articles, opinion pieces, commentaries etc) and grey literature (blogs, reports, policy documents etc), using an iteratively developed search strategy, approved by an information specialist. Searches were limited to 2022-24, with no limit to study type, publication source, language or geographical area. Data sources included both journal databases and websites of research funding and professional organisations, such as research councils. Articles were included that focussed on the 'utility and potential of AI' and/or 'the considerations of risks of AI' or both. The review aimed to map the evidence rather than evaluate the effectiveness of any particular intervention.

> The application of AI cannot be overbearing, scary or to be feared. For AI to work, it must be made strategically simple.
>
> Blatch-Jones, A. J., Church, H., & Crane, K. (2024) p. 21).

In total, 122 articles were included in the review, with the majority of papers originating in Europe (54) or the Americas (49), with 85 articles being from peer reviewed journals. Numerous areas of research funding organisation's operations were identified as having potential for transformation, or improvement, by the use of AI technologies, such as 'data processes, administration, research insights, operational management, and strategic decision-making.' However, the evidence revealed a complex picture as organisations were at different stages in the overall adoption of AI into their organisations and were using AI for a variety of different tasks / areas of their business. **Key areas for funding bodies to consider for transformation of their work were reported under four themes**, with accompanying evidence summary statements from the underpinning evidence

[34] Blatch-Jones, A. J., Church, H., & Crane, K. (2024). Exploring the potential benefits and challenges of artificial intelligence for research funding organisations: a scoping review. *medRxiv*, 2024-09. https://www.medrxiv.org/content/10.1101/2024.09.26.24314280v1

1. ‘Data driven initiatives and industry 4.0 driving innovation’ (such as exploring data at scale and using data management frameworks.)
2. ‘The integration of AI technologies, systems, and tools’ (such as the importance of FAIR interoperability to ensure that the dovetailing of different data formats and systems and the importance of ‘Explainable AI’ to help reduce ambiguity and deliver more robust and reliable results across large amounts of data, increasing consumer confidence and understanding.)
3. ‘Optimisation and innovation for organisational and user efficiencies’ such as (the need to render the application of AI strategically simple in order to benefit from reduced bureaucracy in public sector workflows which in turn would increase the acceptance and use, and a reminder that the use of AI is not a replacement for human critical thinking.)
4. ‘Strategic direction and focus’ (such as the value of using data analytical tools and methodologies to gain useful insights into AI performance to inform evidence based decision making and such as the need for stable financial resources sustain long term use of AI.)

The review also highlighted the opportunities for funding organisations to manage and analyse large data sets to improve customer relationships, automate administrative tasks (such as using conversational AI for consumer- staff communications), and the use of predictive and performance analysis such as gamification tools to enhance HR practices such as training and recruitment.

> There is a tendency to believe that automated decisions are bias-free, when in reality they are subject to the intrinsic biases of the humans involved in AI modelling.
>
> Blatch-Jones, A. J., Church, H., & Crane, K. (2024) p. 25).

The review also synthesised 104 papers reporting **risks for funders using AI technologies**. These were reported under four themes:

1. **Support the AI readiness of organisations:** for AI to be a ‘data norm’ changes in personnel will be required to include ‘data scientists, librarians or archivists, data management experts and AI/ML

researchers', and similar barriers were reported by organisations already using AI, pertaining to 'AI knowledge and digital skills, data quality and data privacy and protection'.

2. **Support the AI readiness of data:** with data management being the most cited barrier to AI adoption, with the highest risk being data integrity, and the fundamental requirement and underestimated importance for data to be 'cleaned and verified before it can be fed into algorithms to prevent bias and errors.'
3. **Support accountability and fairness in AI:** such as, the need to involve the public and end-users in AI decisions to get a wide variety of perspectives. Biassed AI decisions can also lead to prejudice towards certain groups in society and an 'erasure of diversity from data'. There is also an erroneous belief that any automated decision is without bias, when in fact AI systems are subject to the human biases of the AI modellers who created them.
4. **Governance and ethical use of AI**: such as, GDPR now provides rules for information governance in AI and requires that human participation is required where AI is used in the processing of personal data, also, there are increasing reports of ethical failures in the use of AI and there are increased calls for organisations to introduce 'sound and transparent controls for the systems they use'.

Finally, the review also highlighted a prevalence of unanticipated evidence that strongly indicated: *'facilitation of shared AI efforts through collaboration and partnerships is important in achieving AI readiness and successful implementation in the research sector.'* (p. 29).

As reflected in this review, a range of research funders, across the globe, are currently exploring the use of AI approaches, with a particular focus on ML, to address diverse motivations and strategic aims. The need for more efficient operation of selection processes and grant management appear as the current most prominent drivers for the adoption of new technologies. Several funders are also looking at how AI methods can help increase the effectiveness of their funding, e.g. to ensure that it actually meets the goals set by boards and governments. Recent examples include the algorithm developed to check the quality of peer review at the Swiss National Science Foundation (SNSF) and the use of algorithms to secure more consistent tagging of grants at RCN, among others. AI tools may also help with identifying inherent biases in peer

review so that these can be discussed and corrective measures applied if deemed necessary. The label of "AI" encompasses a wide variety of approaches, all drawing on a common theme of using knowledge about the world to guide computation and analysis. This knowledge may come in the form of expert-built rules and resources, community-sourced knowledge bases, or real-world data, among others. Both symbolic reasoning and data-driven machine learning have significant potential in a setting combining domain expertise and large-scale data, as is the case in research funding. While less well-developed in the literature, hybrid approaches to combine expert systems with ML may offer further benefits in targeted settings.

However, recent findings, such as in the NIHR review presented above,[35] have demonstrated that many popular AI technologies, especially ML-based approaches, can reflect and even amplify social biases such as racism, sexism, and ableism.[36] These approaches also often depend on subtler limiting assumptions about the types of varieties of data they are used to analyse,[37] as well as the assumption that past patterns are indicative of desired future outcomes.[38] Progress on addressing these issues has mostly focused on the design of core research tools, with comparatively little effort devoted to the application of AI/ML tools to specific decision contexts. In developing good practice for research funders, who aim to serve both the public interest and the advancement of ethical and responsible science, it is therefore necessary to focus on practical guidance that speaks to ethical practice in real-world decision making.

Having presented the context of AI in research and more specifically in science funding, the next sections will focus on the GRAIL project, beginning with the project methodology and methods.

---

[35] Blatch-Jones, A. J., Church, H., & Crane, K. (2024). Exploring the potential benefits and challenges of artificial intelligence for research funding organisations: a scoping review. *medRxiv*, 2024-09. https://www.medrxiv.org/content/10.1101/2024.09.26.24314280v1

[36] Hoffman et al., (2024). AI generates covertly racist decisions about people based on their dialect. *Nature*, 633:147-154. https://doi.org/10.1038/s41586-024-07856-5

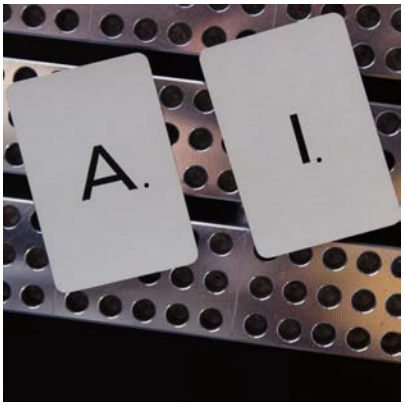
O'Connor and Liu, (2024). Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. *AI & Society*, 39:2045-2057. https://doi.org/10.1007/s00146-023-01675-4

Newman-Griffis et al., (2023). Definition drives design: Disability models and mechanisms of bias in AI technologies. *First Monday*, 28(1). https://doi.org/10.5210/fm.v28i1.12903

[37] Dalmer et al., (2024). Configuring data subjects. In *Dialogues in Data Power* (Jarke, J., Bates, J., eds). 10-30. https://doi.org/10.51952/9781529238327.ch001

[38] Sahiner et al., (2023). Data drift in medical machine learning: implications and potential remedies. The British Journal of Radiology, 96(1150), p.20220878.

2. Codesigning responsible uses of AI in research funding and evaluation



Our methodology in the GRAIL project brings together thirteen funder partners, including:

Nine government research funders – Australian Research Council, Austrian Science Fund, Dutch Research Council, German Research Foundation, Research Council of Norway, Social Sciences and Humanities Research Council of Canada, Swedish Research Council, Swiss National Science Foundation, and UK Research & Innovation; and:

Four philanthropic funders – "la Caixa" Foundation, Novo Nordisk Foundation, Volkswagen Foundation, and Wellcome Trust.

Each funder is at a different point in exploring, adopting, deploying, and evaluating AI/ML approaches in their work. The GRAIL project functions as a space for these funders to come together to exchange knowledge and experiences—to celebrate successes, but also to share confusions and challenges that are vital to developing best practice but often challenging to discuss in public fora.

The GRAIL project team works co-productively with funder partners as a core practice of the project. GRAIL is led by a Steering Group of key partner representatives, and discussions and feedback are sought from a wider Working Group including relevant staff across all partners in addition to the project team.

Together, we are working to accomplish three aims:

1. **Build cross-funder knowledge** and seed shared practice around the responsible use of AI/ML in research funding and evaluation;

2. **Produce new insights and recommendations** on how research funders and other public bodies can effectively navigate the *sociotechnical* systems and processes[39] required to bring AI and ML technologies to bear effectively in their work, whilst maintaining the highest standard of ethics and social responsibility.

3. **Understand how funders are currently using AI/ML** in their work and identify opportunities for synergy and/or application of shared practice.

2.1. GRAIL workshop series

The core activity of GRAIL, addressing ***Aim 1 (Build cross-funder knowledge)*** is a series of **virtual, co-productive workshops** held with the GRAIL Working Group of staff from partner funding organisations.

Workshops may be led by the project team or hosted by one of the project partners to focus on a particular topic of interest to them. Each workshop is organised around a specific area of AI/ML application in research funding and/or a particular challenge for research funders in effectively and ethically managing AI/ML use. **Table 1** lists the topics of the GRAIL workshops held to date as of October 2024.

Workshops are **closed sessions operating under Chatham House rules, with limited external data sharing and a strong focus on protected conversation** with the freedom to discuss challenging topics and experiences. The host organisation for each workshop may invite external presenters and additional guests as relevant, with all attendees agreeing to abide by a co-produced set of ground rules established at the outset of the project.

[39] I.e., systems and processes that combine technical implementation with organisational and social context. See Whitworth, B., 2009. A brief introduction to sociotechnical systems. In *Encyclopedia of Information Science and Technology*, Second Edition (pp. 394-400). IGI Global.

Table 1. *GRAIL workshops held as of October 2024.*

Month	**Topic**
June 2023	ChatGPT/Generative AI and the research funding ecosystem
November 2023	AI and research evaluation
January 2024	GRAIL & AI guidance
February 2024	Natural language processing in research funding
April 2024	Policy and responsible use of AI/ML
June 2024	Applying AI/ML tools to improve research assessment
July 2024	Guidelines for the use of Generative AI in research funding processes
September 2024	Responsible AI principles for research funders

Workshop discussions are noted by the project team, with anonymised versions of notes produced for sharing to attendees after the workshop. Attendees are also invited to complete an anonymous feedback survey reflecting on the presentations/activities and discussion topics in each workshop and highlighting specific learning to carry forward.

Workshop notes are reviewed by the project team to identify emergent themes and recurring topics pertaining to funder use of AI/ML technologies, and to pinpoint specific needs or recommendations for using these technologies effectively and ethically.

2.2. Research funder's AI handbook

The key output of the GRAIL project, drawing directly on the workshop series, is a **Research Funder's Handbook of AI** [working title], addressing ***Aim 2 (Produce new insights and recommendations).***

The GRAIL handbook follows the co-productive model of the project and is being written with joint input from the project team and the partner funding organisations. The experiences, analyses, and recommendations in the handbook draw on the discussions shared in the GRAIL workshop series, the professional experiences of the project team and partner funders, and the rich space for cross-sector discussion created in the GRAIL project.

The aim of the GRAIL handbook is to **provide tangible, reusable guidance on using AI/ML techniques that helps funders build resilience to the changing technological landscape.** The handbook focuses on the key *processes and considerations* involved in applying AI/ML methodologies in the research funding context, drawing on Newman-Griffis' AI Thinking framework.[40]

Table 2 gives the overall structure (subject to change) and goals of the handbook as of October 2024. Development of the handbook will continue with collaborative dialogue between the project team and the GRAIL partners into 2025, with anticipated publication in June 2025.

Table 2. *Outline of GRAIL AI handbook (subject to change).*

Chapter	**Description**
Introduction	Establish handbook focus on the context of AI change in the research funding and evaluation ecosystem.
Funder contexts	Situate the handbook in broader discourses around technological change and responses in science policy to AI.

[40] Newman-Griffis (2024). AI Thinking: A framework for rethinking artificial intelligence in practice. *OSF Preprints*. https://doi.org/10.31219/osf.io/7xtz2

AI/ML implementations	Outline the steps and key considerations of the process that goes into each defined use of AI/ML.
Management challenges	Discuss key challenges and considerations for funders in managing AI/ML as an organisational competency.
Case studies	Illustrate examples of how funders have gone about implementing and managing AI/ML in practical use cases.

2.3. Surveys on AI/ML applications

Finally, we are working to better understand the global landscape of current AI/ML use in funding organisations. We are doing so through **two surveys**, addressing ***Aim 3 (Understand how funders are currently using AI/ML).***

The first survey is being administered by the AGORRA project in RoRI Phase 2, and is exploring attitudes and practices towards responsible research assessment (including the use of AI) amongst Global Research Council members.

The second survey is being administered within the GRAIL project itself, to include documentation of current, past, and anticipated uses of AI/ML within the GRAIL funder partners.

Together, these surveys will provide an initial mapping of range of AI/ML methodologies and applications currently being explored by funders, as well as key considerations in their implementation and management for the GRAIL AI guidance to help support.

3. Facilitators and barriers in managing AI/ML - lessons from the GRAIL project

Discussions in the GRAIL workshop series have been highly generative for both the project team and participating partners. Whilst discussions are still ongoing as the workshop series continues into 2025 and we bring together the workshops with data collected on current AI/ML applications, clear themes are already emerging to inform planning and best practice for using AI/ML in research funding and evaluation.

3.1. AI guidance that works in context

GRAIL discussions have repeatedly highlighted that **AI guidance must be pragmatic and fit for purpose.** Whilst AI policies and guidelines have proliferated in the last few years as AI technologies have become more mainstream, few of these documents provide **specific and actionable guidance** that guide complex organisations in implementing good AI practice. AI guidance that works must not only speak to process as well as principle, but must be **grounded in current and near-term technologies** rather than hyperbolic expectations of unproven future AI capabilities.

AI guidance must also be **adaptable to diverse audiences.** As described further below, implementing and managing AI is a complex process bringing together many different components of an organisation. To support this process, cross-cutting AI guidance (such as what GRAIL aims to produce) must focus on **general, actionable principles that can be adapted** to specific audiences and contexts, both within and external to AI-using organisations.

AI implementation is always local, and must be adapted to different regulatory environments, business processes, knowledge cultures, and more. Enabling this adaptation requires recognising the different dimensions of implementation and management that diverse AI

stakeholders must be sensitive to, and the practical challenges that arise in day-to-day AI use. As a result, AI guidance must be **co-produced with AI users**, as we have adopted in GRAIL.

Bringing AI guidance into practice is a multilayered process, as external best practices are adapted to specific contexts and disseminated to diverse stakeholders throughout an organisation. One emerging model for achieving this is through the work of **local champions for AI best practice**. Knowledgeable individuals, often with some familiarity with AI and lived experience in their organisational structure and culture, are often the best placed to identify how AI guidance should be adapted and might be disseminated. Champions can then be the first in a **series of mediation layers** to bring AI best practice to different audiences, using approaches discussed further below.

AI guidance and its adaptation must be **sensitive to the broader contexts surrounding AI use**. For example, many research funders are already operating in a broader policy context of public sector digitisation, and integration of AI/ML must be approached within the specific pressures and opportunities this creates. In these cases, public organisations investing in new IT infrastructure may be more able to meet computing requirements for AI systems, but focus on data transparency policies may be higher pressure than adopting new technologies.

Finally, producing and disseminating AI guidance must be **sensitive to, but not dominated by, current generative AI trends.** The rise of consumer-facing generative AI technologies since 2022 has dramatically changed the public discourse in AI, and where AI use might have been perceived as a risky liability in organisations before it is often now seen as an asset. Many funders are now also engaged in developing their own **guidance**[41] **and policies**[42] **on generative AI in research funding**. However, the fundamentals of implementing and managing AI pertain to broader use cases and technologies than generative AI alone, and must be approached in terms of a **broader landscape of AI methods and technologies.**

[41] NWO publishes preliminary guidelines for the use of AI (January 2024). https://www.nwo.nl/en/news/nwo-publishes-preliminary-guidelines-for-the-use-of-ai

[42] DFG Formulates Guidelines for Dealing with Generative Models for Text and Image Creation (September 2023). https://www.dfg.de/en/news/news-topics/announcements-proposals/2023/info-wissenschaft-23-72

3.2. Strategies for mobilising AI/ML in funding organisations

Discussions with GRAIL partners show a clear consensus that the **primary factor for success in adopting effective and ethical AI/ML in funding organisations is a cross-competency approach**. AI/ML are tools for helping to **learn from data to do things better or differently**—so using AI/ML in the work of research funding or evaluation is not just a technical task, but also one that involves data managers, process owners, scientific officers, and others across the organisation.

There is not just one way to **bring people together across competencies and roles** around the AI/ML implementation and management. Whether working as councils of department heads or team leads, or via internal collaboration between teams, funders must create **regular, collaborative spaces** for diverse stakeholders to come together on exploring and applying AI as part of a funder's business process or research. **Collaborative workshops** that bring together technical and data teams with potential users to explore applications of interest can help identify new approaches or identify specific utility for methods with clear potential. Most importantly, AI/ML leaders must **ensure that AI users feel involved** and that their input is valued in a collaborative approach.

While collaborative work is necessary to bring together the diverse stakeholders involved in using AI in funding organisations, this work is often best **led by data teams who serve as local champions.** Data teams have the technical expertise to be able to recognise 'shared DNA' between different problems or requests, and build more robust approaches that can be used in common across multiple settings. **User engagement workshops** can help ground this process in practical user needs, and **value mapping exercises** can help identify potential high-value applications of AI and prioritise for low-hanging fruit.

Effort on AI implementations must be paired with **getting buy-in on AI/ML** from stakeholders throughout the organisation. A key strategy for building value for AI/ML and ensuring that its use stays on target is to **focus on the problems to be addressed, not the AI to be used.** A

problem-oriented focus helps reduce the risk of being carried away by hyperbolic AI discourse, whilst providing a clear path to action in response to the pressures of adopting AI/ML. An internal **AI board**, tasked with discussing problems, approaches, and anticipated value, can help smooth communication across the organisation and provide much-needed clarity to upper management.

Finally, mobilising AI/ML in funding organisations is often a significant change, and must be **paired with careful change management.** Many staff in funding organisations will need training in basic AI literacy and skills to bring them up to speed, whether or not they are directly using AI/ML in their work. With the high-pressure environment surrounding AI currently, many people struggle to talk about AI, for lack of understanding or fear of not being current on the latest developments. To counteract this, funders must **create space for uncertainty and mutual learning** around AI, for staff to navigate changes together.

3.3. Implementation & management of AI/ML

Bringing AI methods into practice requires tackling a wide range of integration and implementation challenges. First and foremost, **effective AI is a matter of data**. In an ML-driven AI environment, it is often said that "without data, there is no AI"[43] – for a heavily knowledge-based environment like research funding, this is better stated as "**without good data, there is no useful AI**." But good data is a matter of context and goals: what works for one funder's processes and guiding principles may be inappropriate for another funder's approach. Funders must therefore ensure that **data used for AI are context-sensitive** and that **staff working with AI tools have sufficient data literacy** to guide their AI understanding.

In addition to selection of data, **data use with AI must be governed by data policies and good, responsible practice.** Funders operate within complex data provenance and data governance structures, and bear responsibility to ensure that their use of AI is aligned with policies on intellectual property, confidentiality, negotiated agreements, and other terms. Funders must also **set expectations for working in sensitive data contexts**: some external AI tools will have Terms

[43] What do we mean by "without data, there is no AI"? The Open Data Institute (2023). https://theodi.org/news-and-events/blog/what-do-we-mean-by-without-data-there-is-no-ai/

of Use that conflict with institutional data policy (e.g., saving of data entered into a web form), and inter-funder collaboration on AI/ML tools may be limited.

Implementing AI systems in practice involves **navigating familiar IT management and integration challenges**. For example, funders need good information management practices for resources such as source code and trained AI models, in addition to the data those models were trained with. To be useful, AI systems must also be **integrated with existing tools, software, and workflows**, with integration layers often making or breaking the utility of an AI/ML solution. Notably, for funders first exploring use of AI/ML in their work, there are significant **IT infrastructure requirements** that must be considered, including not only data storage but the development of integration layers and Trusted Research Environments for secure experimentation.

Just as there is no single best source of data for AI/ML use, funders must **explore a variety of AI/ML methodologies**. In many cases, simpler, cheaper models may match or outperform complex models–even cutting-edge Large Language Models. Funders therefore need the ability to empirically assess and compare different methodologies, attending not only to task performance (e.g., accuracy of topic classification) but also to the data requirements involved, alignment of the methodology with intended users, and computation and energy demands for model application.

Perhaps the most important aspect of AI implementation is that **AI is always context-sensitive – there is no 'one-size-fits-all' solution.** The effectiveness of different AI approaches will be different across the specific data and needs of different funders, and will also vary for individual users and applications. A practical AI/ML system is therefore one that **offers flexibility**: 'off-the-shelf' commercial solutions which cannot be adapted to specific organisational contexts are simply discarded or avoided entirely. User-level flexibility in terms of model selection, performance metric to optimise for (e.g., improving diversity of a reviewer pool vs topical fit), and practicalities of use are essential for ensuring AI use delivers value.

3.4. Assessing and managing AI value

When implementing AI/ML use in practice, as with any other intervention, the key question eventually becomes: **does it work?** However, evaluating AI use is not a straightforward process, and needs for AI evaluation face pressure of time and expectations that make traditional evaluation studies not fit for purpose.

AI implementation is an iterative process of testing, improvement, and re-testing. To find the stopping point and determine whether or not to put a prospective AI technology into practice, funders must be able to **determine if an AI implementation is good enough to use**. As with other aspects of AI implementation, this depends heavily on the context where AI is being used and the purpose it is put to – what is good enough when suggesting potential reviewers may not be when prioritising applications for funding. In addition to accuracy, **AI systems must also be reliable** to justify their continued use; however, there are as yet no clear definitions of AI reliability, nor strategies to measure it. Reliability must also therefore be explored and assessed by each individual funder according to their standards.

The **external validity of AI systems** is necessary for them to be adopted and useful, but is often difficult to assess. The best strategies for assessing validity rely on working with scientific officers, as the primary users of AI systems in research funding organisations, and performing **user acceptance testing** to determine if AI systems are seen as fit for purpose. These assessments must take into account practical aspects of use, such as technology infrastructure and data availability, as well as performance in controlled experiments.

AI use always involves tradeoffs. **Risk of bias** is essential to monitor for and prepare against, such as through data cleaning and selecting for appropriately-representative training data for ML, but also needs to be compared against known human biases in research funding and assessment processes. AI technologies lack understanding and attend to different nuances of language and data than humans, which may help avoid some biases whilst raising others. **Pre-development workshops** to explore anticipated effects of AI use can help to identify potential issues early on, including risks that have little to do with the use of AI itself.

The benefits of AI use are often presented in terms of **efficiency**. However, GRAIL discussions have highlighted the potential for unexpected benefits of AI use, in **gaining insight into funding and assessment processes** and helping to refine them.[44] The role of AI and algorithmic systems is to expand the options available and make more information accessible to users: assessing AI impacts should thus focus on what will be most beneficial to users, including scale as well as quality.

Assessing and articulating the value of AI/ML use also requires **expectation management of AI.** With the current levels of excitement around AI, hyperbolic claims abound and expectations of new AI systems may be unrealistic. **Adapting AI solutions to specific contexts**, and transferring the learning of ML models, is a difficult task and not guaranteed to work. AI use may also encounter a wide range of attitudes: some users are eager to use AI for all aspects of their work, whilst others find any flaw to indicate failure of the entire AI enterprise. AI teams must work across these diverse attitudes and experiences of AI, and focus on the problems AI systems will be used to tackle and the expected risks and benefits.

3.5. Emerging uses of AI/ML in research funding and evaluation

Whilst survey data collection is still ongoing, GRAIL discussions have already highlighted a number of important use cases for AI/ML in the work of research funding organisations:

AI is being explored by many funders for **improving reviewer matching**, to help identify a wider variety of more appropriate expert reviewers. This is particularly salient for the growing portion of research that is interdisciplinary, and thus requires crossing multiple reviewer pools to assemble relevant expertise for review.

[44] Holm et al. Big data for big investments: Making responsible and effective use of data science and AI in research councils. In *Artificial Intelligence and Evaluation*. (2024). https://doi.org/10.4324/9781003512493

Topic classification and tagging of proposals is another key area of AI/ML application. Automated classification can significantly enhance the manual tags assigned by researchers and/or funder staff, and enrich the relationships between proposals and fields.

Generating non-expert descriptions of research plans and outputs is a highly valuable component of many funders' mission to inform the public about current research. Large language models provide significant benefits in generating these descriptions more easily and effectively than by hand.

Finding similar funding applications is valuable for identifying emerging directions and potential synergies, as well as flagging potential duplicate applications.

Mining funded outputs that are not appropriately linked to research awards brings significant benefit to evaluation of research and funding programmes; automation with AI is making this task much easier to perform at scale.

These examples illustrate some of the present and emerging ways in which AI and ML are being used in research funding. As our data collection continues, we will develop a fuller picture of the types of use cases for AI/ML funders are exploring around the world, and the key considerations that have informed that use.

4. Next steps with RoRI's GRAIL project

This scoping paper was prepared for the GRAIL workshop held jointly by RoRI and the Research Council of Norway in Oslo on 31st October 2024. It outlines the context, goals, and current status of the GRAIL project on responsible AI and machine learning in research funding and evaluation, and highlights emerging findings from the ongoing project.

Each of the three components of the GRAIL project will continue to develop as we work towards the conclusion of the project in June 2025.

1. The **GRAIL workshop series** has included eight workshops to date, and will continue into 2025 with six further planned workshops addressing topics such as "Human in the loop,"

2. Co-productive development of the **GRAIL AI handbook** is continuing apace, with input from both the project team and partner funders continuing into 2025, with reflections in and responses to the discussions in the workshop series.

3. **Survey data collection** is continuing via the GRC survey on responsible research assessment (in collaboration with RoRI's AGORRA project) and collection of example AI/ML use cases among GRAIL partner funders.

Discussions to date have surfaced several major themes for co-developing effective guidance to support responsible use of AI/ML for funders.

1. To be effective, any **guidance must be pragmatic and context sensitive**.
2. **Mobilisation of resources, skills, and will to adopt AI/ML** effectively requires working collaboratively across competencies, and may be supported by a variety of different strategies.
3. **AI implementation** requires good data, integration with organisational process, policy, and other IT systems, and requires experimentation to find the best approaches for each context.

4. **Assessing the value and impact of AI/ML methods** requires addressing practical considerations of when systems are 'good enough' and sufficiently reliable to use, assessment of external validity and potential risks, and expectation management with internal and external stakeholders.
5. **Funders are exploring the use of AI/ML** to inform thorny challenges of scale and efficacy in core functions of research funding and evaluation, but best practice and effective strategies are still evolving.

Our meeting in Oslo will provide the opportunity to dive deeper into these themes and reflect on partner experiences with AI and with the ongoing work of the GRAIL project. Following the Oslo workshop, the GRAIL project team will analyse the discussions and topics that arise, and the project Steering Group will meet to reflect on the ideas that emerge from the meeting. We will take this learning forward into the development and finalisation of the GRAIL Handbook for launch in June 2025, and disseminate our findings and the emerging best practices from the GRAIL project to both academic and funder audiences.

http://researchonresearch.org