RoRI RESEARCH ON RESEARCH INSTITUTE

Funding by *Algorithm*

A handbook for responsible uses of AI and machine learning by research funders

Denis Newman-Griffis, Helen Buckley Woods, Youyou Wu, Mike Thelwall and Jon Holm

Acknowledgments

This report is published by the Research on Research Institute (RoRI) in partnership with the Research Council of Norway (RCN).

For citation: Newman-Griffis, D., Woods, H.B., Wu, Y., Thelwall, M., & Holm, J. (2025). *Funding by Algorithm - A handbook for responsible uses of AI and machine learning by research funders* (ISBN 978-1-7397102-2-4). June 2025. **DOI 10.6084/m9.figshare.29041715**

This publication 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 GRAIL partners for their engagement and support: the Austrian Science Fund (FWF); Australian Research Council (ARC); Dutch Research Council (NWO); German Research Foundation (DFG); 'La Caixa' Foundation (LCF); Novo Nordisk Foundation (NNF); Research Council of Norway (RCN); Research England/UKRI; Social Sciences and Humanities Research Council of Canada (SSHRC); Swedish Research Council (SRC); Swiss National Science Foundation (SNSF); Volkswagen Foundation (VWF); and Wellcome Trust.

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), Alexander Bondarenko (VWF), 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.

Thanks also for invaluable contributions to the case studies in Part 5 from:

Gabriel Okasa and Anne Jorstad
Swiss National Science Foundation

Carla Carbonell Cortés
'La Caixa' Foundation

Emma Olsen and Sandra Schluttenhofer
Novo Nordisk Foundation

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.

Design and production

cplone.co.uk

 ISBN 978-1-7397102-2-4 **DOI** 10.6084/m9.figshare.29041715

Contents

Foreword

Artificial intelligence (AI) technologies are expected to transform almost all aspects of contemporary life. Research systems are not immune to this change: since the public launch of ChatGPT in November 2022 made cutting-edge generative AI technologies widely available and easily accessible, technologists and researchers have been forecasting almost complete transformation of all aspects of research. The AI revolution, it seems, has arrived for research.

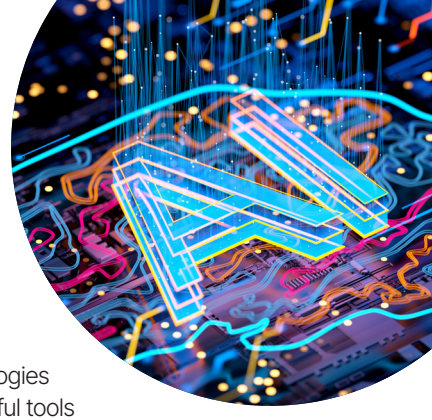
However, clear pathways to translate new advances in AI into tangible changes in the diverse day-to-day work of research systems have proven elusive. Dramatic successes in applying AI technologies to outstanding problems in molecular biology, mathematics, medicine and other fields have demonstrated the clear potential value of AI technologies as a tool for enhancing research. Still, many explorations of AI in research fail to achieve meaningful benefit, and may expose researchers and other actors in research systems to significant risk of bias, loss of data and intellectual property (IP), and even misinformation. New guidance, insights from experience and shared practice are needed to support more effective, beneficial and systematic use of AI. These resources cannot focus on the research process alone, but must be sensitive to the complexities of the wider research system.

In 2021, RoRI and the Research Council of Norway convened three workshops to discuss emerging practice in the use of AI and machine learning in research funding. The rich discussions in those workshops, and the questions and opportunities they raised, identified a clear need for a broader investigation of how AI and machine learning technologies could be used effectively, ethically and equitably in the work of research funders (Holm et al, 2022).

In response, RoRI initiated a project called *GRAIL: Getting Responsible About AI and Machine Learning in Research Funding and Evaluation*. From 2023-25, the GRAIL project has worked closely with a global consortium of 13 research funders, including public and private funders from three continents. The project aimed to investigate two questions: how are research funders using AI and machine learning now; and how can we build on those experiences to create shared, sector-level knowledge and new practices to support effective, responsible use of AI technologies in the work of funding and assessment?

This handbook is our answer to those questions. The result of two years of close collaboration and intensive discussions, the resource in your hands illustrates the diverse experiences of funders exploring and applying AI, some of the benefits AI use can produce in funding and assessment processes, and the challenges that funders and other actors in research and innovation systems must grapple with around AI use. We outline the key steps and decision processes involved in AI applications, and provide a starting point for funders to build their own practice from a strong base of shared understanding.

This handbook is titled *Funding by Algorithm*. This does not mean we advocate the use of AI algorithms to make automated decisions about funding; far from it. This is not a sales pitch for AI in research funding, nor is it a manual prescribing specific steps to maximise AI use. Rather, we use the term algorithm as shorthand for a more data-driven approach to research funding where funders turn these new research tools back onto their funding portfolios. We hope for this handbook to become a critical companion and an inspiration to funders and all who are shaping research systems worldwide.



AI technologies are powerful tools to inform and drive change in research systems, and we aim to provide the key knowledge, questions and challenges to enable funders to explore and apply AI in a way that reflects their own experiences and needs. We hope that this may support their efforts to drive high-quality, effective research funding and unlock research potential.

We warmly invite comments on this handbook at **hello@researchonresearch.org**, and look forward to seeing how research funder experiences, experiments and emerging practice with AI and machine learning grow in the years to come.

Denis Newman-Griffis
Senior Lecturer and AI-enabled Research Lead, University of Sheffield
Research Fellow and GRAIL project lead, RoRI

Jon Holm
Special advisor, Research Council of Norway
Chair, GRAIL steering group

Katrin Milzow
Co-Director, Swiss National Science Foundation
Co-Chair, RoRI

Summary

This handbook is intended as a starting point and go-to reference for exploring and applying artificial intelligence (AI) and machine learning (ML) technologies in research funding and assessment. It is primarily written for research funding organisations, but other organisations throughout research and innovation ecosystems will also find valuable guidance and information in these pages.

The handbook can serve as a reference for funders at any stage in their AI/ML journeys. For those just setting out and learning how and why they might go about applying AI/ML in their work, we share invaluable guidance from experienced research funders on the policy contexts and organisational processes involved in getting started with AI and ML. For funders with AI experience, including global leaders in AI/ML innovation, we provide a clear and straightforward consultation reference for planning, managing and evaluating new AI/ML applications, and for sharing insights with other funders and stakeholders in research.

What this handbook is for

Practical guidance for research funders to help:

- **Identify applications for AI/ML** in research funding and evaluation
- **Select appropriate AI/ML technologies** for specific problems
- **Bring people, resources, and motivations together** around AI/ML applications
- **Manage organisational complexity** in bringing AI/ML into practice.

Outline

- **In Part 1,** we give a **short introduction to AI** for research funders and outline the **foundational principles for responsible use of AI/ML** and **resilience to technological change** that should underpin all funders' work in this important area.
- **In Part 2,** we describe the **wider contexts motivating funders to explore AI/ML** and shaping the use of these technologies within increasingly complex research policy aims.
- **Part 3** outlines the **key steps involved in applying AI/ML** in a given situation. We focus on foundational aspects of AI/ML applications that are relevant across different technologies and purposes.
- **Part 4** widens the lens to discuss **key organisational issues in implementing AI/ML** in an effective and sustainable way.
- **In Part 5,** we share **real-world case studies** drawn from the experiences of RoRI partner funders, showcasing AI/ML applications in practice and how funders have approached the process.
- **Finally, in Part 6,** we reflect on **key insights** from the process of shared learning and knowledge exchange that underpinned RoRI's GRAIL project and the production of this handbook. We conclude by highlighting specific examples, arising from discussions in the GRAIL project, of **future directions for funders to experiment** with adopting AI/ML methods in research funding and evaluation.

Scope of this handbook

This handbook focuses on the use of AI and ML by research funders in their work as organisations: that is, primarily, in research funding and research assessment.

Research funders also have other roles – for example, informing the development of science policy at national and international levels, or helping to shape research cultures in the national research systems in which they operate.

Funders interact with AI in five broad areas:

- **AI in research**: researchers have studied AI as an object of research for many decades and used AI methods as tools in various areas of research for nearly as long. Funders play a fundamental role in shaping the course of this research through awarding funding to support it.

- **AI regulations and policy**: as convenors of scientific expertise, funders make significant contributions to the development of regulations and science policy regarding AI, as with any new technology.

- **AI in funding processes**: funders are actively exploring – and, in some cases, already applying – AI methods to support the processes of awarding, managing and reporting on research funding, as well as assessing the outcomes of funded research.

- **AI in analysis**: AI methods are proving useful for deepening funders' analysis and understanding of changing research landscapes, as part of the funder's role in strategic planning and insight.

- **AI in administrative processes**: there is growing use of AI technologies, particularly generative AI, to support administrative processes in funders, as in other organisations. These include processes such as human resources, data management, or producing documentation.

This handbook focuses primarily on **AI in funding processes**. We include brief discussions of other areas where relevant (for example, in Section 4.4 we discuss the role of funders in helping to guide researchers' use of AI). The roles and responsibilities of research funders in this wider view of AI in research systems is the subject of ongoing policy development, addressing requirements for applicants and evaluators, training, and financing research on AI (cf. Directorate-General for Research and Innovation, 2025). We refer to other resources in this evolving space to supplement our discussion.

Recent developments in AI also have much wider relevance throughout research systems. In 2025, researchers, research managers, publishers, and other actors in research ecosystems are actively exploring wider uses of AI in the conduct, management and dissemination of research. Using generative AI to help write scientific publications, AI-assisted hypothesis development and experimental design, AI-based analysis of large-scale data; these and other applications are reshaping the day-to-day conduct of research around the world. This is an area of dynamic and evolving professional practice, and a discussion of the wider role of AI in research systems is out of the scope of this handbook. Nonetheless, the processes and challenges outlined in this publication, and the questions we raise, have wider relevance to people and organisations throughout research systems. This handbook, though focused on funders and AI in funding, can therefore serve as a starting point for future, wider explorations of AI in research.

> "There is growing use of AI technologies to support administrative processes in funders, as in other organisations"

Foundations for AI/ML

To lay the groundwork for our discussion of AI and ML in research funding, we first need to establish a common language and understanding of what exactly AI, ML, and the various terms and ideas that are discussed around them mean in practice.

This first part outlines the key terms, concepts and distinctions that underpin the remainder of our discussion of AI and machine learning in research funding and assessment. We also describe the approach we take throughout the handbook to build a strong foundation for funders to approach the changing landscapes of AI/ML technologies from a place of responsibility and resilience, and the perspectives we advocate for funders to adopt in responsible exploration and application of AI.

1.1 A SHORT INTRODUCTION TO AI

Artificial intelligence is a broad umbrella, encompassing a wide variety of technologies and approaches. As the AI field has grown, new technologies and applications have proliferated, and the term 'AI' has come to mean many different things to many different people. This section provides a brief overview of key terms and concepts in AI, and describes how these are used to guide the content in this handbook.

1.1.1 AI, MACHINE LEARNING AND DATA SCIENCE

Terms such as *AI*, *machine learning* and *data science* are often used interchangeably, but there are important distinctions between them that are essential to working with AI approaches in practice. Here, we give operating definitions for these and other key terms, which ground the remainder of our discussion in this handbook.

There are many definitions for each of these terms, but a useful way to think about the differences between them is in terms of the purpose they aim to serve. Figure 1.1 illustrates the relationship between these terms.

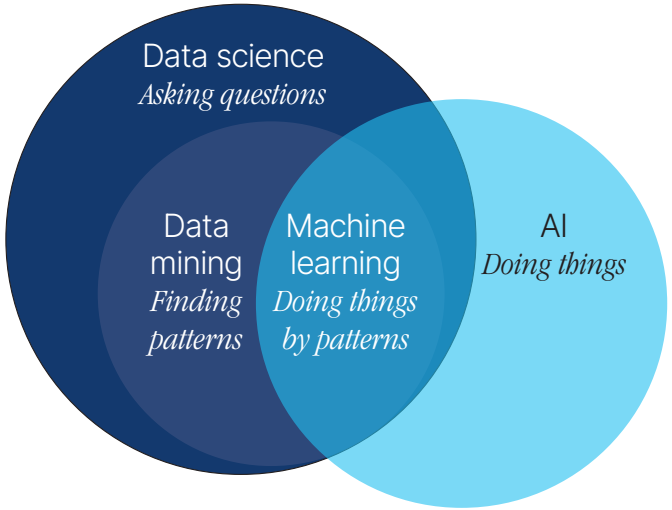


FIGURE 1.1 THE RELATIONSHIP BETWEEN AI, DATA SCIENCE, DATA MINING AND MACHINE LEARNING

Artificial intelligence (AI) focuses on *doing things*. John McCarthy, one of the founders of the AI field, defines AI in terms of the computational ability to "achieve goals in the world" (McCarthy, n.d.).

- In this handbook, we use 'AI' broadly to reflect diverse computational methods 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.

Data science focuses on *asking questions*. Data science is a very broad and multidisciplinary field, encompassing "collection, storage and processing of data in order to derive important insights into a problem or a phenomenon" (Shah et al, 2020).

- In this handbook, data science focuses primarily on funders' use of data to learn about or help inform action in their work.

Data mining focuses on *learning things*, and is a subset of data science. Data mining encompasses a wide variety of approaches, but use of data mining most frequently relies on statistical approaches from large datasets.

- Data mining has been used in various forms by funders for many years, and is not a primary focus in this handbook.

Machine learning (ML) sits at the intersection of data mining and AI, and focuses on *doing things by patterns*. ML models can also be used as one component of more complex AI systems. Machine learning is currently the most widely used form of AI, and it is generally safe to assume that a reference to an 'AI model' is more specifically referring to an ML model.

- In this handbook, machine learning is the primary focus of discussions of AI, and we address a variety of machine learning methodologies and applications relevant to funders.

1.1.2 AI TERMS

Discussions of AI frequently refer to several interrelated terms, such as algorithms, models, methods, technologies and systems. These terms are often used in different ways – and, at times, interchangeably. We provide brief operating definitions of these terms here.

AI algorithms is a loosely defined term that generally refers to automated systems in which AI plays some part. Understanding of the term 'algorithm' varies between disciplines: for example, an algorithm in computer science is a

technical term referring to a reusable process for solving a problem (Cormen et al, 2022); while, in critical data studies, an algorithm may include the larger sociotechnical assemblage behind a particular application of technology (Kitchin, 2019). In this handbook, we follow the most common general-purpose usage of 'algorithm' to refer to a particular technical system used to accomplish a specific purpose (ie, a funder may experiment with two different algorithms, or algorithmic systems, for AI-assisted matching of proposals to reviewers).

AI systems are also variously defined, but typically function in a similar way to our operating definition of 'AI algorithm': a particular application of AI to help solve a particular problem. However, we take a slightly broader view of an AI system to be not just the technical implementation, but also "the broader system of people, data and processes that motivate and make use of AI computation" (Newman-Griffis, 2025).

AI methods, or methodologies, refers to the scientific and engineering methods involved in developing AI systems. These may be methods for modelling, data analysis, performance improvement, or more, all falling under the AI umbrella.

AI technologies refers to particular technical implementations used in AI systems. These are typically implementations of specific AI methods: for example, ChatGPT is an AI technology implementing a language modelling methodology.

AI models, or machine learning models, are typically statistical models of a particular relationship between data inputs and outputs. For example, a model may represent a mapping from funding applications to thematic categories, or from a research output to an assessment score. AI models are the heart of most AI technologies.

1.1.3 KEY AI DISTINCTIONS

AI methods and technologies come in many different forms. Here we outline several important distinctions for understanding the current landscape of AI technologies and how they apply to the funder use cases discussed in this handbook.

Types of AI outputs: deterministic vs probabilistic systems

AI systems produce outputs in response to specific inputs. This process may be *deterministic*: given an input, a deterministic system follows a particular, repeatable process, and the same input will always produce the same output.

Or an AI system may be *probabilistic*: given an input, a probabilistic system will still follow a particular process, but an element of random chance will be involved, and the same input may not always produce the same output.

For machine learning-based systems, this distinction applies both during the training of a model and during its application. In deterministic training, using the same training algorithm with the same training data will produce exactly the same model. This is unusual in practice: most ML models use probabilistic training, which includes some random factors in the training process and will produce slightly different models when trained twice using the same dataset. However, at application time (also referred to as 'test' time), the reverse is true: most

Model training \ Model application	Deterministic	Probabilistic
Deterministic	● Same training data produces same model ● Same test data produces same output ***Example:*** Automated research assessment scorer	● Same training data produces same model ● Same test data may produce different output ***Example:*** Model of peer review scores with reviewer variance
Probabilistic	● Same training data may produce slightly different model ● Same test data produces same output ***Example:*** Topic classification for funding proposals	● Same training data may produce slightly different model ● Same test data may produce different output ***Example:*** Self-assessment tool to produce synthetic peer reviews

FIGURE 1.2: DISTINCTIONS BETWEEN DETERMINISTIC AND PROBABILISTIC AI SYSTEMS, AT TRAINING AND TEST TIME, WITH INDICATIVE EXAMPLES OF EACH

ML models are deterministic in how they produce output, while probabilistic output is typically used only for situations where the output is intended as a sample from a distribution. Figure 1.2 illustrates these distinctions with example applications for each.

For funders, these different approaches have different implications for the use of AI systems in making various kinds of decisions, each of which will have its own requirements for transparency and accountability. Deterministic approaches are a better fit for applications where clear, traceable and repeatable processes are needed, such as in making final funding decisions; probabilistic approaches are valuable for simulations, where variation is part of the intended purpose.

Forms of AI: expert systems vs machine learning

Expert systems and machine learning describe two distinct approaches to AI. An *expert system* typically comprises a set of rules that represent the knowledge of human experts related to a particular task, with the complexity of the rules functioning as the system's 'intelligence'. For example, in a funding advice program crafted from the knowledge of an expert scientific officer, one rule might be that a person expressing an interest in flight mechanics might target aerospace engineering funding. Expert systems are typically, though not exclusively, deterministic.

In contrast, a *machine learning* system extracts patterns from data, typically with respect to a particular task, and tries to learn the patterns that associate with the desired task outcomes. A machine learning funding adviser might be fed with thousands of applicants' CVs and their funding histories in order to learn patterns that associate the two. An example pattern might be that CVs mentioning flight-related words and evidence of a physics or engineering background are associated with success in aerospace engineering funding.

Machine learning approaches: supervised vs unsupervised learning

Machine learning systems also include different approaches. The most common method is called *supervised* machine learning, because the training process (ie, estimation of the statistical model) is 'supervised' by use of expected outputs for example inputs; that is, the system is guided towards patterns that associate with specified targets. Expected outputs may be categorical (eg, funding areas), in *classification*, or numeric (eg, funding amounts), in *regression*.

An alternative approach is *unsupervised* machine learning, in which machine learning models are fitted to descriptive patterns observed in data, without being steered towards a given target. The most frequent type of unsupervised learning is clustering algorithms. For example, an unsupervised learning algorithm applied to a set of CVs might find an association between flight-related words and aerospace engineering, but could equally find associations between funding success and seniority, between gender and hobbies, or between chemistry and physics qualifications.

Generative AI, discriminative AI, and large language models

Most AI systems in practice are *discriminative*: that is, they are built to tell similarities and differences between particular inputs. This may be assigning categories, placing inputs into clusters, predicting output values, and so on.

In contrast, *generative AI* generates new outputs in response to an input prompt, rather than selecting from a predefined set or predicting a single value. Whereas a discriminative ML model might recommend funding areas in response to a CV, for example, a generative AI system might produce a novel paragraph of text about potential future funding opportunities.

The most common form of generative AI currently in use is the group of *large language models* (LLMs). These are unsupervised ML systems that have been trained with very large amounts of text data and represent complex patterns in this text, which enable them to respond appropriately (or at least plausibly) to novel prompts.

These systems may be used to perform a traditionally discriminative task, relying on the large quantities of knowledge captured in the text patterns on which the LLM was trained. For example, an LLM could be prompted to perform a ranking task with a prompt such as: "Which potential reviewer CV matches most closely the content of this grant proposal?" This prompt will produce a text response that can be interpreted by a human reader to identify the answer to the question, though it will not produce the same type of automated categorical response that a discriminative system would.

General-purpose models vs bespoke machine learning

LLMs are one type of *foundation model*: an AI model that is trained once on a large amount of heterogeneous data and reused for a wide variety of applications as a general-purpose tool. More broadly, ML models of any kind may often be reused for other purposes, whether intended as a foundation model or not.

For example, a model trained for topic modelling in funding proposals by one funder may be reused by another, as long as the proposals to be analysed can reasonably be expected to be similar between the two funders.

In contrast, bespoke machine learning models are developed on specific datasets for specific applications. Bespoke models can be highly attuned to particular characteristics of the context in which they will be used: for example, a funder may choose to use a bespoke model to help with reviewer matching in an unusual funding programme. Bespoke ML models are not generally portable to other settings, but can be very powerful tools for capturing fine-grained distinctions within a consistent setting.

1.1.4 AI LAYERS: FROM CONCEPT TO USER INTERFACE

AI applications are complex endeavours, involving many layers of selection and implementation to get from a concept or a goal to a usable system with a user interface. Each of these layers involves different decisions about what to purchase or build, and how to manage an AI system, so it is important for all stakeholders interacting with AI systems to be aware that the interface they see is not the same as the technology underneath it. We briefly outline key layers in the implementation of AI systems here, illustrated in Figure 1.3:

- The **goal for AI use** is the motivation behind a funder's use of an AI system in the first place; for example, matching reviewers to funding applications.
- The **AI task** is the specific problem within this goal that a particular AI system is intended to solve. There may be multiple tasks involved in a given goal: for example, reviewer matching may involve modelling the information in application materials, modelling reviewer history, and calculating similarity scores.
- The broad **modelling approach** used to tackle the given AI task. There may be multiple possible approaches: for example, modelling the information in application materials might be done with language models based on the Transformer architecture (Vaswani et al, 2017) or with word-level representations (cf. Mikolov et al, 2013).
- The particular **model structure** used to implement the modelling approach. Much of AI research focuses on developing better model structures (as well as innovating new approaches). For example, many recent innovations in generative AI have come from the generative pre-trained transformer (GPT) model structure.
- The specific **model** trained to perform the AI task based on the chosen model structure. It is at this point that specific training data and a training algorithm are chosen, and a reusable model developed. For example, ChatGPT 4.0 is a specific model based on the GPT model structure.
- Finally, the **AI interface** is a user-facing system built on top of the trained model. Any given model may be used with multiple interfaces: for example, ChatGPT 4.0 can be

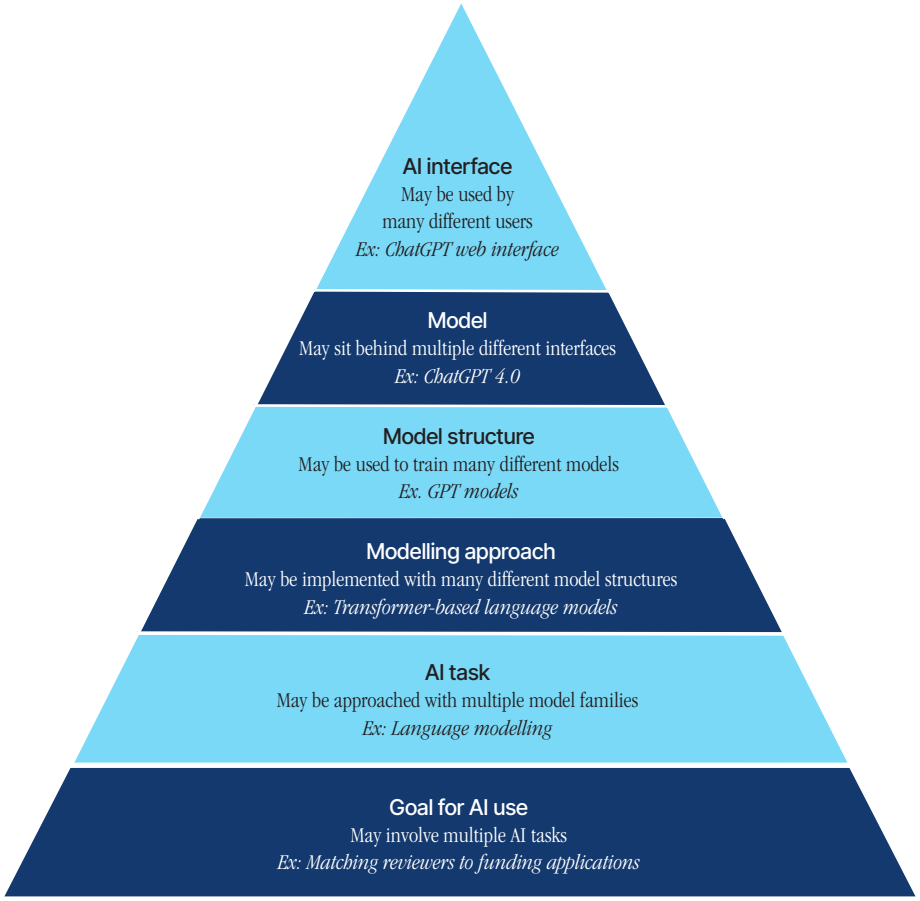


FIGURE 1.3 IMPLEMENTATION LAYERS OF AN AI SYSTEM, FROM GOAL TO BE ACHIEVED TO USER INTERFACE

accessed through a web interface, a smartphone app, or an application programming interface (API).

The choices involved in each of these stages, and the process of developing specific AI models, are discussed in greater detail in Part 3.

1.1.5 AI EVERYWHERE: WHY DO WE USE 'AI/ML'?

AI technologies are everywhere. Many processes and tools with which we interact on a daily basis include some elements of AI: web search uses dozens of different AI models; spelling and grammar correction in word processing software is increasingly AI-based; and even everyday technologies such as automated number plate recognition make use of AI.

However, with the rapid growth of consumer-grade generative AI technologies – such as ChatGPT, Gemini, Claude, and others – the term **'AI' is often used to mean chat-based generative AI** systems in particular. These systems have been influencing new ways of working with AI and new potential impacts for research funders, which we discuss further in this handbook.

However, many **more tools within the AI umbrella** are equally relevant for research funders. Simpler, lower-cost and more easily-managed approaches are often better fits than large commercial generative AI systems for common funder tasks such as reviewer matching, screening of proposals for funding, and grant management.

This handbook takes a **wide lens on AI**, focusing on guidance relevant to any AI methodology and highlighting issues specific to generative AI where relevant. Machine learning has particular relevance to funders, as evidence-based organisations with increasing pressures to be data-driven. ML methodologies are also often discussed separately from overall 'AI' for data-driven work, and may not always be perceived as part of the AI umbrella.

We therefore refer to both AI and ML throughout this handbook, to reflect the importance of referring to both approaches in the practical, day-to-day work of research funders.

1.2 RESILIENCE IN CHANGING AI/ML LANDSCAPES

AI has a long history: the term 'artificial intelligence' was coined in the 1950s and key ideas of AI are often traced to the work of Alan Turing in the first half of the 20th century. This history is one of constantly looking ahead to dramatic future transformations, and of constant iteration and change in AI technologies and methods. The development of public-facing generative AI platforms, most notably the public launch of ChatGPT in 2022, has accelerated discussions of AI transformation across many sectors, and has opened up new avenues for exploring the use of AI technologies to support a wide variety of tasks and applications. In a landscape of rapid change, funders need stable and reliable ways of thinking about the use of constantly evolving technologies that enable them to build resilience to turnover and disruption in AI.

These changes have significant implications for research funders. **Funders often set the course for the cutting edge of research** and need to be actively engaged with the latest developments affecting research systems. **Funders are also important knowledge institutions**: they hold vast records of data on research inputs and outputs, and bring together extensive experience and cross-cutting expertise within their walls. They are tasked with leveraging these data with their knowledge and experience to be as nimble and responsive to changing research landscapes as possible, and AI technologies can be very powerful tools to achieve this. At the same time, **funders have vital responsibilities to protect the confidentiality and integrity of sensitive research-related data**, and must always be able to demonstrate clear accountability and value to society for the work that they do.

The recommendations and guidance in this handbook are founded on three core principles for responding to these changing landscapes:

- **A problem-based approach**: we focus on AI as a set of tools with which to respond to specific problems and challenges funders wish to solve, and goals they wish to achieve. We emphasise this approach over being *solution-led* and seeking out opportunities to apply AI.
- **A focus on skills and process**: we primarily address the skills and competencies involved in applying AI tools to specific problems, and the processes that make this successful. We advocate for fundamental, reusable skills over specific technical guidance for particular technologies.

- **A competency-based approach**: we focus on the broader underlying competencies involved in AI use, rather than developing technical expertise in specific AI tools and technologies, allowing funders to build expertise in AI that grows over time and as technology changes.

Together, these principles prepare funders and other organisations to respond productively to new developments in AI technologies by assessing their capabilities, fitness for purpose, and potential impacts in each organisation's unique context.

Part 3 outlines the steps involved in a problem-based approach to AI/ML use for funders. Part 4 describes key competency models and strategies for emphasising skills and process in AI use. The case studies in Part 5 illustrate real-world examples of specific organisational knowledge and technical decisions that inform practical AI use. Together, these resources provide a valuable starting point for funding organisations to move forward on building AI expertise that is robust and resilient over time.

1.3 PRINCIPLES AND PRACTICES FOR RESPONSIBLE USE OF AI

Responsible use of AI in research funding organisations meets three criteria:

It is **effective**: applications of AI address the goals for which they were designed, and do so in a way that supports a well-functioning organisation.

It is **ethical**: AI applications are designed, implemented and managed to minimise harm to those who may be affected (eg, staff of funding organisations, researchers, members of the public), and to support the good of society.

It is **equitable**: the use of AI helps to close opportunity gaps for individuals participating in the research system, and does not unfairly disadvantage one group over another.

Achieving these goals amid rapid growth of AI adoption, the proliferation of AI start-ups and innovators, changing regulatory structures, and frequent new technologies requires funders to adopt a **strong foundation of good practice**.

Used appropriately, AI technologies can offer a variety of benefits to funders, including efficiency improvements in decision-making, enabling funders to leverage and learn from a greater amount and variety of information about research, and enabling data-driven discovery about the research system. However, poorly implemented or poorly contextualised use of AI poses significant risks, such as: entrenching structural biases in research systems by mimicking past, unfair patterns; misusing researcher data and measures for unfair assessment; and learning to value some disciplines, methodologies, etc over others.

> "Poorly implemented use of AI poses significant risks, such as entrenching structural biases in research systems"

Achieving the benefits of AI use while avoiding its risks requires funders to build a strong base of responsible AI practices and translate principles of AI for good into day-to-day action.

This section highlights key international principles we recommend to guide funders' responsible AI efforts, together with practice-based recommendations developed from discussions with funders in the GRAIL project. The implementation of these principles in the context of complex funding organisations is further discussed in Part 4.

1.3.1 BENCHMARK PRINCIPLES

Responsible AI is a very active space of research and discussion, with many competing visions of responsibility and what responsible action looks like. We particularly recommend that funders refer to two international public policy documents outlining key aspects of responsible, ethical and trustworthy AI, developed by the European Commission and the Organisation for Economic Co-operation and Development (OECD).

Ethics Guidelines for Trustworthy AI, presented in 2019 by the European Commission's High Level Expert Group on Artificial Intelligence, describes three foundations of trustworthy AI:

1. It should be **lawful**, complying with all applicable laws and regulations.
2. It should be **ethical**, ensuring adherence to ethical principles and values.
3. It should be **robust**, both from a technical and social perspective, as – even with good intentions – AI systems can cause unintentional harm. (*European Commission, 2019; p5*)

This is underpinned by four principles for ethical and robust AI:

1. Respect for human autonomy
2. Prevention of harm
3. Fairness
4. The principle of explicability
(*European Commission 2019, pp12-13*)

To achieve these principles, the guidelines outline seven requirements that must be met by AI systems, individual users and societal contexts:

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental wellbeing
7. Accountability
(*European Commission 2019, pp15-20*)

The guidelines also provide a pilot Trustworthy AI Assessment List for organisations to measure themselves against these seven requirements (*European Commission 2019, pp26-31*).

OECD Principles for Trustworthy AI
The OECD's Principles for Trustworthy AI were published in 2019 and updated in 2024, and have been adopted as an intergovernmental standard by 47 governments at the time of writing. They endorse five key values-based principles to guide AI actors and provide recommendations for policymakers:

1. Inclusive growth, sustainable development and wellbeing
2. Human-centred values and fairness
3. Transparency and explainability
4. Robustness, security and safety
5. Accountability
(*OECD, 2024*)

1.3.2 RECOMMENDATIONS FOR RESPONSIBLE AI/ML PRACTICE

Drawing on the discussions and research conducted in the GRAIL project, we make four specific recommendations for research funding organisations to help operationalise the benchmark principles above:

Recommendation 1: Assess AI use in terms of its impacts on the broader research system
Research funders have significant influence over the research systems in which they operate, helping to determine what research is funded, new investments in strategic directions for research, and how research policy is shaped and implemented – and, often, directly affecting the development and progression of researchers.

European Commission Trustworthy AI guidelines

Any use of AI by funders must be sensitive to this influence and must be assessed in terms of the kinds of broader impacts funders aim to have on research. For example, AI-based assessment of researchers or research proposals may replicate past inequities, such as perpetuating gender disparities or preferential outcomes for specific institutions. AI systems will always exhibit biases towards particular outcomes: the role of funders is to assess what biases are present, which are harmful and should be mitigated, and which (if any) can support positive change in research systems.

Recommendation 2: Use existing policies, principles and values to guide new applications of AI
AI is often presented – and perceived – as a source of novel, dramatic transformation for organisations that use it. However, many practical aspects of AI use are already well addressed by existing policies and best practices. For example, data governance and confidentiality agreements provide established frameworks for managing data use in AI applications. Established organisational values can guide how

organisations shape what responsible use and management of AI systems looks like in their contexts.

Funders should therefore approach AI use with an eye towards leveraging existing policies, principles, processes and values wherever possible. Where new approaches are needed to deal with unique challenges posed by AI, these should be addressed with the minimum additional burden on the organisation, to maximise the likelihood of uptake and compliance with new policy. This approach will enable funders to be more responsive to changes in AI systems and applications, and to integrate responsible use of AI into existing culture.

Recommendation 3: Proactively manage changes to internal process, culture and trust from using AI

The introduction of AI systems into an organisation's work has knock-on effects on the processes informed by AI use, the trust staff and end users place in those processes, and the internal culture within the organisation. For example, established practices for transparency in decision-making may not be easily adapted for the use of AI, and introducing AI into a well-understood process may cause its legitimacy to be questioned. These issues are exacerbated when AI tools or expertise are sourced externally and may be less flexible or adaptable to an organisation's needs.

Funders must be proactive in countering these risks to internal and external perception of their work with AI. Clear communication – beginning from exploratory stages, before any AI solutions have been procured or implemented – helps staff and end users engage with the intended purpose of AI use. Working directly with staff, researchers, policy audiences and other stakeholders to shape how AI is used helps to build trust in AI applications, as well as the ability to constructively challenge them. These strategies may be of equal benefit to those with little AI expertise, who do not know what questions to ask, and to those with high AI expertise, who may bring expert scepticism to its use in their work.

Recommendation 4: Treat responsible AI as an evolving, iterative process

Getting AI right is not about a single decision. AI technologies will continue to evolve, as will the data they are used to analyse, the purposes to which funders put them, and the policy contexts in which funders operate. Use of AI technologies, especially in settings with significant financial impact, such as research funding and assessment, will also be shaped by evolving regulatory environments around the world.

Funders should therefore approach their policies and practices for the responsible use of AI as an iterative and continually evolving process. New benefits and risks from AI use will continue to emerge as funders explore AI for a wider variety of applications. Building a practice of repeated assessment, monitoring and mitigation will enable funders to be agile in responding to changing AI technologies and changing contexts of AI use.

> "Funders should approach their policies and practices for the responsible use of AI as a continually evolving process"

1.3.3 PUTTING THE RECOMMENDATIONS INTO PRACTICE

The remainder of this handbook presents key information, best practices and real-world examples to help research funders apply these principles and recommendations in their day-to-day work.

The principles and recommendations presented here do not cover all aspects of AI use, and funders making use of AI must develop their own specific practices to ensure that their AI use is effective, ethical and equitable. However, the approaches outlined in this chapter will provide a strong foundation on which funders and other organisations working in the research system can build, and create a shared understanding of responsible AI in research funding and assessment.

The case and context for AI/ML

AI/ML use by funders is motivated and affected by the wider contexts in which research funding organisations operate. Here, we examine three key questions to help set the stage for our subsequent guidance and case studies:

1. What motivates funders to explore AI/ML use?
2. What areas of the work of funding organisations provide the clearest opportunities for AI/ML use?
3. How does AI/ML use in funding organisations relate to the wider landscape of AI in research systems?

2

2.1 AI/ML MOTIVATIONS FOR FUNDERS

The shift to being 'data-driven'. AI and machine learning applications are part of a larger shift towards data-driven approaches in research funding and assessment. The integration of big data analysis and AI is transforming business processes across the private and public sectors: for example, a recent OECD report forecasts that the use of AI by the public sector "can increase productivity, responsiveness of public services, and strengthen the accountability of governments" (OECD, 2024). This is based on an assumption that collecting and analysing data from customers and relevant markets provides a competitive advantage to organisations that are able to make good use of these data in decision-making and business development (Iansiti and Lakhani, 2020).

Funders evaluate research ideas at a larger scale than any other research organisation. They **collect and process data**, often through proposals for funding. They aim to **evaluate these proposals** efficiently while minimising the use of the public funds they distribute. Additionally, they must **maintain transparency** to ensure fair treatment of all parties involved, including taxpayers and applicants.

Until recently, the knowledge base developed and used by research councils was dominated by the expertise of scientific peers, and other relevant stakeholders, taking part in programme development and assessment of proposals for funding. A transition from mainly experts-based governance to a more data-driven system for decision-making is in its early stages around the globe. To succeed, this transition will need careful planning, and to be mindful of the ideals of academic freedom and self-governance of science.

The need for data-driven funding and assessment. The role of a research council is to stimulate the best possible knowledge creation, development and discovery that may serve present and future needs in the private and public sectors. Public research policies are increasingly focusing on directing research investments towards societal goals, as seen in the development of the European framework programmes for research and innovation (European Commission, 2023).

At the same time, selecting proposals for funding is still very much a bottom-up process, in which project proposals are selected by scientific peers based on scientific criteria. Recent research has highlighted the limits of peer review in assessing the potential societal benefits of research proposals (Oxley and Gulbrandsen, 2025). When designing research programmes aimed at tackling societal challenges or strengthening strategic innovation capabilities, research funders may need to extend their knowledge base beyond the expertise of scientific peers.

The role of AI in achieving the shift to 'data-driven'. This handbook provides examples of how big data analysis and AI techniques are already used by research funders around the world, in programme design, selection processes and grant management. These are multi-stakeholder tasks that draw on a range of knowledge: for example, the first two tasks typically involve academic expertise and, in the case of challenge-oriented research, expertise from the user side. In all cases, stakeholder involvement brings with it interests and ideologies that go beyond 'pure' scientific knowledge.

In current research systems, important parts of the knowledge base for priority setting and funding decisions are often hidden. Big data analysis and AI offer new opportunities to develop systematic empirical knowledge on the pathways to societal impact for different research disciplines and various configurations of research projects. For example, a recent RAND analysis of impact cases submitted to REF2021 in the UK mapped the influence of research on society using text search, bibliometrics, topic modelling and policy citations (Stevenson et al, 2023).

What do funders need for this shift?
Adopting a data-driven approach to decision-making will require a new set of skills and organisational transformations in research funders. A recent scoping review on AI for research funding organisations explored the potential benefits and challenges that AI presents for funders (Blatch-Jones et al, 2024). The study identified multiple areas of funders' operations with potential for improvement with AI, including "data processes, administration, research insights, operational management, and strategic decision-making". However, achieving this improvement is not straightforward, and supporting AI transformations requires adaptable guidance for organisations at different stages in AI adoption and different levels of familiarity with AI.
The review identified four primary themes of concern, which we address throughout this handbook:

1. **AI readiness of organisations**: making the most effective use of AI requires access to and collaboration between teams with a wide range of expertise, and the organisational

infrastructure to leverage this. We discuss considerations and strategies for this collaboration in Section 4.3.

2. **AI readiness of data**: data management is the underpinning challenge for AI and often underestimated. We address key data processes for AI implementations in Section 3.2 and provide real-world examples in Part 5.

3. **Accountability and fairness in AI**: in addition to organisational stakeholders using or affected by AI, members of wider publics and end users must be involved in designing and assessing AI applications. We discuss key strategies for funders in Section 4.1.

4. **Governance and ethical use of AI**: funders must operate in a variety of regulatory contexts, such as data privacy legislation (eg, General Data Protection Regulations [GDPR]) and individual rights in algorithmic decision-making, as well as ensuring ethical principles and societal benefit in AI use. We discuss key considerations and recommendations in Section 1.3.

2.2 AREAS OF AI/ML APPLICATION

To help put the cases of best practice presented in Part 5 in context, we briefly outline the benefits of data-driven methods during the typical phases of response-mode research funding: 1) Strategic planning; 2) Project selection; and 3) Grant management.
For any research funder considering using AI/ML methods, the starting point will be to ask:

- What data are we collecting during the phases of the financing cycle?
- Is this data well structured and available for analysis?
- Are there external data sources we could use to enhance our data, such as publications databases or patent registries?

Without data there is no AI, so rather than drawing a sharp line between AI and other quantitative analysis, we emphasise how data-driven methods, including the use of AI, might enhance and complement traditional peer review in strategic planning, project selection and grant management.

2.2.1 DATA-DRIVEN STRATEGIC PLANNING

The research sector is one of the best-documented parts of society. The OECD makes available R&D statistics for international comparison and benchmarking every year (OECD, n.d.). Many countries provide key performance indicators – for example, the UK Research Excellence Framework (REF, n.d.) and the European Higher Education Sector Observatory (European Education and Culture Executive Agency, 2024). Finally, there is an increase of available data on the research process itself through openly shared data and open-access publishing. Metadata on academic research publications is now widely available through open sources such as OpenAlex (OpenAlex, n.d.).

Advanced data analysis and AI offer new opportunities for analysis of **trends in research publications** in terms of thematic scope (topic modelling) or cooperation patterns (citation clusters). There are also open-source tools – such as VOSviewer (CWTS Leiden, n.d.) – for visualising citations-based or topical clusters as network maps. A network analysis of research publications may be used to show how a specific set of funding instruments influences the thematic focus of research over time, or to follow the development of cooperation patterns between disciplines, organisations and countries. This type of analysis may help a research council to identify relevant topics or actors within research that are missing in an existing grant portfolio, and that might be targeted in future calls for proposals.

Another use of bibliometric data in strategic planning is the **categorisation of research publications according to societal goals**. Several services are already available linking research papers to the UN Sustainable Development Goals (SDGs). Many funders and other organisations have similar needs to document alignment with national research priorities. Furthermore, some research funders have started to use data on research uptake in their reporting and strategic planning. Such data, often referred to as **altmetrics**, include citations in patents (indicating innovation potential), citations in policy documents (indicating policy relevance) and data on clinical trials (cf. Research Council of Norway, 2022).

Caution: It is important to acknowledge that the adoption of big data analysis in a research funder context does not make domain-specific knowledge redundant. For any type of quantitative analysis, it is essential to anchor the interpretation of the data in a good understanding of the context of the research activity

targeted by a funding instrument. Decisions on strategic priorities and programme design should always be based on a broader understanding of the research system at hand, either based on research on research, evaluations, or other qualitative-domain expertise.

2.2.2 DATA-DRIVEN PROJECT SELECTION

Many cases in this handbook focus on evaluating funding proposals, which is often the most resource-intensive aspect of financing. In addition to seeking efficiency gains, however, it is also important to consider how the use of AI may affect the direction of travel, by changing the chances of getting funded for various groups of proposals. Current applications include **automated matching of reviewers and proposals, similarity check between proposals, eligibility check and quality assurance of expert feedback**. Uses are mostly limited to the preparation and support of peer review. In fact, the division of labour between administrative tasks (carried out by funder staff) and assessment of project content (carried out by peers) might still serve as a line of demarcation for legitimate use of AI in research assessment.

While there are valid concerns about using AI to assess research quality, advanced data analysis can still aid peer review by comparing funded and non-funded proposals. An AI model trained on successful proposals can identify systematic differences in research topics, investigator gender, track records or hosting institutions between successful and unsuccessful projects. These differences might indicate research quality or relevance, but they could also reveal potential biases in peer review. One of AI's key strengths therefore lies in uncovering previously unknown biases.

Another way of checking the fairness and **consistency of peer review** is to analyse variation in grading between peers for a specific proposal and between proposals for a specific peer. Such an analysis can be done by classical statistical methods, but may serve – together with new methods from data science – to open the black box of peer review.

Caution: The use of algorithms for reviewer matching may have an impact on the choice of reviewers that, again, may alter the chances of getting funded. Most importantly, the matching will be affected by the amount of data available on various disciplines in the training data. The use of AI may therefore favour experts from more common disciplines, which, in turn, risks compromising the diversity of active reviewers. Another limitation when training AI models to recognise patterns in the historical portfolio of projects is that the model will probably work poorly for projects with radical new ideas/competencies or configurations of methods, disciplines and project partners.

2.2.3 DATA-DRIVEN GRANT MANAGEMENT

As with project assessment, grant management is a resource-intensive task for any research funder. Organisation of the task varies across councils, but, in general, there is less involvement of external experts and more standardised procedures to be followed. This makes the grant management process a potential area for automation. We first distinguish between three types of tasks within grant management: compliance to contract; data collection; and portfolio analysis. All of these tasks may be automated and/or enhanced by AI to some extent.

The **follow-up of contractual requirements** (eg, deliverables and milestones) may be automated if the requirements are sufficiently defined for the case-handling system to check the compliance of reporting to the requirements. This may include reported publications, progress reports, etc. In such an application of AI/ML, case handlers may be involved at multiple levels; in the case of highest automation, they would only intervene in cases where the AI system cannot verify compliance.

Some funders seek to reduce their costs by outsourcing the data collection from projects to external service providers. However, outsourcing data collection does not relieve PIs of reporting duties. Funders should therefore explore **alternative methods for data collection** to improve system efficiency. For example, AI/ML use could strengthen the **grant-to-publication link** based on funder attributions that are already present in publications. These links are already retrieved by several publications' databases. Other existing data sources include **policy citations**, **patent citations** and **news media coverage**, which may serve as indicators of early uptake of research and so limit the need for more costly qualitative reporting and evaluation. For examples, see case studies 5.4 and 5.5.

The potential for use of AI is even greater in the task of **portfolio analysis**. In general, a portfolio analysis classifies projects into relevant categories based on various properties such as funding lines, research disciplines and

themes, business sectors, and types of research. Then, at the next level, results and outcomes may be analysed for a set of projects defined by these categories. AI models may be especially useful in **classifying projects** according to established taxonomies – list of disciplines, research themes, etc – or in creating new categories based on previously classified projects.

Funders regularly use **descriptive statistics** for portfolio monitoring and reporting to governing bodies. For deeper insights, they can analyse **correlations** between funding types and research outcomes to understand the effects on quality and impact. Using complete historical project data, a research funder can implement **predictive modelling** to design programmes and formulate calls for proposals. Predictive modelling can assist funding bodies in selecting the best funding instruments and call formulations to meet specific research policy goals.

Caution: There are still important limitations on what we can learn from the success of previous projects, especially when it comes to predicting societal impacts from specific properties of funded research. The societal impact of research often depends on user-side factors, and typically emerges from the collective efforts of multiple projects and collaborations across research institutions and user organisations. Additionally, outcomes from funded projects may disseminate through informal channels or via the career trajectories of the researchers involved. It is important to acknowledge that not all dissemination of research can be formalised or traced. Consequently, when employing quantitative models, it is essential to consider the broader context of the funded initiatives.

2.3 AI/ML AND THE BROADER RESEARCH ECOSYSTEM

In developing good practice in the use of AI for research funders – who aim to serve both the public interest and the advancement of ethical and responsible science – there is a need for practical guidance that speaks to ethical practice in real-world decision-making.

2.3.1 SYSTEM RISKS OF AI/ML USE BY RESEARCH FUNDERS

The potential benefits of AI/ML use are balanced by significant risks from inappropriate or ill-informed use. General risks of AI/ML use have been studied and reported widely: for example, the MIT AI Risk Repository (MIT, n.d.) captures more than 1,600 distinct AI risks at time of writing. However, there are also particular risks to research systems from the use of AI/ML in funding processes, which we outline here.

When we use historical data on our portfolio of funded projects to train algorithms to support selection processes for new projects, we run the risk of **recycling the problems of the current research system**. Recent findings, such as in the National Institute for Health and Care Research (NIHR) review presented above, have demonstrated that many popular AI technologies, especially ML-based approaches, can reflect and even amplify social biases such as racism, sexism and ableism. These approaches also often depend on subtler limiting assumptions about the types of data they are used to analyse, as well as the assumption that past patterns are indicative of desired future outcomes.

Such problems are addressed in current research policy initiatives such as the Coalition for Advancing Research Assessment (CoARA), which promotes a shift from quantitative to qualitative assessment of research and researchers: *The vision of CoARA is to recognise the diverse outputs, practices and activities that maximise the quality and impact of research through an emphasis on qualitative judgement in assessment, for which peer review is essential, supported by the responsible use of quantitative indicators* (CoARA, n.d.).

As AI/ML use grows among research funders and researchers, there is an increasing risk of three kinds of **feedback effects**:

- First, there is a danger that applicants with better access to AI, or better skills in using AI, gain an unfair advantage in the funding process. That is, AI use could become 'an inappropriate deciding factor' in funding allocation.
- Second, the use of AI/ML in the reviewing process has implications for evaluation panels. AI/ML systems may struggle to replicate the established, expertise-based approaches to composing evaluation panels, making recommendations that fail to capture the disciplinary focus or the interdisciplinary breadth needed for different panels.
- Third, as funders and researchers increasingly make use of the same AI models (particularly generative AI systems), there is a risk of 'like judging like,' in which applications become more similar to one another and to funder expectations in order to succeed.

2.3.2 PEER REVIEW AND AI/ML

Peer review will probably continue to be the gold standard for the assessment of research quality for years to come. Efforts to introduce elements of automated decision-making in established assessment exercises such as the UK's REF, even in a highly controlled manner, have proven difficult. This is less because it is not possible for an AI system to make a good approximation of the evaluation results, and more because of issues of legitimacy and prestige connected to intermediate results, where the AI predictions are not entirely consistent with human peer review (see Case Study 5.6).

Still, peer review is a limited resource, and comes with its own flaws. Going forwards, it would be irresponsible not to investigate the potential for better data-driven methods that could 'inform' and even provide checks and balances on peer review. While current research reforms have focused on limiting inappropriate use of bibliometrics, initiatives such as a 'revisiting' of *The Metric Tide* report (Curry et al, 2022), commissioned as part of the Future Research Assessment Programme (Future Research Assessment Programme, n.d.), point to the possibility of expanding the database available in research assessment by collecting what the authors call "data for good". In the UK REF, this response addresses the need for value-led indicators to assess the research ecosystem and its community impact. Examples include teamwork volume, collaboration, co-produced research, open research indicators, and policy impacts through citations in policy literature.

Advanced data analysis could even be used to improve the use of peer review in a research funding context. In a series of webinars hosted by RoRI in 2021, Dr Kuansan Wang pointed out that biases in peer review are often caused by cognitive limitations. Machines may compensate for these limitations by processing superhuman amounts of information, extracting useful patterns, and making precise computations (Holm et al, 2022). The Swiss National Science Foundation (SNSF) offered another example in the same webinar series: an algorithm trained on a dataset of annotated grant peer reviews rated by experts for different aspects of review reports (evaluation criteria, focus of comments, statement type and reasoning). This could be used to analyse the contents of review reports at scale and, potentially, give automated feedback to peers on the characteristics of their reviews, thereby contributing to more helpful comments being given back to applicants (Okasa et al, 2024).

Caution: Funders wishing to increase the efficiency of their funding processes through the use of AI should also be able to document the fairness and effectiveness of the AI-assisted processes. This is partly a legal requirement. Under European privacy legislation, a data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her (Regulation (EU) 2016/679, Article 22). In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to: obtain human intervention; express his or her point of view; obtain an explanation of the decision reached after such assessment; and challenge the decision (Regulation (EU) 2016/679, Recital 71). More generally, transparency and explainability are key to the responsible use of AI/ML in research funding processes. Methods and processes for documenting fairness and ensuring transparency are developed further in Parts 3 and 4.

"Efforts to introduce elements of automated decision-making in established assessment exercises have proven difficult"

Practical guide to applying AI/ML

This part outlines the key steps in planning and implementing any single, defined application of AI/ML. AI applications vary widely: in the technologies, purpose, intended users, measures of success, and more. However, there are key steps and considerations that are common across most or all applications of AI, and a good understanding of these provides funders with a strong starting point for developing specific applications in their context.

We break these steps down into four phases, in approximately chronological order:

1. Motivating a specific application for AI/ML
2. Selecting and assessing data to use with AI/ML
3. Implementing the first version of the application
4. Ongoing evaluation and management of AI/ML over time

For each phase, we present best practices and highlight critical questions for teams and organisations. Brief examples are included to illustrate specific steps in the process; for more comprehensive examples of these steps in practice, refer to the case studies in Part 5.

3

3.1 MOTIVATING AI/ML USE

Before writing any code, analysing any data or training any models, a new AI application must be driven by the need it will address and the specific context in which it will function. We break this process down into three stages:

- Initial **problem formulation**, to identify the need for AI
- **Making the organisational case**, to tie AI use to specific strategic and organisational contexts
- **Assessing expected impacts and risks** of introducing a particular use of AI

3.1.1 PROBLEM FORMULATION

Effective AI/ML implementations start from a **clearly defined problem** and some initial idea of **how AI might be used to address it**. These may be considered in terms of three Ps: problem, priority and paradigm.

Problem identification. Specific problems and potential AI solutions can be identified through reactive and proactive approaches:

A ***reactive approach*** is when prospective users of AI share their specific challenges and collaborate with data science and AI experts to assess whether an AI intervention is feasible and beneficial. For example, scientific officers responsible for assigning peer reviewers to grant applications may struggle with slow and inconsistent matching because of the high volume of submissions. To address this, they explore AI-driven automation to streamline the process and improve matching quality. Teams working with specific funding lines may not always recognise where AI could optimise processes or improve decision-making. A ***proactive data science team*** can uncover opportunities for AI-driven solutions to add value. To bridge this gap, the data science team can arrange organisational roadshows to showcase AI capabilities and spark discussions about potential applications. Another effective approach is embedding data scientists in departmental meetings, allowing them to engage directly with teams, observe workflows and discover pain points – even when AI is not initially considered. For example, in a grant assessment panel meeting, data scientists might observe that reviewers frequently miss evaluating certain aspects of a proposal, such as impacts. In response, they could propose an automated system that prompts reviewers to cover all evaluation points.

Prioritisation. As AI potential expands, the demand for applications often exceeds an organisation's capacity to implement them all, making prioritisation essential. One approach is to focus on solutions that address multiple challenges at once. For example, categorising funding proposal topics is a recurring task for scientific officers, policy teams and administrative staff. Developing a scalable AI classification tool could streamline this process across teams. However, a more comprehensive solution means greater resources and planning – an inevitable trade-off. An alternative is to implement a quick fix to address immediate needs while placing the long-term solution in the development pipeline.

Paradigm. AI methods are generally organised around specific types of tasks, each defining how the system transforms inputs into meaningful outputs. These task types, or paradigms, are essential for guiding the design, development and evaluation of AI systems. For example, in *classification*, the system assigns inputs – such as the text of a funding proposal – to one or more predefined categories, such as research topics. In *information extraction*, the goal is to identify and extract specific elements from unstructured text, such as identifying the methods proposed, study populations or policy relevance within a grant application. Choosing the right paradigm is crucial, as it shapes how data is prepared, how models are evaluated, and how outputs will be interpreted by end users.

3.1.2 MAKING THE ORGANISATIONAL CASE

After identifying a problem, the next step is to **develop a case for an AI solution**. This involves several stages of working with users and other stakeholders within the organisation to articulate the role of the proposed AI system in specific funding or assessment processes.

Scoping. An internal scoping process aims to envision the end product – what an AI application will do and how it integrates into daily workflows. For example, an AI-driven reviewer-matching system could recommend reviewers by analysing the alignment between their recent publications and the proposal's content. It could automatically rank potential reviewers based on relevance, experience and past assignments. The tool might function as a decision-support system, generating ranked recommendations for manual selection, or it could fully automate assignments.

Proposal. Building on these discussions, data scientists and engineers can propose a feasible solution. They can provide examples of similar AI implementations to illustrate practical use cases. Data scientists also play a role in managing expectations and clarifying the solution's capabilities and limitations. This includes constraints related to data availability and quality, variations in performance across different use cases, and the long-term sustainability of the solution. Additionally, the organisation should assess whether an existing vendor solution meets the requirements or if developing a bespoke internal model would be more suitable. Part 4.1 elaborates on how to make these decisions.

Prototyping. Many important discussions remain uncertain when working with a hypothetical AI application. To quickly ground them in practical insights, a small-scale prototyping phase can be conducted between technical experts and users to validate initial assumptions and identify potential roadblocks, such as missing tools, insufficient data or gaps in expertise. This is then followed by a more extensive iteration, focusing on broader scoping and detailed planning.

Linking to organisational goals. For AI initiatives to gain support, they should clearly link to the organisation's overarching objectives, such as improving efficiency, enhancing transparency or accelerating research impact. For example, gaining insights into funding impact is a priority for many funders. Developing an AI tool to better track the translational impact of funded proposals can provide greater visibility into how research contributes to real-world advancements.

3.1.3 ASSESSING EXPECTED EFFECTS AND RISK

One key part of initial scoping for a new AI application is to **evaluate potential risks** associated with it and **establish mitigation measures** that can guide the development process.

As an example of this process, the Wellcome Trust has designed a pre-development impact assessment workshop to facilitate a structured discussion on the topic independent of the technical solution (Spengeman et al, 2022). We summarise the main points of this assessment process, which provides a valuable template for approaching any new AI/ML application:

- Discuss the **function and intended use** of the AI tool
- Discuss the **uses and misuses** of the system
- Discuss **risks and minimising the risks**
- Assess **impacts** on different groups of people/disciplines
- Define a **strategy to minimise the risk** of negative outcomes
- Define **criteria for fairness** of data and algorithms

Consider an AI-driven proposal-screening system designed to conduct an initial review and automatically filter out lower-ranked applications based on text analysis. If the model is trained on past funding decisions, the tool risks reinforcing existing patterns, potentially favouring conventional research topics over novel ideas, or biasing against certain researchers based on their background.

To mitigate these risks, drawing on the principles and recommendations outlined in Section 1.3, the AI model should be trained on diverse datasets, excluding applicant characteristics, and be regularly audited for bias. It should function as a decision-support tool rather than an automated filter. Human reviewers should retain oversight and can override AI-generated classifications. Establishing fairness criteria is essential, with continuous monitoring to prevent systemic bias against underrepresented groups or research areas.

3.2 DATA

3.2.1. IDENTIFYING AVAILABLE DATA SOURCES

Once the project scope is defined, the next step is to evaluate the available data for AI tool development. This includes assessing both internal and external data sources. Here, we briefly describe common data sources. Part 4.2 elaborates on possibilities for sharing and reusing data across organisations.

External data sources. External sources provide data on various aspects of the research ecosystem, including on researchers, publications, funding, policy documents, patents, and media uses. Some datasets are freely and openly accessible, while others require subscriptions. Platforms such as CrossRef, OpenAlex and Dimensions provide comprehensive datasets that link publications, researcher characteristics, citations, and funding information. Alongside these, some platforms specialise in research metrics; for example, Overton focuses on policy documents, while PatentsView provides patent-specific data.

For an overview of commonly used external data sources, refer to Liu and colleagues' paper (2023), in which Figure 1 outlines data domains, key metrics and example sources. For instance, funding-related data come from platforms such as National Institutes of Health (NIH), Dimensions, CrossRef and UMETRICS, and include metrics such as principal investigator details, grant abstracts, funding amounts, and resulting publications.

Internal data sources. Internal data typically includes funding calls, funding proposals, applicant and reviewer characteristics, funding outcomes, and peer reviews. In particular, funders have access to submitted proposals outlining new research that are not systematically available to any other actors in research systems. For many research funders, who may hold decades of historical data, these represent an invaluable source of information for more data-driven operation with AI and ML. This data can be further processed to generate additional metrics, such as manually labelled research topics.

3.2.2 EVALUATING DATA FOR USE WITH AI/ML

Once potential data sources have been identified, the next step is to assess if they are fit for purpose for use with AI/ML, whether for training new models or applying existing ones.

Assessing fit and relevance. To assess data fit and relevance, examine the defined problem and how its main components can be measured and mapped to available data. For example, in reviewer matching, the core task is aligning reviewer expertise with the content of a funding proposal. A benchmark for reviewer-proposal alignment can be established using historical reviewer assignments as the ground truth. The internal data consists of funding applications and their assigned reviewers. Reviewer expertise can be represented through past publications sourced from external bibliometric databases.

Evaluating data quality. Once potential data sources are identified, it is essential to evaluate their quality, considering factors such as sample size, errors and noise, data format, biases, and time dependency. In the case of reviewer matching, the dataset must be large enough to ensure diverse and representative reviewer-proposal assignments. The quality of existing matches should be assessed, and misaligned pairings identified and removed. The format and structure of the text data should be evaluated – how easily they can be extracted and whether they can be consistently segmented into relevant sections for analysis.

Only after such evaluations are performed does a team have the necessary information to determine if the AI/ML application can proceed as planned, or if plans need to be adjusted because of limitations in the available data.

3.2.3 PREPARING DATA FOR ANALYSIS

Pre-processing. Preparing data for analysis often involves extensive pre-processing steps, such as standardising formats, linking different datasets using common identifiers, or even collecting additional data. For example, in reviewer matching, pre-processing may involve extracting text from funding proposals and reviewers' past publications. It also includes determining an appropriate time window for reviewer expertise – deciding how far back to consider publications to reflect current research focus. Additional data collection may be necessary, such as gathering feedback on reviewer performance.

Data management. Good data management practices are essential for responsible use of AI/ML. Many funder datasets will be proprietary, or will be used internally only, but effective preparation and organisation of data supports better experimentation, clearer

transparency and accountability in model development and use, and more effective reuse. The FAIR principles (GO FAIR, n.d.) are intended for data sharing, but offer a valuable approach even for internal-only data. Data should be:

- **F**indable: clearly organised and accessible to those who need it currently, or who may need it in future
- **A**ccessible: relevant staff within the funding organisation need to have appropriate permissions and infrastructure to access the data they need
- **I**nteroperable: datasets need to be prepared in such a way that they can be cross-referenced with one another and integrated into the same workflows
- **R**eusable: data should be well documented and described, to support later reuse as well as transparency for legal requirements and organisational responsibility.

3.3 TECHNICAL IMPLEMENTATION

3.3.1 BRINGING THE TEAM TOGETHER

Convening. Technical implementation starts with bringing together data scientists and engineers. Data scientists usually focus on formulating the problem, developing models, selecting the right algorithms, pre-processing data and optimising performance. Engineers are usually responsible for integrating the AI system smoothly into existing workflows.

Collaborating. Technical experts and teams should work closely with stakeholders, domain experts and end users throughout the development process. One effective way to secure stakeholders' engagement is by developing early prototypes that demonstrate practical use and actively gathering feedback throughout the process. As one data scientist noted: "For them [stakeholders] to prioritise spending time on this project, we needed to make them feel involved."

3.3.2 DEFINING THE PLAN: TECHNOLOGY, PROCESS AND INTEGRATION

A well-defined implementation plan should cover the choice of AI methods, model development process, appropriate evaluation metrics and integration into existing workflows. This stage should balance technical feasibility, ethical considerations and practical usability.

ML training requirements. Selecting the appropriate AI approach depends on several factors, including performance, explainability, computational cost, and ethical considerations. One decision to make is whether to fine-tune a pre-trained ML model or train a new model from scratch.

Fine-tuning a pretrained model means taking an existing model that has already been trained on a large dataset and adapting it to a specific task with additional, domain-specific data. This approach requires less data and computational resources while still achieving strong performance. An example is fine-tuning SciBERT for reviewer matching (Okasa and Jorstad, 2024). SciBERT is the go-to model for scientific text processing, pretrained on Semantic Scholar's 1.14 million scientific papers. Fine-tuning it on an internal dataset of past grant proposals and reviewer assignments allows it to learn patterns specific to the task. However, caution is needed because the corpus is heavily biased towards life sciences and computational sciences. SciBERT might not be appropriate for tasks related to social sciences and humanities.

Training a model from scratch involves collecting a high-quality large dataset, choosing the algorithms, and training it to recognise patterns specific to the task. While more resource-intensive, this approach offers full control over model design and inputs. It may also be preferable in cases where transparency is prioritised. For instance, Thelwall et al (2023) developed a model to predict the quality of research publications using textual and bibliometric features, deliberately opting not to use pretrained models such as SciBERT. Instead, the team trained its models using interpretable features such as unigrams (single-word terms), bigrams (two-word terms) and journal names – allowing it to understand which elements of the text influenced the model's predictions.

Algorithms and architectures. Following the choice between fine-tuning a pretrained model and training one from scratch, the next step is determining which type of algorithm and model architecture best suits the problem. Machine learning methods

range from simple linear models to complex deep neural networks, each with different trade-offs. Besides performance, interpretability is a critical factor for choosing the algorithm, particularly for decision-support tools in high-stake scenarios. Returning to the earlier example of assessing research quality using AI, it is essential to understand which factors drive the decisions if the assessment is used to determine future funding allocation. Highly complex models, such as deep learning architectures, often struggle with interpretability. In general, simpler models should be preferred if their performance is sufficiently strong, as they are computationally efficient and easier to interpret.

Evaluation. Robust evaluation metrics must be defined to assess the effectiveness of the AI tool. Model performance should be measured through traditional metrics such as accuracy, precision, and recall using a hold-out validation dataset. Beyond raw performance, however, it is important to conduct real-world validation. For example, in AI-assisted proposal screening, the model's outputs should be compared against human assessments to check that important but less conventional research is not deprioritised unfairly.

3.3.3 IMPLEMENTATION CYCLES: ITERATIVE DEPLOYMENT AND REFINEMENT

AI implementation is rarely a one-step process; gradually rolling out the system and iteratively refining it is key to ensuring the system is effective and well integrated into real-world workflows.

Pilot testing. The first stage of implementation is applying the AI tool to a small set of pilot cases, allowing for a focused evaluation before broader deployment. This phase helps identify early challenges, usability issues and potential biases in the system. User feedback from domain experts – such as scientific officers in the case of grant-review applications – provides critical insights that guide refinements. For example, in an AI-driven reviewer-matching system, an initial prototype may rely on unigrams to assess topic expertise. However, pilot feedback might reveal that the same words can have different meanings across disciplines (eg, neural 'network' in neuroscience vs social 'network' in sociology). An improved iteration could incorporate scientific bigrams or phrase-level embeddings to better capture disciplinary nuances.

Beta testing and refinement. After successful pilot testing, the AI tool can be expanded gradually to a wider group of users while its functionality continues to be refined. The expansion process means scaling up data inputs, testing the model's robustness across different cases, and adapting it to domain-specific variations. For example, an AI-driven proposal-screening tool initially tested in one funding programme may later be applied to multiple research domains.

User training. Once the AI tool is fully deployed across an organisation, user training is important for its successful adoption. Hands-on training sessions, interactive guides and ongoing support allow the users to trust the system and understand its limitations. Proper training enables users to incorporate AI into their workflows effectively while remaining critical of its outputs, preventing over-reliance on automated recommendations.

3.4 EVALUATION AND MANAGEMENT

3.4.1 ASSESSMENT AND REFINING

Automated and comparative assessment. Once the AI tool is fully deployed across an organisation, continuous monitoring and evaluation become essential. Regular assessments should measure performance metrics, such as accuracy, precision, recall and F-measure, to confirm that it continues to meet expectations. A common practice is to periodically collect a subset of cases and evaluate them using conventional, non-AI methods, and then compare outcomes. For example, in categorising funding proposal topics, a sample of proposals should be manually classified by scientific officers at regular intervals and compared against AI classification.

System analysis. Understanding where the model fails or underperforms helps refine and improve it. Regularly reviewing misclassifications, false positives and false negatives provides insight into areas where the model needs tuning. In a reviewer-matching AI, if certain fields (eg, interdisciplinary research) consistently receive poor matches, this could indicate the need for

retraining the model with better-labelled datasets or adjusting how reviewer expertise is represented.

Assessment over time. AI systems must also be maintained over time and refined based on how well they work in real-world settings. Developers must proactively track how the model is being used over time. One consideration is whether the scope of the data has shifted – for instance, a model originally developed for biomedical proposals may end up being applied to social sciences. Channels for structured user feedback should be established to identify usability gaps and unexpected AI behaviours. An effective approach is to enable real-time error reporting, allowing users to flag AI mistakes as they occur. For example, scientific officers could report incorrect topic classifications or reviewer assignments, providing valuable data to detect patterns of misclassification and refine the model accordingly. Additionally, AI systems should adapt based on the actual workflow. A peer-review automation tool, for instance, may initially suggest high-quality reviewers, but feedback may later reveal that senior researchers are less likely to accept review invitations. Incorporating this feedback by giving preference to qualified junior researchers could increase the reviewer acceptance rates while maintaining quality selection.

3.4.2 RECORDING, DOCUMENTATION AND REPORTING

Effective AI management requires clear and consistent documentation. Two widely recognised frameworks – datasheets (Gebru et al, 2021) and model cards (Mitchell et al, 2019) – offer structured approaches to recording details about AI systems.

Datasheets provide a systematic way to document datasets, detailing data sources, collection methods, intended uses, biases and limitations. By including standardised information, datasheets help AI practitioners and stakeholders assess whether a dataset is appropriate for a given application and identify potential risks associated with its use.

Model cards serve a similar purpose for AI models. A model card should detail the model's purpose, intended applications and potential risks, including biases and ethical concerns. It should also provide information on training parameters, training data and evaluation results.

Versioning and change tracking are important. As AI models evolve through fine-tuning and retraining, documenting updates, parameter changes and dataset modifications ensures that past decisions remain interpretable. The use of version control systems such as Git is invaluable for tracking changes in the source code for AI systems, but this must be paired with good data management of model files and training logs.

Stakeholder reporting. In addition to internal documentation, reporting to stakeholders – such as funding bodies, administrators or policymakers – can enhance trust and accountability. Reports should communicate how the AI system is performing, perceived challenges and planned improvements, ensuring that stakeholders remain informed about the system's reliability and impact.

"Documenting updates, parameter changes and dataset modifications ensures past decisions remain interpretable"

3.4.3 CONTINUOUS IMPROVEMENT

AI tools need continuous improvement to align with evolving needs, changing technologies and new developments in the data they analyse.

Retraining. Consider retraining the model periodically so that it reflects current trends, new data distributions and shifting user needs. Over time, research landscapes change and older training data may no longer represent the diversity of topics, funding priorities or reviewer expertise. Retraining on newly collected, high-quality data helps mitigate these issues. For example, in a grant proposal categorisation system, updating the model with recently submitted proposals helps represent emerging research areas.

Benchmarking. Beyond retraining, AI systems should be benchmarked against state-of-the-art models to assess whether training new models on the same data could enhance performance. Advances in large language models, domain-specific pre-trained models or more interpretable architectures may offer improvements in efficiency, fairness or transparency. If newer techniques outperform existing models significantly, evaluating the feasibility of adopting them – through fine-tuning or replacement – becomes essential.

Organisational perspectives, cooperation and collaboration

Research funders exploring or applying AI/ML are faced with important organisational concerns in addition to technical implementation. Some of these are specific to the research-funding context, while others are more general concerns faced by any organisation using AI/ML.

Here, we summarise five key organisational concerns that research funders face when developing, using and evaluating AI and ML tools:

1. Keeping humans in the loop with AI/ML applications
2. Bringing AI expertise together
3. Competency-based collaboration in AI/ML teams
4. Cross-funder cooperation and reuse of AI/ML resources
5. Guiding AI use in research systems

For each of these themes, we summarise challenges and key questions arising in the practice of developing and using AI solutions.

4

4.1 KEEPING HUMANS IN THE LOOP

Keeping the 'human in the loop' is a common phrase used in the development of AI tools. This generally refers to the inclusion of human judgement alongside AI (eg, having human oversight of decisions where AI is involved), but it is important to include human critical perspectives at all stages of the AI process.

4.1.1 WHAT HUMAN-IN-THE-LOOP LOOKS LIKE

Human expertise may be 'in the loop' at many different points in the development and application of an AI system.

- At the outset of a new application, human expertise is vital to starting with people, not technology. Rather than seeking out a use for an existing tool, human knowledge and experience (eg, of funder staff) is the best starting point to identify an existing problem and then seek out a solution. If appropriate, that solution may involve AI/ML.
- In identifying data sources, human expertise helps to select appropriate data, ensure a good fit with existing workflows, and check results for appropriateness and fitness for purpose.
- In the process of training, validating and adjusting ML models and AI systems, human expertise is the best guide to which differences will have a meaningful impact on the organisation and which will not.
- When an AI/ML system is being applied, human expertise can provide an essential oversight role, ensuring that system outputs are used appropriately, checking to make sure outputs make sense, and providing additional accountability for sensitive decisions.

Human-in-the-loop is more of a spectrum of strategies and decisions than a single approach. The key organisational decision for each new application of AI/ML systems is to identify the degree to which systems are:

- Human-**in-the-loop**: with active interaction and judgement from human experts, and intervention to control and adjust what an AI/ML system does
- Human-**on-the-loop**: human experts serving in a supervisory or oversight role, but not involved in the decision-making process
- Humans-**out-of-the-loop**: when AI/ML systems are autonomous and human experts simply make use of system output.

Each of these may be used at different times and for different purposes, including in an iterative process of refining a funder's model for human-AI collaboration. Each of these decisions, however, is an essential part of managing the technical and political trade-offs involved in AI/ML use.

4.1.2 EXPLAINABILITY AND TRANSPARENCY

Explainability and transparency are essential parts of keeping the human in the loop, as well as ensuring a funder's legal responsibilities are upheld when AI is used. The benefits of AI tools need to be balanced with the challenge of human users not always having the necessary information to explain how results are produced by AI use.

Explainable AI is an active research area encompassing a wide variety of methods (Burkart and Huber, 2021). While all explainable AI methods have their limitations, they can provide valuable insights into what a model is responding to when producing a particular output. This can be very helpful for dealing with the more unstructured way ML models work in comparison with structured human reasoning, particularly for highly complex model structures such as LLMs.

However, explainability of AI outputs is only part of the picture. Research funders making use of AI/ML must also be able to provide transparency into the wider process of how AI is used – for example, how the system was designed, what data was used in training, and how the outputs of AI systems are integrated into decision-making processes. A broader practice

of transparency around AI/ML use is therefore essential for research funders, including clear and regular reporting on the design, implementation, integration and ongoing assessment of AI systems in practice.

Research funders have, in many cases, a legal duty to be able to explain their decisions to the public and their applicants. For example, the European GDPR legislation imposes a legal requirement of explainability on any automated system deciding on matters of importance to citizens. In parts of the funding process that do not involve automated decision-making, such as scoping calls for proposals, AI tools may be used without GDPR implications, but transparency remains a best practice for the organisation, researchers and other stakeholders in the research system. Funders also bear responsibility for maintaining public trust in research systems, and explainability and transparency are key elements of supporting this trust. Importantly, non-explainable AI can align with accountability and transparency ideals if its use is well documented, with risks evaluated and mitigated.

4.1.3 STRATEGIES FOR DEVELOPING HUMAN-IN-THE-LOOP APPROACHES

Clear strategies for communication and collaboration between stakeholders are important to manage risks with new systems and develop the best AI solutions. It is important for data teams to ask specific questions of end users to gain specific feedback. It is also essential to integrate users into the process early. For example, an impact exercise when planning an AI intervention must be sensitive to the role of the proposed AI system within the business and what the real problems will be, and developing this understanding and effectively managing expectations around AI use requires actively engaging with how users work and how the product will be used in day-to-day work. This understanding helps to identify approaches for human oversight and intervention in AI use that will work with established processes and ways of working.

Funders should also set expectations around how those responsible for developing AI solutions will communicate with colleagues and what level of detail they will provide on methodologies, internally and externally. For any AI application by a research funder, there may be many interested people who might want to explore AI use in their own contexts. It is therefore prudent to have accessible descriptions prepared that will meet this information need and build trust in the AI system, and in the process of developing it. One good strategy is to develop an illustrative example of what using the AI system would look like in practice – eg, explaining a simple algorithm using a table. This is especially important with AI applications that are likely to be used repeatedly and/or in different contexts. If developing a system for matching expert reviewers, for instance, an illustrative example will show the detail of how this works to any interested parties and help clarify how human oversight is involved.

> "Developing effective, ethical and equitable AI applications requires bringing the right people together from the beginning"

4.2 BRINGING AI EXPERTISE TOGETHER

Developing effective, ethical and equitable AI applications requires bringing the right people together around the discussion table from the very beginning. Funders face three key challenges in this area:

1) Balancing internal and external AI expertise
2) Engaging stakeholders from around the organisation
3) Managing interprofessional and transdisciplinary working

4.2.1 BALANCING INTERNAL AND EXTERNAL AI EXPERTISE

Like any organisation, research funders must always balance trade-offs in how they use resources, including staff time. To explore AI use, funders may make use of relevant internal expertise, but often also need to source external expertise (particularly technical expertise). Internal and external expertise each have their own benefits and drawbacks for AI applications, which often come with tight timelines and high expectations.

Benefits of engaging external expertise:

- External consultants specialise in building something that addresses a brief and delivering a product that shows immediate value.
- AI solutions developed by external experts can be adapted to the internal context by in-house 'champions', people who are invested in AI use and willing to experiment.
- External expertise is invaluable in getting started with AI, and funders can learn from consultants and build from there.
- Good relationships can be developed with external partners over a longer term, whereby product development becomes a shared learning process akin to working with internal colleagues.

Benefits of investing in internal expertise:

- It is difficult to get the full benefit of an external product without the internal competencies to take over the work and adapt and refine it over time.
- A product from an external consultant will provide an immediate solution, but may not be maintainable or easily reused. Funders must be able to answer the question of what happens when an external consultant goes away.
- External consultants do not necessarily know the context of the organisation and the full background of the range of operations within research funders. Internal teams are better prepared to explain AI solutions to colleagues and help build trust in the use of AI.
- Internal teams are better placed to reuse effort: for example, if an internal team produces one AI system and multiple additional users become interested, they can adapt and expand the solution without rebuilding for multiple clients.
- Internal experts understand how to use a funder's complex data and the value of using this data internally. Processes such as standardising data, capturing data, and storing and managing data need internal expertise. Internal data scientists can also benefit the organisation by raising awareness of how to use internal data.

In practice, funders can combine elements of internal and external expertise. Combining in-house development with targeted projects done by external consultants can help to strengthen understanding of AI and build up methodological competencies within funding organisations, and can help to use data and expertise that external consultants may not have. The balance between internal and external expertise can, and does, shift over time: for example, many funders around the globe are working to grow their internal expertise around data and AI (Rushforth et al, 2025).

Engaging internal stakeholders is vital for navigating internal and external competencies. Internal customers can identify needs for new AI and data development, challenge what is being provided, and shape the production of bespoke solutions.

4.2.2 ENGAGING STAKEHOLDERS

Stakeholder engagement is a vital part of navigating AI exploration and application. This includes engaging internal stakeholders – such as scientific officers, strategic decision-makers and policy experts – as well as external stakeholders, such as researchers, professional staff at research organisations, science policymakers, and members of the public.

Planning for stakeholder management is essential, as it is time-consuming and complex to navigate the network of stakeholders for a given AI application. There are also many different demands to manage, including strategic and operational considerations and managing risk. Many funders are developing frameworks for stakeholder engagement based on their own structures and ways of working, using methods such as consultative workshops, organisational roadshows and embedding AI experts within teams.

This engagement is needed at all stages, from scoping a potential application of AI/ML through to assessing the impact of a deployed AI/ML intervention. Focusing on specific use cases and the user journey is a powerful enabler for a partnership-based approach, which

is more likely to achieve successful AI/ML applications. The engagement process is also necessary to put AI/ML use in context, including identifying existing structures, processes and policies that can inform an application, as well as pinpointing the specific utility that AI/ML can bring. Proactive and consistent engagement with the diverse stakeholders of AI/ML use helps to establish clear lines of accountability, and ensures minimal disruption, less duplication and quicker progress.

4.2.3 INTERPROFESSIONAL AND INTERDISCIPLINARY WORKING

AI can be considered a knowledge technology as much as an information technology. AI/ML use involves collecting and processing information to produce knowledge and action: as a result, using AI/ML inherently requires working across boundaries and skill sets. Technical teams that manage infrastructure and implementation must work with data teams that understand the data funders have and how it is managed; operational experts who know the processes in which AI/ML is to be used; and strategic roles with oversight of organisational goals.

Working across these boundaries is an important challenge. Individuals with different backgrounds and roles may have different notions of AI performance and of what is considered a correct or desirable output from AI use. Exploring and applying AI will always be a team-based process: the first step of any AI team is to establish a shared understanding of the goal for AI use and how the outcomes will be assessed. Teams must also recognise that individuals may understand AI terms and concepts differently, so the process of working together needs to include navigation of these differences.

A powerful tool for structuring the process of working across professional and disciplinary boundaries is to focus on the specific decisions and competencies involved in AI application, discussed in the next section.

4.3 COMPETENCY-BASED COLLABORATION IN AI TEAMS

Once AI/ML expertise has been brought together around the table, the next challenge is to work together effectively. This includes addressing two key aspects of the competencies needed to drive effective, ethical and equitable use of AI/ML:

1) How AI/ML will be used in practice, as part of a funder's wider work.
2) Whose input should inform or affect AI/ML use in practice.

"The first step of an AI team is to establish a shared understanding of the goal for AI use and how the outcomes will be assessed"

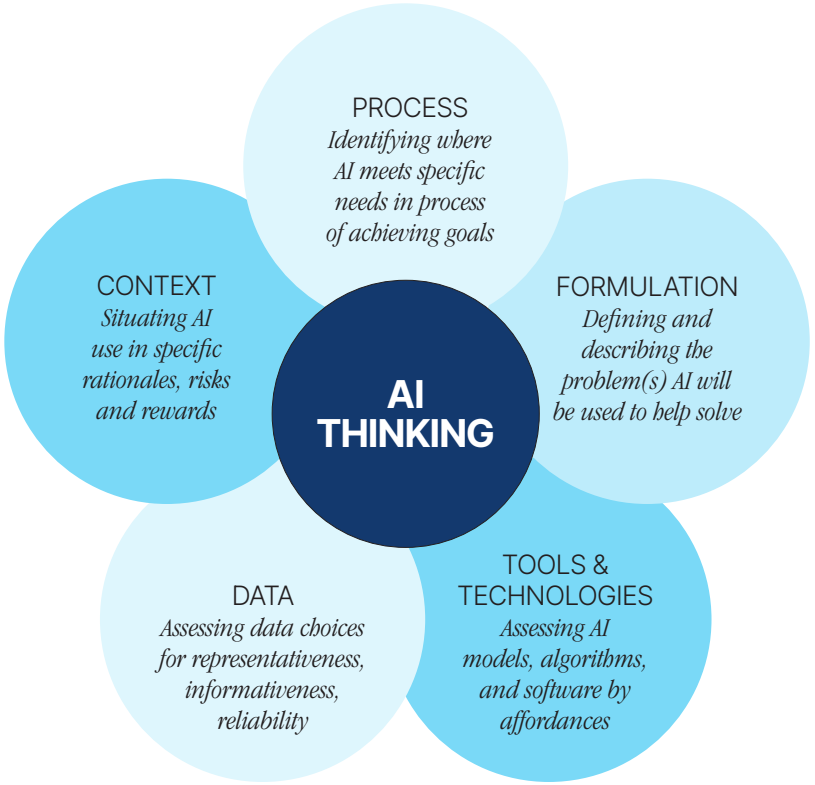


FIGURE 4.1. THE AI THINKING MODEL (REPRODUCED FROM NEWMAN-GRIFFIS, 2025)

4.3.1 HOW: AI THINKING FOR USE IN PRACTICE

AI and ML can mean many different things to different people. For some, AI is technical; for others, it is an area of policy and regulation; for many, it is a tool to use for other purposes. These different lenses on AI relate to differences in disciplinary and professional background, and can lead to people speaking at cross-purposes.

The AI Thinking model (Newman-Griffis, 2025) provides an example framework for developing shared understanding of how AI is used in practice, to help bridge these differences and improve communication. The model reflects the multiple competencies that are needed in a team to make AI work, and is a good starting point to help assemble and structure a team, and recognise each member's contribution.

AI Thinking describes five primary aspects of AI use in practice:
Process: identifying where AI use would meet the needs of specific processes to achieve organisational goals

Formulation: defining and describing the problem(s) AI will be used to solve

Tools and technologies: assessing different AI models, algorithms and software in terms of their fit with the target application and goals

Data: assessing the data used for training, evaluation and application for representativeness, informativeness and reliability

Context: situating AI use in specific organisational goals, operational and policy contexts, and risks and rewards

This framework provides a springboard for funders to bring together an AI team and ask application-specific questions when exploring a new application of AI/ML. As a tool for reflection on the AI application process, AI Thinking can also help funders evaluate and identify impacts of AI use, track and critique decisions, or untangle the compound effects of AI use across the organisation.

4.3.2 WHO: EXPERTISE AREAS IN FUNDERS INFORMING AI/ML USE

To answer these questions about AI/ML use in practice appropriately, funders need to draw on a wide range of expertise within the organisation. Expertise is needed to address three primary areas:

1. ***Governance and oversight*** of AI use, in organisational, national and international contexts.
2. ***Development and use*** of AI, including bespoke in-house guidance on buying AI solutions, the design and build of AI agents, and training and evaluating solutions.
3. ***Efficiency*** of AI adoption and integration, including leveraging existing organisational policies and processes for procurement, roll out and testing, data protection assessments, training and change management, and so on.

AI/ML use is not only the domain of technical and data teams therefore. It encompasses input from, among others:

- **Information technology**: including the local computing hardware, cloud computing solutions, AI implementations, and software management
- **Data experts**: including data collection, management and

integration, as well as information governance practices
- **Scientific officers**: to provide subject-matter expertise in the areas in which AI will be applied, including expertise on funding processes
- **Operational experts**: to understand and inform how AI/ML use will interact with the complex network of existing operational processes within a funder
- **Strategic experts**: to articulate how AI/ML use fits within the strategy of the organisation, and how it will help or harm the achievement of a funder's long-term strategic goals
- **Legal and regulatory experts**: to identify potential regulatory impacts or concerns from a specific application of AI/ML
- **Finance experts**: to inform appropriate levels of resourcing that are available to support a particular AI/ML application
- **Human resources experts**: to identify and manage the diverse expertise informing an AI/ML application.

By broadening the sense of whose perspectives are relevant and necessary to inform a particular AI/ML application, funders will be better prepared to integrate AI/ML into their workflows effectively, and to proactively manage ethical considerations and equitable impact from AI/ML use.

4.4 CROSS-FUNDER COOPERATION AND REUSE

Research funding organisations vary greatly in the types of research they fund, the level of human and financial resources they have, and the size of organisation. They also operate under different legal frameworks depending on national and international context and whether they are publicly or privately funded. However, they share similar missions, challenges and experiences, and bridging the differences between funders to create shared, cooperative learning about AI/ML has the potential to produce valuable insights and improvements in research systems.

As a sector with shared missions and concerns, funders have the opportunity to share:
- The **data** that drives AI/ML
- The **tools** that implement AI/ML
- The **knowledge** of how to use AI/ML responsibly.

4.4.1 DATA

Funders hold substantial data to which no other actors in research systems have access. This includes the full details of funding applications (successful and unsuccessful), research assessment exercises, and records of research portfolio management. Sharing these resources with other funders could help all research-funding organisations gain better insights into the research they fund, and support better-informed AI/ML models. However, most of these data are provided to funders – or created by them – in strictest confidence, and funders have a legal and moral responsibility to protect the security, privacy and integrity of these data. It is therefore highly challenging for funders to share data with each other, whether across national boundaries or between public and private funders. Here, we highlight emerging strategies for funders to overcome these barriers and pursue more open data sharing.

Open datasets. The OpenAlex database is a large, open-access resource of research works, and already a valuable source of data for funders. Funders have the opportunity to share non-sensitive data with each other by enriching OpenAlex or similar datasets, or to create an open-access database specifically for funders. Standardised open datasets are invaluable for funders to develop and test analyses before adapting their methods to their own internal use cases with private data. Any open datasets should adhere to FAIR principles to ensure effective reuse.
See more: OpenAlex – **openalex.org**

Secure, anonymous data. Anonymising internal funder data into a common format enables cross-funder analysis, and can support further linkage to publicly available data on research outputs and outcomes for extended analysis. The RoRI Funder Data Platform provides a platform and legal framework for funders to share their internal data for analysis on specific problems, topics and research questions. It also facilitates opportunities for standardising and anonymising data, and then the possibility of linking closed and open data.
See more: RoRI Funder Data Platform – **researchonresearch.org/project/funder-data-platform**

The need for ongoing resourcing. Infrastructure such as shared repositories needs to be maintained, costing time and money. The RoRI Funder Data Platform provides some of the benefits of an open platform, but its primary role is as a secure platform to conduct specific analyses using confidential data on a project basis. Expanding this or similar infrastructure to a repository model that people apply to access creates additional value, such as enabling more systematic study of funding data and analysis of what is shared, or different, between different funding contexts. This type of effort requires significant resources to develop,

however, and an ongoing commitment to maintain.
Example: The Swiss National Science Foundation provides data, including textual data, for approved grants publicly via the Data Portal: **data.snf.ch/datasets**

4.4.2 TOOLS

As well as data, there are benefits to be realised from sharing AI/ML tools between funders. These may include:

Pre-trained models for adaptation by other funders. There are multiple web platforms available for sharing and distributing pre-trained machine learning models, which can form the basis of targeted applications of AI/ML via fine-tuning. These models can be used without sharing underlying data and/or code, though some models may have the potential to leak sensitive information and must always be examined thoroughly.
Example: The Swiss National Science Foundation shares pre-trained models on Huggingface: **huggingface.co/snsf-data**

Off-the-shelf software for direct reuse. Additional engineering is required to make a pre-trained ML model directly usable as a dedicated software application, but it can produce enormous value by removing technical barriers to access of the models and software. This is a common approach for research tools.
Example: CWTS Leiden has released VOSviewer as a standalone application: **www.vosviewer.com**

Source code for AI/ML applications. Sharing source code provides transparency into the operation of AI/ML systems and serves as a starting point for other funders to adapt an existing approach. Though the utility of source code for ML models is limited without the data, the code nonetheless provides valuable opportunities to enhance and adapt approaches, as well as learn more about the research system. The Git version control system is an essential resource for sharing source code within a team and outside of a team.
Example: The Swiss National Science Foundation shares its code publicly on GitHub: **github.com/snsf-data**

4.4.3 KNOWLEDGE

Finally, funders benefit from sharing knowledge of how to conduct AI/ML experiments, use AI/ML to improve processes, communicate with stakeholders about AI/ML use, and many other aspects of practice.

Knowledge is, in some ways, the least complex resource to share, but the one that most lacks clear mechanisms for sharing. Here, we highlight three ongoing mechanisms for funders to consider in taking their own actions on AI/ML knowledge exchange:

> "Knowledge is the least complex resource to share, but the one that most lacks clear mechanisms for sharing"

Workshops and discussion forums. RoRI's GRAIL project has provided an international forum for funders to share knowledge and experiences about AI/ML use with one another. A similar, workshop-based structure can help funders share focused discussions with one another to address particular aspects of AI/ML in the work of funding and assessment.

Publications and reports. Some funders are already publishing on their explorations and applications of AI/ML methods. Others have issued technical reports, or presented in conferences and meetings with other funders. These are invaluable mechanisms for disseminating aspects of AI/ML knowledge that are non-sensitive and open for public consumption.

Working groups. Working internationally helps funders connect with each other across the global research ecosystem and build relationships that can support further knowledge sharing. For example, discussions in GRAIL have contributed to working groups associated with the Barcelona Declaration, helping to raise awareness and promoting discussion about funder data sharing.

4.5 GUIDING AI USE

Research funders have a role in setting good practice and clear guidance on AI use in two ways: internally to the organisation (ie, use of AI by funder staff, peer reviewers, etc) and externally within broader research systems. Funders are key to shaping broader research culture, and clear policies around AI use are important for informing the role of AI technologies in research.

Guidance on AI may come in many different forms: as standalone documents, position statements, or as part of wider IT or data policies. In many cases, important aspects of AI use are already addressed in existing policies, particularly on data privacy and confidentiality. Currently, policy development on AI is focused on the use of generative AI, but funders should also think more broadly about the use of other kinds of AI systems in research.

Developing policy and guidance must be specific to the needs and challenges of individual funders, but, here, we address some shared starting points:

1) Example approaches to developing AI policies
2) Key considerations on developing AI policies and education resources
3) Guiding principles for AI use in funding contexts

4.5.1 APPROACHES TO DEVELOPING AI POLICIES

Many aspects of the use of AI in research funding may be covered by existing policies, but new policies will be needed to address some gaps. Two examples illustrate what this process may look like:

Dutch Research Council (NWO). The NWO guidance on the use of GenAI (NWO, 2024) was underpinned by the Netherlands Code of Conduct for Scientific Integrity and informed by other relevant policies. The spark for creating the guidance was an observation that the use of GenAI was increasing greatly in the scientific community and that there was a need to make a statement on how these technologies should be used in the funding process. The aim was to give a balanced view to recognise the potential value of GenAI, but also recognise the risks to security, privacy and intellectual property (IP).

Social Sciences and Humanities Research Council of Canada (SSHRC). The Draft Guidance on the Use of Artificial Intelligence in the Development and Review of Research Grant Proposals, created jointly by SSHRC, CIHR and NSERC (Government of Canada, 2024), was also informed by existing guidance coming from the centre of government, such as a Directive on Automated Decision-Making, aimed at making service delivery more efficient within the core principles of administrative law, and the Guide on the Use of GenAI within government itself, which encouraged exploration of potential uses of AI to improve operations while being risk aware.

Although generated in different geographical and organisational contexts, these two cases followed a similar process for policy creation: the new policy met an identified need not provided by existing policies, but complemented and built upon what was already in existence. Both organisations are public bodies and draw on their respective government's existing policies, such as on research integrity, data protection and IP. Overall, policy development benefits from multiple stages of the input and feedback process. For AI guidance, targeted guidelines for different stakeholder groups – eg, applicants and employees – is more useful than blanket statements.

4.5.2 KEY CONSIDERATIONS FOR AI GUIDANCE

1. **Staying relevant** in the fast-moving AI landscape. Guidance that is overly specific about particular technologies rapidly becomes out-dated. Guidance must be relevant and specific enough to be useful, but not ephemeral because of a high level of granularity or reference to specific technologies. A process-based approach can help achieve this.
2. **Responding to evolving AI misuse** in funding processes such as application and peer review. Funders may opt to establish general principles and give examples, or take a more thorough approach and list all examples of misuse cases as they are observed.
3. **Disclosure requirements** are necessary to maintain transparency and accountability, but funders need to set guidance on what needs to be disclosed and with what attributions.
   - **a.** Students and trainees often play a role in application preparation, as well as AI.
   - **b.** Guidance should not bias reviewers against applications using or not using AI.
   - **c.** AI technologies may be used for many different purposes: spellchecking, grammar correction and style correction are very

different from using a tool to generate new content.

4. **Disclosure mechanisms** for reporting AI usage by applicants, reviewers and other AI users. This may be done through tick boxes, fixed sets of options, or more detailed descriptions of usage.
5. **Educating applicants, reviewers and other users** as to which tools involve AI, particularly generative AI. Users may not disclose if they are unaware that they are using an AI application.
6. **A joined-up approach** between research funder guidance on AI usage and disclosure and guidance provided by research-producing organisations. Researchers may receive conflicting guidance and funders need to be clear on what is required (eg, for applying for funding) and what is a recommended practice.
7. **Research community concerns** about AI use and potential negative impact on research integrity and originality, and about privacy and potential IP loss. Funders can help to set expectations and best practice on these concerns.
8. **Role of AI in equity and accessibility**. Generative AI technologies have significant potential to help close accessibility gaps and improve equity in research systems – eg, for neurodiverse researchers or those working in a non-native language. However, achieving these benefits requires access to AI tools and the skills to use them, which funders can help support.
9. **Perceptions of and trust** in AI use are essential parts of addressing wider ethical concerns around AI. For example, researchers, funders and others may differ on whether it is appropriate to use a generative AI system as part of assessing a funding application. Researchers may also have conflicting feelings on if they were assessed by an automated system, or if a competitor was successful in a funding application that they prepared using AI.
10. **Reliability and provenance of information**. Generative AI systems are known to 'hallucinate' inaccurate or false information, and the probabilistic nature of generative AI platforms means that the same prompt may produce two different answers. Funders need to identify their desired interventions via policy to regulate AI use and education to help researchers be better informed.

4.5.3 STARTING POINTS FOR AI GUIDANCE

As a starting point for developing AI guidance specific to their national and international contexts, funders will benefit from starting with existing resources, such as the European Commission's Living Guidelines on the Use of Generative AI in Research (European Commission, 2025). These provide broad guidance to researchers, research organisations and research funders, centred on four key properties of ***reliability***, ***honesty***, ***respect*** and ***accountability***.

As funders adapt and expand AI resources, they should provide clear guidance on the mechanisms for engaging with and demonstrating these principles for responsible use. We emphasise four essential aspects of this:

Attribution. Actors in research systems need clear principles and processes to guide attribution of authorship and ideas. Many funders have developed guidance for co-authorship and teams; similar guidance is needed for attributing contributions from human authors and AI systems. The iterative process between a researcher and AI technologies means this will rarely be clear-cut, so funder guidance must be flexible.

Accountability. Funders must clearly establish who bears ultimate responsibility for research applications, outputs and material. Researchers retain ultimate responsibility for what they produce, so the use of AI does not remove the requirements of funding applications or other processes in terms of privacy, confidentiality, data security and protection of IP.

Data protection. Research-funding processes involve many actors working with confidential, proprietary and/or otherwise sensitive information. These include funding applications, reviews, research outputs, administrative documents, and more. All actors – including applicants and authors, reviewers, funder staff, consultants, and other external assessors – must have clear guidance on what is and is not permitted regarding the use of AI technologies to handle sensitive data in funding processes. For example, no one should input application information into commercial generative AI tools, as this can result in a breach of privacy and the confidentiality of intellectual property.

Transparency. Researchers need to have clear criteria and mechanisms for transparency about AI use. This includes reporting where and how AI tools were used – eg, to help produce funding application materials or research outputs, or in conducting peer-review activity (where permitted). Guidance from funders should provide clarity on what is expected, how it should be reported, and who may make use of the reported information.

Case studies

Many research funders around the world are actively using AI/ML technologies or have experimented with them in the past. Learning from the experiences of these funders is the best way to illustrate how the principles and processes described in this handbook can be put into practice, and how other organisations have managed the complexities of integrating AI/ML into their work.

Here, we present five case studies shared by research funders as illustrative examples of AI/ML use in practice:

1. Swiss National Science Foundation
2. 'La Caixa' Foundation
3. Novo Nordisk Foundation
4. Research Council of Norway
5. UK block grants funders

5.1 WHAT WE AIM TO LEARN FROM CASE STUDIES

The pace of AI adoption varies across research funders: some funders are established experts in AI/ML use, while others are just beginning to explore potential AI/ML applications. However, there are shared patterns in how expert knowledge, organisational practice and AI have been integrated and used as AI has entered the research funder's toolbox. In addition to the technological aspects of AI, funder experiences illustrate the human and organisational perspectives involved in AI/ML use. The use of AI is increasingly part of organisational strategies, and AI use is proliferating across departments and work areas. As AI development has become more systematised, processes have been developed to aid transparency and better interprofessional communication between technical and non-technical personnel (see the resources discussed in Part 3.4).

In these case studies, we describe five different applications of AI/ML to support the work of research funders. These case studies illustrate a range of goals, funders, strategies and funder types. Each case study describes the following elements:

- **Motivating factors** behind the reported use of AI/ML
- **Data used** for training, analysis and/or evaluation with AI/ML systems
- **Implementation overview** of how the AI/ML system was developed and put into practice
- **Evaluation/management strategies** for assessing the AI/ML application and managing its use over time
- **Organisational reflections** on the experience of this particular use of AI/ML methods

The purpose of this chapter is to give a flavour of the challenges and potentials of AI to support the work of research funders, and to share some general insights into the issues that must be considered by funders, including some that might be easy to overlook at first.

5.2 SWISS NATIONAL SCIENCE FOUNDATION: MACHINE LEARNING FOR REVIEWER MATCHING

Matching reviewers to proposals is a common and labour-intensive task for research funders. Although Google and expert recommendations are still used, computer-assisted methods are valuable and increasingly used. The issue is similar to the process of finding reviewers for submissions to academic journals. This case study describes how the Swiss National Science Foundation (SNSF) uses AI to help match evaluation panel reviewers with proposals.

Motivating factors

Matching potential reviewers to proposals is a core process for SNSF, so making this process more effective or efficient is a clear win for the organisation. It is also a natural task for automation, given that it is based, at least partly, on semantic similarity – between the reviewer's expertise, as expressed through their previous publications, and proposal texts.

Data used

The titles and abstracts of the grant proposals are used to capture more information about the proposals than just their keywords.
The titles and abstracts of reviewers' previous publications are used. Although full texts would give richer data, they are not always available online or in a format that is easily usable for natural language processing. This data is accessed from Dimensions.ai via its API, making it relatively easy to obtain. Which subsets of publications should be used to best represent each individual reviewer is an open question (how many years of data, how many publications should be of a similar topic, etc).
This case study is implemented for the situation where all potential evaluation panel reviewers are known in advance, and their bibliometric data can be downloaded and processed together.

Implementation overview

The SNSF Data Team includes research data scientists with experience in machine learning and text analysis.
A Python script (shared on Github) was used to identify reviewers.
Overall, the grant proposals and the reviewer text data (titles and abstracts from all publications combined) were converted into vectors encoding their content using a variety of methods.
The text similarity between reviewer publications and grant proposals was then calculated with a standard metric, the cosine similarity measure. The text-encoding methods tested were basic bag of words (recording every word in a different vector position, weighted using the term frequency–inverse document frequency [TF-IDF]) and several variations of semantic word embeddings (transformer models representing each word with a complex vector encoding itself and its meaning within the text). In theory, the latter methods should give a much more precise and context-aware encoding of the texts into numeric vectors.
The overall system attempts to solve the entire reviewer-proposal matching problem in one combined and fully automatic procedure, by incorporating

rules that balance the reviewers' individual workloads and avoid conflicts of interest.
The final stage (human-in-the-loop) is the checking and validation of the results by SNSF scientific officers.

Evaluation/management strategies

This use case is already used in practice, although experiments aimed at improvement are ongoing.
Different word embedding algorithms were tested (pre-trained transformer models: BERT, SciBERT, SPECTER from huggingface.co). Experiments were applied separately to the life sciences (LS), mathematics, informatics, natural sciences and technology (MINT), and social sciences and humanities (SSH) research domains. System accuracy was measured by comparing the overlap between the top five system reviewer and human-selected reviewer choices.
The results showed a similarly high level of accuracy (as measured by overlap probability) for LS (81%) and MINT (85%), but substantially lower accuracy for SSH (67%), with the basic word embedding (bag of words TF-IDF). As expected, the results were better with the word embedding vectorisation methods, but only when they had been pre-trained on scientific data (eg, SPECTER2 > SciBERT > BERT), and were actually worse with standard BERT. In the best case, SPECTER2, the accuracy was LS (92%) and MINT (88%) – but, again, there was substantially lower accuracy for SSH (68%). More details on the evaluation results can be found in the corresponding research article (Okasa and Jorstad, 2024).

Organisational reflections on the use case

To be most effective, sufficiently many representative texts must be available for all reviewers. Typically, less bibliometric data is available for SSH reviewers, generally because of a higher focus on book publishing (data not systematically included in bibliometric databases) and partial publishing in languages other than English (which would confuse the semantic similarity aspects of the system).
It is a practical and political challenge to have a system that works substantially less well for one of the three research domains served by SNSF. It is hoped that further algorithmic improvements, such as fine-tuning and incorporating more representative bibliometric data, can partially improve these results.

5.3 'LA CAIXA' FOUNDATION: AI FOR PRE-REVIEW GRANT PROPOSAL SCORING

The 'La Caixa' Foundation (LCF) is one of the biggest charities in south Europe. It funds and promotes social, cultural, education and research programmes as part of its mission to build a better future for everyone. It is also a research funder, because it puts out highly competitive calls for biomedical research, split into different thematic areas. There is an overview talk on YouTube (Carbonell Cortés, 2024; Carbonell Cortés et al, 2024). The pre-review grant proposal system discussed here went live in 2023 and is currently (2024/25) in use.

Motivating factors

Each research proposal submitted to an LCF funding call must be assessed by external experts as part of the selection process. The situation is highly competitive, with more than 600 proposals but only 25-33 funded. While most of the applications were very strong, LCF noticed that some were not, and it seemed wasteful to have to recruit and use the time of three experts for submissions that were relatively easy to reject as unsuitable for funding. This led to the idea to test AI solutions as a form of triage to identify the weakest submissions for rejection without a full evaluation. The goal was not to score the proposals with AI, but to identify ones that were likely to get a low score from the human experts. Although reviewers are paid, the primary goal was to improve evaluation by reducing reviewer workload, rather than to save money on the reviewing process.

Data used

The data used is the full scientific text of research proposals in structured format, excluding personal data and any section related to the team research background. Proposals from previous years and their scores are used to train models, which are then applied to the new proposals from the current year to predict their probability of NOT being selected.

Implementation overview

LCF had already used AI as part of the reviewer characterisation process and to match proposals to potential reviewers, so the team had experience with successful use of AI. This was helpful in both the design and implementation of the system. Nevertheless, LCF worked with ITHINK as a technical partner for the system and SIRIS for an external review of the process.
The system was built in Python, mainly with Hugging Face, Torch and transformers as a natural language processing task. The domain is biomedicine, so the language is very

technical. Language processing is supported by BioBERT, BioELECTRICA, or BioBERT with adaptor blocks.
The system only went live after a deep initial analysis, a pilot test, consultations with stakeholders, tests of effectiveness, and board approval. The system operates on private servers to prevent leakage of the information in the proposals into the public domain or into the training data of external systems.
At time of writing, the system is live, but models are re-trained every year with the new data, to keep the process up to date within the research environment.
The system produces rejection probabilities from three different algorithms. Proposals flagged by the algorithms as having a low probability of being selected are sent to two human reviewers and only rejected if both reviewers are fairly sure they are not good enough to be funded.
This is an explicitly human-in-the-loop process, with no intention to make it fully automatic. Computer-based decisions are not legally allowed in any case. LCF thinks that a human-in-the-loop safeguard is essential to avoid overlooking excellent proposals that are out of the box and may trigger a low probability from the algorithms.

Evaluation/management strategies

The system self-evaluates in the sense that each rejected proposal is a 'win', in that no further human input is needed, having received a flag from the AI and two expert confirmations that it is unlikely to be fundable. During the pilot test, only one out of 86 proposals rejected by this process was funded, representing a high level of accuracy, but still an imperfect system. This was an unusual/uncommon proposal.

Organisational reflections on the use case

Ethical issues and informed consent were at the heart of the implementation. LCF held meetings with stakeholder groups – reviewers were part of the pilot and a meeting was held with a group of applicants to share the process and possible concerns. Applicants are informed of the process and given the possibility to opt out. LCF would like more insights into how the AI identifies proposals as likely to be weak and has internally approved a project to start studying this.

5.4 NOVO NORDISK FOUNDATION: USING AI TO IDENTIFY MATCH-FUNDED RESEARCH OUTPUTS

The Acknowledgements Project involves identifying research outputs that have been funded and the grant by which they were funded, even if this information is missing from the outputs.
The project was carried out by the Novo Nordisk Foundation (NNF), drawing inspiration from a similar initiative by the Wellcome Trust. The initial exploration phase was conducted by Raquel Roses. It was later revisited and refined for reproducibility by Emma Olsen during her internship at NNF, in collaboration with Sandra Schluttenhofer and Rasmus Lund Jensen.

Motivating factors

Each year, hundreds of academic papers published in scientific journals acknowledge the funding support provided by the NNF. Accurate tracking of funded publications is not only essential for impact management, but also contributes to a transparent understanding of how private foundations drive innovation and societal benefit. To enhance its assessments and shape future policies, the NNF seeks to maintain a comprehensive record of publications resulting from its funding, and the link to corresponding grants.
NNF guidelines require funded researchers to explicitly acknowledge the foundation's support and include their grant reference in all research outputs. Researchers are also required to report their research outputs directly to the foundation. However, compliance is frequently incomplete because of a lack of awareness among researchers or administrative oversights. Bibliometric database searches have revealed numerous publications acknowledging NNF funding that were not formally reported. This incomplete reporting undermines the accuracy of the foundation's evaluations and limits its ability to fully assess the societal impact of its grant-giving activities.
To address this, the foundation developed an AI-driven solution leveraging advanced text-mining techniques to match publications with their likely grants, improving the accuracy and comprehensiveness of its impact reporting. This strengthens the foundation's ability to evaluate its contributions and societal impact.

Data used

The project aimed to match publications to grants by using two primary data sources: publication data and grant application data.
Publication data was extracted from the bibliometric database Dimensions using its API. This dataset included publication

titles, abstracts, publication dates, author information, affiliation details and journal names. To ensure relevance, only publications explicitly acknowledging the NNF in their acknowledgments section, or listing NNF in the funders field in Dimensions, were included. Given the various spellings and abbreviations of NNF across publications, regular expressions (RegEx) were employed to identify all relevant mentions accurately. The grant application data included project titles, brief descriptions, grant references and programme areas submitted by applicants seeking NNF funding. This dataset was crucial for linking publications to their corresponding grants, as it provided the contextual information necessary for accurate matching.

Implementation overview

a) Identifying unreported publications

Once all publications mentioning the NNF were identified, the next step was to determine which of these had already been reported to the foundation. For the five-year period from 2019 to 2024, 3,427 publications acknowledging NNF were found that had not been submitted through the reporting system. Of the publications that had not been submitted, 1,876 (55%) explicitly included a grant reference, making them straightforward cases that could be mapped directly to the corresponding grant. However, the more complex cases involved publications that mentioned NNF but did not specify a grant reference. These cases required more sophisticated AI tools to establish the relationship between the publication and a specific grant.

b) Verifying funding acknowledgements

Not all publications mentioning the NNF in their acknowledgments section represent actual funded projects. Some references are related to disclosures of competing interests or other reasons unrelated to direct funding. For instance, a publication might state: "[The author] reports receiving unrelated grants from [...] and the Novo Nordisk Foundation." Such mentions, while referencing the foundation, do not indicate that the research was directly funded by NNF. Given the large volume of publications, manually reviewing acknowledgments to distinguish genuine funding acknowledgments from unrelated mentions was not feasible. To verify the funding acknowledgement, the open-source large language model (LLM) LLaMA was used. By employing refined prompt engineering and few-shot learning techniques, the LLM analysed acknowledgment sections and automatically flagged publications to determine whether the NNF mention indicated direct funding. While this automated process reduced manual effort significantly, ambiguous cases still required human verification to ensure accuracy.

c) Ranking based on similarity scores

To match unreported publications to their most likely grants, we employed natural language processing (NLP) techniques to compare the textual content of publications and grant descriptions. Various embeddings – TF-IDF, Word2Vec, and transformer-based models (eg, SciBERT) – were used to calculate cosine similarity scores, providing a measure of how closely a publication matched a particular grant. While transformer-based models offer advanced semantic understanding, they struggled in this context because of input length limitations and high computational costs. Conversely, the simpler, context-unaware TF-IDF demonstrated strong performance. Its effectiveness likely stemmed from its ability to process the full text and emphasise rare, highly informative terms, which are often key to distinguishing between potential grant matches. Importantly, we ensured that comparisons were only made between grants and publications that were date-compatible, thereby enhancing

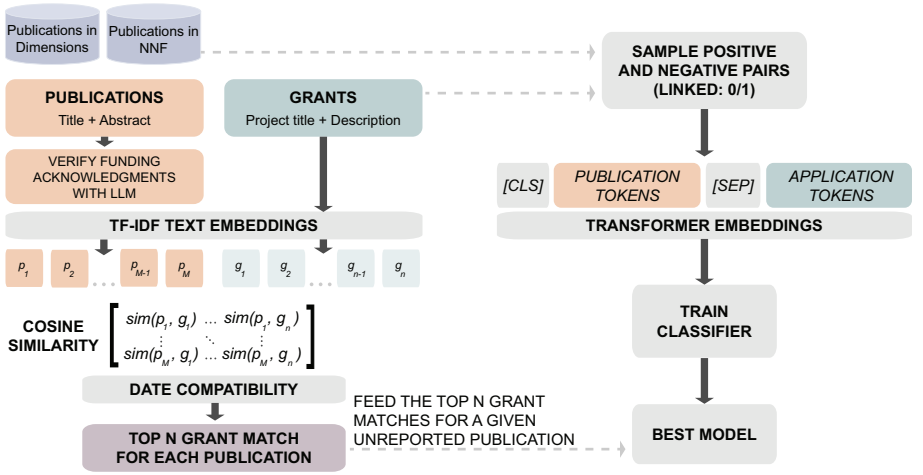


FIGURE 5.1 IMPLEMENTATION OVERVIEW OF NOVO NORDISK FOUNDATION SYSTEM FOR MATCHING PUBLICATIONS TO GRANTS

the relevance of potential matches. The similarity scores between an unreported publication and all possible matching grants were ranked, with the highest rank indicating the publication-grant pair with the greatest similarity. To evaluate this approach, we tested it on reported publications where the correct grant reference (ground truth) was already known. The method identified the correct grant reference as the top-ranked match in 35.67% of cases and included the correct grant within the top 10 ranked matches in 65.54% of cases. These results provided a benchmark for the effectiveness of text-similarity-based methods and set the stage for further refinement.

d) Using a classifier for publication-to-grant matching

The classifier was designed to determine the probability that pairs of texts – specifically, a publication (represented by its title and abstract) and a grant application (represented by its project title and brief description) – are semantically linked. Its task was to classify each pair as either a 'match' or 'no match', while providing an uncertainty score for this classification. To train the classifier effectively, a balanced dataset of positive and negative examples was created. Positive pairs consisted of publications matched correctly to their grants based on known associations in the NNF's database. Negative pairs were synthetically generated. These false pairs were crafted to closely resemble positive examples by sharing features such as being within the same programme area or having similar author or applicant names, but were definitively unrelated. This careful design ensured the classifier was trained on challenging cases, enhancing its ability to distinguish subtle differences between true matches and plausible but incorrect ones.

The classifier was based on a pre-trained LLM fine-tuned for a binary classification task. By leveraging semantic embeddings, it analysed the textual content of publications and grant applications to predict whether a pair was linked based on semantic equivalence. In addition to analysing text, the classifier incorporated an additional feature: the names of the authors listed in the publication and the applicants or co-applicants associated with the grants. By comparing these names, the model gained another layer of information to assess potential links, particularly in cases where textual similarity alone was insufficient to make a clear determination. The output of the classifier included both a classification ('match' or 'no match') and a confidence score, providing an interpretable measure of uncertainty for each prediction. The classifier was evaluated using 5-fold stratified cross-validation, achieving an accuracy of 85.76%, precision of 94.12% and recall of 76.31%.

e) Final approach: combining text-similarity ranking and the classifier

While the classifier produced more accurate results in matching publications to grants, it was computationally expensive and slow. This limitation stemmed from the need to compute a classification for each possible pair of publications and grants, a process that becomes increasingly resource-intensive as the number of pairs grows.

To address this challenge, we adopted a hybrid approach that combines the efficiency of text-similarity ranking with the precision of the classifier. First, for a specific unreported publication, we used the text-similarity ranking method to identify the 100 most similar grants. This step, based on cosine similarity between semantic embeddings, is computationally fast and provides a narrowed down list of candidate grants. With ranking alone, the correct match is expected to be among the top 100 candidates 86% of the time, making this threshold a deliberate trade-off between computational efficiency and matching accuracy. Once the 100 most similar grants were identified, the classifier was applied to these pairs. By limiting the classifier's use to a smaller subset of highly probable matches, we significantly improved performance while maintaining the classifier's accuracy. This hybrid strategy allowed us to balance computational efficiency with high-quality results, ensuring a practical yet effective solution for matching unreported publications to their likely grants.

Evaluation/management strategies

The solution developed through this project is intended to be used within the organisation in both the short and long terms.

In the short term, the matching of unreported publications to grants will be used to contact grantees directly. This outreach will serve four purposes:

(1) To confirm whether the identified matches are correct.

(2) To request that grantees add the publications to their reports if they were inadvertently omitted.

(3) To understand why these publications were not initially reported.

(4) To get feedback from the grantees on whether we have identified the correct grant and to investigate the cases where our methodology didn't work. This feedback will be the starting point for further improvements to the model and will be an iterative process, done after each reporting period.

This approach not only improves the

completeness of reporting, but also signals to grantees that we actively use and value the data they provide. Furthermore, it offers valuable insights into reporting behaviour, enabling us to identify and address potential barriers. In the long term, the solution will be integrated into our reporting system. When grantees are asked to submit their reports, they will ideally be presented with a pre-generated list of publications in which they have acknowledged the NNF. This streamlined process will allow them to confirm and add relevant publications with minimal effort, reducing the reporting burden significantly and saving time for our grantees.

Organisational reflections on the use case

The Acknowledgements Project represents a significant step forward in improving the NNF's ability to track and assess the impact of its funding. By leveraging advanced AI techniques such as text-similarity ranking and classifiers, the project addressed the challenges of incomplete reporting and unreported publications. The hybrid approach successfully balanced computational efficiency with precision, ensuring a scalable and effective solution for matching publications to grants.

5.5 RESEARCH COUNCIL OF NORWAY: ASSESSING SOCIETAL IMPACTS OF RESEARCH

This section summarises the results of a study conducted in 2021 by Technopolis for Research Council Norway (RCN) – Study to Establish a Methodology to Assess the Societal Impact of Research and Research-based Innovation (Technopolis, 2024) – which RCN is currently following up. This is also discussed in greater detail in a book chapter (Holm et al, 2024). The heart of this case study is the need to have a pipeline and classification process to reliably identify impacts that relate to specific societal impact goals.

Motivating factors

RCN is, among other things, the major research funder in Norway, and supports basic and applied research and innovation. As part of this, it would like to assess whether the research that it supports has a beneficial impact on society. The task of tracking the influence of research on society is notoriously difficult most of the time, as lag times can be long, the influence of research is often not directly acknowledged, and multiple studies may contribute in complementary ways to any measurable societal benefit. Qualitative approaches are often used to reflect the complex ways in which influence may occur (Miettinen et al, 2015). RCN was motivated to start this project by a recognition that capturing societal-impact evidence of its supported research would be very useful to help guide funding policy, even though it was recognised as being a complex task that could not deliver comprehensive results.

Data used

Preliminary work by Technopolis identified three main impact dimensions for RCN (economy, society, environment) and nine types or pathways for these three impacts. It ran a pilot study to trace the effects of RCN's funding of three types. The basis of the methodology was identifying (a) the funding (b) the outputs produced by the funding and (c) early evidence of the uptake of the outputs and influence.

The data sources used were:

(a) The RCN internal project database. Text mining was used on English-language abstracts (when available) to match projects with the impact types investigated.

(b) The Dimensions.ai bibliometric database through its API.

(c) The Lens.org (patents), Dimensions.ai (patents), Altmetric.com (blog and social media citations, Wikipedia citations) and Overton (policy documents) databases. These were used for their complementary coverage of different document types. For example, Overton specialises in policy documents and was used to identify citations to RCN-funded outputs from policy documents.

Implementation overview

Team overview: RCN did not have the in-house data science expertise to address this challenge, so commissioned Technopolis for it.

Technology and processes: While (a) and (b) should exist within funder databases, the latter needs a technical solution to identify publications (in the Dimensions.ai bibliometric database) that are not reported to the funder but are known to be connected to the funder (eg, mentioning them by name) but do not contain grant IDs. AI solutions involving topic modelling (of publication text) and keyword matching were suggested for this task.

The main technical challenge is (c) and the proposal to address it was documentary: identifying relevant citations from patents, the mass media and blogs, Wikipedia, and policy documents. Here AI is needed to classify the citing documents by topic to assess whether they are relevant to the types of impacts considered for the study (eg, protection of ecosystems). The TextRazor API was used instead of locally implemented topic modelling

for text classification. This has the advantages of being multilingual and a service, so it does not need to be implemented locally.

Evaluation/management strategies

The Technopolis project consisted of three pilot studies and was primarily evaluated for technical feasibility, rather than formally evaluated for accuracy and coverage. The problem here is that the project is attempting to capture something that is otherwise not captured or measurable, and therefore the results cannot be comprehensively evaluated. They could be assessed for plausibility and for the presence of obvious errors and omissions, but partial coverage and some errors are inevitable in this process. Nevertheless, the system as a whole was clearly successful at identifying large numbers of documents that were potentially reflective of societal impacts of the types sought. The system was considered to be able to find more evidence of impact than even an expert would know (Holm et al, 2024).

Organisational reflections on the use case

RCN has broadly accepted the Technopolis recommendations and is moving towards a production system. Once this system is in place, it will become possible to assess the extent to which the results can inform RCN policy and be used for other purposes. RCN is working on implementing this system and trying out different classification approaches, including local topic modelling to assign categories that are more closely aligned with funder needs. RCN considers that an advantage of the approach developed is its flexibility, in the sense of possibly being able to adapt a system designed to identify one type of impact into a system designed to identify another type (Holm et al, 2024).

> "The system was considered to be able to find more evidence of impact than even an expert would know"

5.6 UK BLOCK GRANTS FUNDERS: AI FOR ACADEMIC JOURNAL ARTICLE QUALITY ASSESSMENT

This section reports a study by the University of Wolverhampton, UK, funded by Research England, the Scottish Funding Council, the Higher Education Funding Council for Wales, and the Department for the Economy, Northern Ireland, as part of the Future Research Assessment Programme. The goal was to assess whether traditional machine learning methods could support or partly replace expert assessments for the post-publication research evaluation of academic journal articles, which is used to allocate the majority of block grant research funding in the UK. The results were published in an overall report and a key journal article (Thelwall et al, 2022, 2023).

Motivating factors

The UK funds higher education institutions (HEIs) and various other bodies for research using block grants that are allocated on the basis of performance over the past 5-7 years. This performance-based funding is driven by the results of the Research Excellence Framework (REF), which last ran in 2021 and is due to run again in 2029. Details vary between iterations, but for each REF, the approximately 157 eligible institutions submit a portfolio of evidence for the quality of their research. This fits within three components: research environment (narrative report); impact case studies (a set of five-page structured reports); and outputs (a list of 185,594 journal articles, books and other products for REF2021). Submissions are split into 34 units of assessment (UoAs), which are broad fields of research (eg, physics, area studies) or clusters of related small areas. For REF2021, each output and other submission aspect was scored on the following four-point scale by a team of more than 1,000 experts over a year (**2021.ref.ac.uk**).

4*: Quality that is world-leading in terms of originality, significance and rigour.

3*: Quality that is internationally excellent in terms of originality, significance and rigour, but which falls short of the highest standards of excellence.

2*: Quality that is recognised internationally in terms of originality,

significance and rigour.
1*: Quality that is recognised nationally in terms of originality, significance and rigour.
The results were then put into a formula to calculate the block grant research funding that each submitting institution would receive annually for the next seven years.
A key problem with the REF is that scoring 185,594 research outputs is extremely time-consuming (and expensive) because the experts required must be re-assigned from other research-related activities. So, any use of AI to support or partially replace the human labour involved could, potentially, be highly beneficial. This is why the main four block research funders combined to commission a study to assess whether there was any scope for AI support.

Data used

Although other types of outputs are in scope for the REF, the study focused on journal articles because these are more widespread and standardised as UK research outputs, both of which are advantages for AI approaches. The REF team supplied the Wolverhampton researchers with a list of journal article details (eg, title, authors, DOI, journal), submission details (institution, UoA), and the provisional score given by two or more of the 1,000-plus experts after the internal REF score norm-referencing exercises. HEIs are only told their score profiles and not the scores for individual outputs. These scores are private to the REF process, so the Wolverhampton team had to sign confidentiality agreements and destroy the data after two months.
The data supplied by the REF team was paired with metadata from Scopus, including citation counts, article abstracts and Scopus-allocated (journal-based) field classifications. This was stored in a flat file (tab delimited) and split into training/testing subsets by the (Python) program used for the analysis.

Implementation overview

The University of Wolverhampton team consisted of scientometricians – with experience and expertise in the use of quantitative methods to support research assessment, such as various forms of citation analysis – together with qualitative researchers and academic experts on machine learning. The team had been selected by an open competition for a team and analysis strategy.
The technology chosen for this task was traditional machine learning, implemented in Python, primarily the Python package Scikit-learn. At the time of the bid (January 2021), large language models had not emerged as a serious challenger to traditional machine learning, so were not considered. Machine learning was selected instead of other forms of AI because the tabular format of the data is mostly well suited to it.
The core experiments compared the accuracy of 32 different machine learning algorithms for the task of predicting the REF score of a journal article from its metadata (number of authors, field normalised citation count, field normalised journal citation rate, team size, team citation and publishing achievements, title, and abstract words and short phrases). The experiments were not conducted as standard in the field (10-fold cross-validation), but in formats that could match potential applications, such as machine training on 50% of the outputs to test whether computer predictions could replace half of the human scoring labour.
Three different overall strategies were also tested: standard; active learning (stratifying the data, with the computer selecting difficult outputs for the experts to score); and prediction by probability (only making predictions when the computer had a high certainty of them being correct).
The results from the best strategies were shared with a series of focus groups of REF assessors to get their feedback on whether the most plausible strategies were acceptable. The most plausible strategy was to accept the automated predictions when the computer had a high degree of certainty, which would only be practical for a small percentage of the outputs. This would result in a small time saving for the assessors overall, but this was not considered enough to outweigh the perverse systemic incentive caused by including journal citation rates as an input, given the UK's official position of strongly opposing journal impact factor data in decision-making. The final recommendation from the report was not to use the technology for REF2029, but to run additional parallel pilot studies to work towards improved accuracy.

Evaluation/management strategies

Because of the negative outcome, this project was not implemented for REF2029, but a future evaluation is being planned for REF2029 data.

Organisational reflections on the use case

Given the substantial expense of the REF assessment process, there has been ongoing political pressure for efficiency in the form of supporting or replacing it by automated methods – formerly citation analysis and now AI. Thus it was politically necessary to evaluate this approach. The results

were promising from the perspective of generating a technically practical way to save a small amount of time in the assessment process, even though, for wider reasons, it was not desirable to implement them.
A key and unexpected organisational issue was that assessors in the focus groups were wary of the computer predictions, even if they were highly accurate. This is because they would still cause small fluctuations in the league tables of universities within UoAs, which are highly valued by them. This issue could be overcome in the future with better knowledge of the variability of human expert judgements. This would allow the computer errors to be benchmarked against computer variability.

5.7 SUMMARY AND REFLECTIONS ON AI CASE STUDIES

These five case studies illustrate a range of goals for AI, strategies for implementing it and stages of development, from pilot testing to ongoing use.

One common factor for the tasks is that they are periodic – neither one-off nor ongoing, but needing to be carried out repeatedly. Of these, the most frequently needed task might be finding reviewers. However, each case study illustrates different challenges and opportunities for monitoring and intervention, as funders go about using, improving and maintaining AI/ML systems as part of their regular operations.

Learning from the case studies, the following general points should be considered when funders are considering AI/ML solutions:

- With the promise shown by ChatGPT, there is increasing expectation by senior managers and end users that AI solutions will be developed to deal with complex tasks.
- Identification of the scope of possible AI solutions is a complex task that may be an entire project in itself.
- Organisations lacking the internal skills or time to develop AI solutions can commission other organisations or researchers to develop solutions. This can have the advantage of enhanced expertise for the task, although perhaps at the expense of reduced task awareness by the developers. The recommended solution might then be implemented in-house, if the expertise exists.
- Unless generic solutions are bought in, ongoing maintenance will be needed on AI systems to keep them functioning efficiently and, perhaps, adapt them for new challenges or upgrade them when new technologies appear. This (including any personnel implications) should be built into the project lifecycle.
- When it is possible to evaluate AI recommendations, this gives useful information, but alternative evaluation strategies should be sought when it is not possible.
- There are many commercial tools and data sources (eg, bibliometric databases) that could form part of integrated solutions that pull from multiple data sources to feed the AI component of a solution.
- Effective AI implementations are designed with a clear understanding of the contexts in which they will be deployed, the ethical implications of their use, and strategies for assessing equitable outcomes. This is specially the case when AI implementations can impact stakeholders such as funding applicants.
- Important potential AI applications that must have a high degree of accuracy might be implemented through a hybrid system where the AI feeds suggestions for consideration by humans, effectively as a form of pre-filtering.

"Identification of the scope of possible AI solutions is a complex task that maybe an entire project in itself"

Responsible AI futures

This handbook serves as a jumping-off point for funders to explore next steps and future directions with AI and machine learning. What this process looks like is not a settled question, and AI exploration and implementation will be different from funder to funder and AI application to AI application.

This concluding portion of our handbook highlights key learnings, recommendations and directions to further guide funders in their next steps with AI/ML, and to help construct clearer futures of responsible AI use in research funding and assessment.

We draw on the collaborative discussions in the GRAIL project, which served as the genesis for this handbook. Here, we present:

1. **Key observations and lessons learned** from the experiences of an international consortium of funders at different stages in exploring and applying AI
2. **Core recommendations** for funders at any stage of their AI journeys, to help achieve more effective, ethical and equitable AI applications
3. **Directions for experimentation** for funders interested in structured, systematic assessment of AI impacts in the work of funding and assessment.

6.1 SHARED LEARNING: KEY AI/ML TAKEAWAYS FOR FUNDERS

The collaborative discussions in the GRAIL project surfaced a wide range of important considerations, challenges and strategies for research funders working with AI/ML systems in practice. Here, we highlight key observations that can inform how funders work with AI/ML in the future, and describe the implications of these findings for funders.

6.1.1 AI IMPACTS FOR FUNDERS AND RESEARCH SYSTEMS

Funders have a unique role in shaping the use of AI in research systems. Funders must balance the goals and expectations of governments or private entities that supply their funding, the research communities that rely on their support, the individual applicants with whom they interact, and wider publics that the work of research funding aims to ultimately benefit.

Each of these groups has different expectations and understandings of the role of AI in research systems, and each looks to funders as a central source and convener of expertise to help inform what AI-informed research systems should look like. Funders must therefore bridge the gap between developing policy on AI use and supporting AI innovation in research.

Many funders also include a broad range of disciplines in their remit, up to and including all of research, so must approach questions of AI use in the context of how it will affect all of these diverse areas of research and practice.

To respond to these pressures, and the lack of peer organisations that many funders experience in national contexts, funders should draw on their international networks and established models for collaboration. Open research and responsible research efforts – such as the Coalition for Advancing Research Assessment (CoARA), the San Francisco Declaration on Research Assessment (DORA), or the Barcelona Declaration on Open Research Information – provide examples of how funders can build international networks and work together to tackle complex challenges in global research, and can serve as exemplars for building collaborations on AI.

Most AI problems are not new problems. The work of funders is situated in complex, ever-evolving networks of science policy, research cultures, organisational values and community expectations. The use of AI technologies interacts with all of these, but does not replace them – or, in most cases, even transform them.

The key challenge for funders exploring or applying AI is not the development of new policy or practice, but integration of AI with existing policies or practices. Policies on data use will often apply as is to use of data with AI; AI use leverages existing IT infrastructures; and established organisational values are the best guide to building successful cultures around AI adoption.

AI use is driven by a wide range of anticipated benefits. Efficiency improvements are the go-to argument for AI adoption in many cases, particularly in the public sector. However, funders explore and apply AI for many reasons, including gaining deeper insight into existing (human) processes, expanding their options and making better use of data, decreasing reviewer workload, and even helping applicants strengthen their applications.

These anticipated benefits come with the challenges of managing expectations, within funders and among external stakeholders. The hype surrounding AI developments, particularly with generative AI, often leads to inflated expectations of solutions being easy, fast and transformative. Teams in funding organisations implementing AI in practice must carefully navigate the collision between these expectations and the reality of AI application, which is often more measured and complex in benefit and impact.

However, AI use can also bring unexpected benefits. Funders using AI have been able to learn about trends that human review would not be able to identify, such as tracing large-scale developments in maritime industries (Holm et al, 2024). AI analysis can also

uncover aspects of existing processes to improve, such as highlighting continued over-reliance on journal metrics in assessment.

AI use presents specific risks and limitations for research systems. Broad risks of AI use, such as reinforcing bias in data or 'hallucinating' inaccurate information from generative AI, are well known. AI use also poses several specific risks in the context of research funding and assessment, spanning technical, operational, strategic and political considerations.

For example, misuse of AI may expose confidential or proprietary data to third parties, or inappropriately share intellectual property. By learning from past successes, machine learning may penalise original research ideas that diverge from what has been seen before. Similarly, if many applicants use the same AI systems, the originality of their ideas may be watered down. Existing problems of reproducibility are magnified by the use of probabilistic AI systems that are difficult to explain.

More broadly, funders and researchers must be cognisant of the limitations AI systems pose in the context of complex funding processes. For example, peer review fulfils a social and cultural function of reinforcing, and at times reshaping, what is accepted research in evolving environments. This is not a function that AI systems can reproduce, no matter how successful they may be at helping to provide specific feedback on individual applications. Similarly, the probabilistic nature of most AI poses fundamental questions for funders about the decisions made in funding processes, though existing practices, such as partial randomisation, provide a basis for helping to think through these implications.

6.1.2 PEOPLE AND PROCESS: WORKING WITH AI

Working with AI involves navigating many types of diverse thinking. Different individuals, within and outside of funding organisations, have different understandings of what AI is and how it may be used, and these are continuously evolving. Individuals also have different perceptions of AI benefits and risks, including different conceptualisations of societal versus individual harms from using AI (eg, erroneously rejecting an application vs enforcing normativity in research). Different individuals and organisations also have different opinions on AI use, from early adopters who are eager to use AI to the fullest to reluctant users seeking flaws in AI applications.

Different funders have developed different practices around AI, as demonstrated in the case studies in this handbook. AI/ML technologies have been put to use for a great diversity of purposes: reviewer matching, topic modelling, mining funded outputs, screening proposals, summarising peer review, and even providing FAQ interfaces for applicants. Each funder exploring or applying AI may learn from these diverse experiences, but must shape their own practice with AI that is based on their own individual context.

AI use is an interdisciplinary process that requires working across silos. Funding organisations employ diverse teams with deep expertise in distinct skill sets: scientific officers are experts in particular research areas; funding administrators are operational experts; senior advisers are experts in strategic thinking; and so on. In addition to their organisational role, funder staff are often experts in widely varying disciplines and research areas, depending on the funder's remit and their own past experience.

As a result, exploring and applying AI in the research funder context requires effective bridging between professional and disciplinary areas that are often siloed, and building collaboration across highly distinct skill sets. This handbook has laid out several strategies and examples for tackling this challenge; the fundamental learning is that AI is a collaborative team effort across roles, and cannot be handled effectively by an individual or single department.

As well as building collaboration, interdisciplinarity brings practical challenges. The ground truth of what is 'correct' for an AI system to predict – an essential aspect of training and evaluating machine learning models – can be difficult to resolve when an application involves different people with differing conceptions of what the system should do. These issues can be exacerbated by disparities in data quality and availability between disciplines, such as the reduced data resources on social sciences and humanities research (Sivertsen, 2022).

Funders use diverse methods to engage internal stakeholders around AI. Funders who have developed or explored AI/ML applications have made use of a wide variety of engagement models for working with internal stakeholders.

Purpose-based workshops are a common method for stakeholder engagement around specific AI/ML applications, including at pre-development stage (to assess potential impacts, benefits and risks early),

with collaborative user engagement workshops during development, and impact analysis workshops as part of deployment and assessment.

More ongoing structures include organisational roadshows, in which data/AI teams attend regular meetings to present and listen to potential user needs; and departmental consultation models, through which data/AI teams have a standing offer for consultation to identify AI applications.

Stakeholder engagement within the funding organisation must be approached as an iterative and ongoing process, particularly given the constantly changing landscape of AI/ML technologies. Engagement models should reflect iterative relationship building, with periodic reassessment and re-balancing of operational and strategic priorities. There is no single right answer for AI implementations in funders' complex contexts, and continuous engagement with stakeholder networks is an essential tool for finding the next best approach.

Initial AI expertise can be sourced externally, but long-term expertise is in-house. External expertise (ie, consultancy or contracted services) is invaluable for funders in kick-starting exploration or application of AI, and enabling faster access to more data and expertise. However, funders' needs for accountability, transparency and customisability in AI applications pose barriers to relying on external expertise alone, which may also pose risks to replicability and a funder's reputation.

Any external solution, if adopted as part of a funder's regular work, requires internal expertise to integrate it, assess it and develop assurance processes, as well as deliver the full value of the AI system in context. Relying on internal expertise presents its own challenges: the interdisciplinarity of AI means this may require access to a large number of people with very limited bandwidth, and significant effort to build shared understanding across them.

A key strategy for building internal expertise and buy-in on AI use is to identify local champions around the organisation who can help to engage directly with colleagues and adapt AI systems to their contexts. Using strategies such as this, and approaching AI expertise as an iterative process involving both external and internal stakeholders, will help funders develop the long-term AI expertise they need.

6.1.3 PUTTING AI INTO PRACTICE

Measuring AI performance in the funder context is multidimensional. The efficacy of AI/ML systems is often measured in terms of task performance: demonstrated ability to produce the outputs that are expected for specific inputs. However, the use of AI in highly complex funding and assessment processes requires more sophisticated evaluation that addresses multiple aspects of how a system performs.

Reliability of AI systems is a key concern for funders, who must be able to provide stable and reliable operations with clear accountability. The efficacy of AI systems may vary from situation to situation – eg, a topic-modelling system may work well for disciplinary research, but fail when presented with an interdisciplinary funding proposal. How to determine when a system is 'good enough' to use is an open question, which must be addressed by each funder in their individual contexts.

The external validity of AI systems, in terms of how well they agree with expert decisions and their impact on research systems, is a key consideration for funders. This can be assessed most directly through review by scientific officers and other organisational users of AI, though this carries with it significant expense in time and money. Engagement with users is essential for managing trade-offs in AI applications, such as between false positives and false negatives, transparency and accuracy, or deeper understanding vs reducing bias.

Funders have a key role in setting AI guidance in research systems. In line with their role as guardians and guides of national research systems, an emerging role for funders is in establishing guidance on the use of AI in research systems. This includes AI guidance for applications within funding organisations and for wider applications by researchers and research institutions. Funders are the best-positioned actors to set AI guidance that bridges the goals of supporting researchers, responding to AI policy directives and laying a strong foundation for future regulation on AI in research.

Guidance on AI from funders should, first and foremost, be adaptable. Just as AI use varies widely within funders, so there is enormous variation in AI application throughout wider research systems. Emerging guidance from funders therefore focuses first on general principles and processes, which enable adaptation to local contexts.

The role of guidance from funders is

twofold: to proactively shape behaviour and articulate a wider cultural stance on AI in research; and to provide the framework for reaction and response to misuse of AI in research. While the specifics of AI guidance documents differ from funder to funder, they generally address a wide scope of AI (not just generative AI) and reinforce existing expectations that researchers are ultimately responsible for what they produce, whether or not AI use was involved.

Developing best practice on AI in research funding needs a community-of-practice approach. Funders often work in relative isolation, because of the lack of peers in national systems and the specificity of the contexts in which they operate. With emerging technologies such as AI, this isolation can lead to different funders repeating the same experiments and relearning the same lessons. The GRAIL project has illustrated the value of creating a community of practice in which funders can share experiences, recommendations and lessons learned.

A community of practice also provides a space for funders to work together to create shared understanding of responsible AI principles in the context of research funding and assessment. Fundamental principles of responsible AI, such as those outlined in Section 1.3, are a jumping-off point for investigating questions that are more specific to research, such as authorship and intellectual property, academic freedom, originality, and research integrity. Working together to understand how AI interacts with these and other long-standing questions for research is essential to the functioning of AI-informed research systems.

6.2 RECOMMENDATIONS FOR SHARED PRACTICE

In response to these learnings, and drawing on discussions and progress made in the GRAIL project, we make the following recommendations for funders exploring and applying AI. Our recommendations address three essential aspects of building strong, shared practice on AI in research funding and assessment: the ***people*** involved; the ***practice*** of using AI; and the ***principles*** that guide its use.

6.2.1 PEOPLE

1. Evolution, not revolution: when exploring AI use or developing new AI/ML applications, funders should leverage their existing organisational values, established policies and the wider cultures in which they participate, such as open research. These are powerful resources to guide new AI practices in a way that is accessible and easily understood by all stakeholders. They also provide valuable frameworks for reflecting the fact that most AI use is a matter of improving existing processes, rather than new transformations.

2. Hub and spokes engagement: responsibility for decisions about AI use is diffused across different teams within funding organisations, but the most effective action on AI is led by a centralised team. AI teams should prioritise active and repeated engagement with stakeholders throughout the organisation, in a 'hub and spokes' model in which they lead by maintaining strong connections to all those working with or affected by AI use.

3. Use/buy/build assessment for AI: new AI technologies and established AI expertise may be sourced internally and externally, with different benefits and drawbacks for each. When considering a new AI application, funders' first step should be to assess whether the best step is to: 1) use an existing technology, whether commercial or open source; 2) buy an external technology – eg, through consultancy; or 3) build an in-house, custom technology.

6.2.2 PRACTICE

4. No 'one size fits all': to function effectively, be implemented ethically and produce equitable results, AI applications must always be adapted to the unique contexts of individual funders. Funders should experiment with multiple methodologies for any AI application, and consider the experiences of other funders (or any pre-built application) as a starting point for their own AI journey, rather than a complete solution.

5. Build a transparency toolkit: there are valuable existing resources for AI transparency, including datasheets, model cards, and some initial templates for impact assessment. As with AI implementations, however, these are only a starting point; each funder must adapt and expand these for their own practices in data preparation, model development and evaluation, and in monitoring and assessment of AI systems.

6. Embrace 'human-in-the-loop': human oversight of decision-making is a legal requirement for some funders, and a wise choice for all. Human experts and AI systems have complementary strengths that human-in-the-loop models can help to leverage. The essential criterion for any effective use of AI by funders is that it is trusted by staff, leadership and applicants, and human-in-the-loop approaches are essential to achieving this.

7. Dedicated oversight structures: while many aspects of AI can and should be covered by existing policies and practices, funders will benefit from having a dedicated structure for oversight of AI applications throughout the organisation. This may come in different forms, such as an internal oversight board or an organisation-wide AI working group. Funders must cultivate appropriate internal expertise to provide oversight and assurance for all AI applications they use.

6.2.3 PRINCIPLES

8. Responsible AI for funders: responsible AI frameworks, such as those discussed in Part 1, are valuable resources, but funders must build on these to develop internal understanding of what responsible AI looks like for each organisation. Funders should work with each other and with internal stakeholders to develop shared understanding of AI accountability, transparency, explainability, agency and oversight, and legitimacy of AI use, as well as how AI affects key concerns for research systems, such as academic freedom and research integrity.

9. Build for the tools of today, not the myths of tomorrow: AI is often presented with far-flung promises of future transformations that are rarely matched by reality. Rather than responding to visions of potential AI futures, funders should focus on tools and technologies currently at hand. These are the technologies that will actively shape research systems and will be most relevant for funders and the researchers with whom they work. Focusing on current technologies also allows funders to build responsible policies and practices from experimentation and concrete experience with specific use cases.

10. Take a problem-based approach to AI: funders should approach AI/ML as a toolbox with which to respond to specific problems, rather than seeking out opportunities to make use of AI/ML technologies. A problem-led approach will ensure that AI/ML use is fit for purpose and shaped by the needs of the organisation, and avoid the risk of unnecessary and counterproductive change driven by the hype of AI transformation.

6.3 DIRECTIONS FOR FUNDER EXPERIMENTATION WITH AI

Finally, we briefly highlight future directions for funders to develop systematic experiments using the methodologies highlighted in our previous publication (Bendiscioli et al, 2022). Use of AI/ML should be evaluated in a similar way to any other process intervention, and both the work of funders and the development of AI/ML technologies will benefit from more systematic experimentation with AI/ML application.

Here, we present seven directions for potential experimentation by funders, beginning with AI applications closest to current use and going towards more speculative directions for AI/ML use. Many of the directions highlighted here reflect those explored in specific case studies in Part 5, which can serve as a template for funders seeking to develop structured experimental evaluations.

6.3.1 AI IN REVIEWER MATCHING

Helping funders identify the best peer reviewers and panel members is currently the most common application of AI/ML in research funding (Rushforth et al, 2025). However, the contributions of AI/ML technologies to this process, and their impacts for the organisation, have not yet been explored systematically.

There is significant scope for exploration and experimentation in the types of AI/ML methods used: for example, different types of data used as input for AI/ML models (reviewer scientific record, application materials, applicant team information, etc); different modelling techniques; and different kinds of outputs and purposes for AI/ML application, such as a ranking of possible reviewers or suggestions of new reviewers to add to an established reviewer pool.

For focused experimentation to compare and improve AI/ML approaches, funders could begin with curated reference sets of applications and 'ideal' reviewer assignments, as selected by teams of scientific officers within the organisation. AI/ML systems could then be evaluated based on their ability to produce these ideal matchings, as an initial step in technical comparison.

To evaluate impact on the full review process, funders could develop a foundational A/B testing experimental framework to perform multiple experiments with different AI-enabled reviewer-matching strategies. For example, funders could set up parallel experimental/control tracks within a single call, to which applications are randomly assigned, and measure the impact of an AI-enabled matching system based on measures such as time to completing reviewer recruitment, number of declined invitations, or scientific officer feedback on the quality of reviews received.

6.3.2 AI IN PEER REVIEWING

The use of AI/ML to help produce and process peer-review reports has been explored in the experimental literature (Price and Flasch, 2017; Checco et al, 2021), but not yet explored systematically within funding organisations. Generative AI technologies have particular potential to help produce and summarise peer reviews, but also present significant risks of producing inaccurate and misleading content. Funders will benefit from developing structured evaluations of generative AI technologies in peer-review processes.

Automated evaluation of AI-generated reviews is a complex process without clear measurement strategies (Yuan et al, 2022). Funders should therefore approach this in one of two ways: expert evaluation of quality and informativeness of generated reviews, relying on scientific officers or expert peer reviews; or comparative evaluation of peer-review processes, comparing manual peer reviews to those where AI is used as part of the reviewing process.

Both of these designs would enable experimentation with different AI-based generation or synthesis strategies, as well as different levels and types of AI use in the peer-reviewing process. However, the reliance on expert evaluation will make these experiments time-consuming and expensive to run, so funders should choose carefully what aspects they wish to prioritise in their experimentation.

6.3.3 AI FOR PRIORITISING FUNDING APPLICATIONS

Case study 5.3 illustrates one recent example of using AI/ML systems in the process of prioritising applications for consideration in the funding process.

As an area in which decisions have significant material consequences (ie, affecting whether an application is awarded funding or rejected), a human-in-the-loop approach is essential. However, the ranking process is a clearly defined problem for AI, and one for which funders have developed clear criteria to guide the process. This means there is real potential for funders to benefit from careful experimentation with AI/ML to support the prioritisation process.

Automated evaluation for piloting different AI/ML systems is quite straightforward for funders to perform with their own historical data. Records of which applications were awarded and rejected, and what scores were assigned by panels, provide data that can function directly for training and evaluating AI/ML systems for ranking applications. This provides a valuable and easily implemented platform for testing and innovating with new AI/ML methodologies, with no impact on new funding decisions.

Experimentation in live funding calls, however, requires a carefully structured approach with a well-developed plan in place for bringing the research community on board from the beginning of the process. A randomised controlled trial is a natural approach, with funding

applications randomly assigned to a pool for manual or semi-automated prioritisation (with human oversight over all decisions), and evaluation based on final awarding decisions and time saved in the process. As a starting point, we refer funders who may consider this direction to the design implemented by 'La Caixa' Foundation in its experimentation (Cortés et al, 2024).

6.3.4 AI IN RESEARCH ASSESSMENT EXERCISES

Case study 5.6 describes recent research evaluating the use of machine learning models to assist in automated scoring of research outputs in a national assessment exercise in the UK. A similar study has explored the use of machine learning in assessing research impact (Williams et al, 2023) and early analysis has examined the use of large language models for research assessment (Thelwall, 2024).

This is an area in which the use of AI/ML techniques shows particular promise, but also significant challenges in how to integrate the use of AI/ML effectively without compromising the quality of assessment or the trust that governments and research communities place in assessment processes.

Funders should build on these initial examples to explore further targeted experimental questions in where and how AI/ML might best be leveraged within assessment processes. As there is already active academic research evaluating specific AI/ML approaches for assessment, funders will benefit from focusing particularly on evaluating which elements of research assessment processes are effective points for AI/ML intervention, and how to strengthen the trust built up with key stakeholders in assessment processes when AI/ML systems are introduced.

6.3.5 AI FOR APPLICANT SELF-ASSESSMENT

As well as their demonstrated value in helping to select submitted funding applications, AI/ML technologies have significant potential for applicants as tools to self-assess and improve applications in development. This may come in the form of assistive writing technologies (eg, dedicated writing assistants, such as Grammarly, or more general-purpose generative AI platforms), or more purpose-built tools to score the likelihood of success and generate feedback on in-progress application materials.

Self-assessment systems might vary widely in terms of their generalisability across funders, but the most useful systems would be funder-specific. Automated evaluation of these systems would be possible to a limited extent by using previously collected applications and the peer reviews they received, but matching specific review feedback to particular aspects of an application is a challenge without good automated solutions for measurement.

Funder experimentation in this area would best be designed with established user experience research methods such as focus groups and user studies. These studies could work with applicants to assess utility and with scientific officers to assess how well aligned the feedback is with funder expectations. These types of evaluations would enable funders to compare different AI/ML strategies for self-assessment.

A more sophisticated evaluation of the impact of these systems on funding processes could be performed in two ways. Funders could use an opt-in approach to evaluate the outcomes of applicants who self-select for using a system compared with those who do not, or could provide a self-assessment system for a subset of funding calls and survey applicants to compare experiences where the tool was and was not made available.

6.3.6 AI FOR NAVIGATING FUNDING RESOURCES

The use of AI/ML technologies is not limited to decision-making or assessment of quality. Funders also provide extensive information to applicants and policymakers, which can be difficult to navigate efficiently without prior experience. Extensive research on interactive AI systems to provide customer-facing information from large textual knowledge bases (Fader et al, 2014; Xu et al, 2024) is increasingly translating into everyday web applications for question answering, which may be highly useful to help navigate funder resources.

Automated evaluation of interactive question-answering systems is not straightforward, but there are strong precedents for developing curated sets of resources, questions to query them, and expected answers that can provide a starting point for funders to experiment internally with different AI/ML methods (Chen et al, 2019).

More direct experimentation with these types of AI/ML applications would best be performed by deploying AI systems as an optional tool on funder websites for users to interact with, and conducting opt-in surveys for users to report their experiences. A randomised controlled

method is possible in this context, with users randomly assigned to a version of the website in which the AI tool is available or to the standard (non-AI) version of the funder website, and both populations surveyed on their experience locating the resources they need.

6.3.7 AI IN STRATEGIC PLANNING

Strategic planning is a highly complex process, and each funding organisation approaches it in its own way. However, the value of AI/ML systems in supporting discovery and learning from large volumes of data suggests potential value in specific areas, such as foresight activities or identifying directions for strategic funding calls. While experimentation with strategic planning is higher risk, identifying routes for success with AI/ML in this context can also be high-reward for funders willing to take on the experimental process.

To experiment with AI/ML in strategic planning, funders need to identify specific elements of their planning processes that may be amenable to AI/ML intervention. For example, landscape analysis can benefit from AI use to help identify emerging trends and patterns in current research (Holm et al, 2024). Funders could also use machine learning analysis of applications and outcomes from past strategic funding calls to identify particular characteristics or directions to which the research communities they serve are especially responsive.

Structured experimentation in strategic planning must be designed on a case-by-case basis. Funder research in this area is more likely to be analytic in nature, using AI/ML technologies to help learn about the research landscape or funding patterns, which can then support hypothesis-driven experimentation with the design of strategic funding calls.

"Where technologies and ways of working change, looking at past experiences provides invaluable insights for the future"

6.4 CLOSING WORDS

The use of artificial intelligence and machine learning in the work of research funding and assessment is an evolving area. Best practices and resources for funders exploring and applying AI/ML will continue to grow and change as AI technologies become more and more commonplace.

This handbook provides a snapshot of the current landscape of AI/ML use by research funders and a starting point for funders to inform their future AI journeys. Drawing on two years of co-productive work and discussion with an international community of funders, we have presented here:

- Working definitions of responsible AI for funders
- The wider context in which funders explore and apply AI/ML
- Practical steps involved in developing AI/ML applications in funding organisations
- Organisational issues and strategies informing the use of AI/ML
- Case studies of diverse real-world applications of AI/ML by research funders
- Takeaways, recommendations and future directions for funders seeking to explore or apply AI/ML in their work.

Some aspects of this handbook will, of course, become rapidly outdated by the pace of technological change. However, our focus on the people, process, practices and principles involved in the application of AI/ML technologies means that the vast majority of the content in this publication will, we hope, remain relevant for many years to come.

Where technologies and ways of working change, looking at past experiences provides invaluable insights for the future. Where new problems emerge, looking to challenges solved before will often help identify established solutions that stand the test of time. The AI journey for research funders is just beginning: this handbook is a stepping stone for any funder to help chart their own path.

References

Bendiscioli, S., Firpo, T., Bravo-Biosca, A., Czibor, E., Garfinkel, M., Stafford, T., Wilsdon, J., Woods, H. B., & Balling, G. V. (2022). The experimental research funder's handbook (2nd edition) (Version 4). Research on Research Institute. doi.org/10.6084/m9.figshare.19459328.v4

Blatch-Jones, A., Church, H., & Crane, K. (2025). Exploring the potential benefits and challenges of artificial intelligence for research funding organisations: a scoping review [version 1; peer review: awaiting peer review]. F1000Research, 14, 126. doi.org/10.12688/f1000research.160142.1

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research, 70, 245-317. doi.org/10.1613/jair.1.12228

Carbonell Cortés, C. (2024). AI-assisted pre-screening of biomedical research proposals: ethical considerations and the pilot case of 'La Caixa' Foundation. Available at: www.youtube.com/watch?v=O2DcXzEtCmg

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. Humanities and Social Sciences Communications, 8(1), 1-11. doi.org/10.1057/s41599-020-00703-8

Chen, A., Stanovsky, G., Singh, S., & Gardner, M. (2019). Evaluating question answering evaluation. In Proceedings of the 2nd workshop on machine reading for question answering 119-124. doi.org/10.18653/v1/D19-5817

CoARA. (n.d.) The coalition for advancing research assessment (CoARA). Available at: coara.eu. [Accessed 16 April 2025]

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2022). Introduction to algorithms. MIT press.

Cortés, C. C., Parra-Rojas, C., Pérez-Lozano, A., Arcara, F., Vargas-Sánchez, S., Fernández-Montenegro, R., David Casado-Marín, D., Rondelli, B. & López-Verdeguer, I. (2024). AI-assisted prescreening of biomedical research proposals: ethical considerations and the pilot case of 'La Caixa' Foundation. Data & Policy, 6, e49. doi.org/10.1017/dap.2024.41

Curry, S., Gadd, E., & Wilsdon, J. (2022). Harnessing the metric tide: indicators, infrastructures and priorities for UK responsible research assessment. (Version 2). Research on Research Institute. doi.org/10.6084/m9.figshare.21701624.v2

CWTS Leiden. (n.d.) VOSviewer: Visualising scientific landscapes. Available at: www.vosviewer.com

European Commission: Directorate-General for Research and Innovation. (2025). Living guidelines on the responsible use of generative AI in research. Publications Office of the European Union. [Updated 15 April 2025]. research-and-innovation.ec.europa.eu/document/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en

European Commission: Directorate-General for Research and Innovation, Cavicchi, B., Peiffer-Smadja, O., Ravet, J., & Hobza, A. (2023). The transformative nature of the European framework programme for research and innovation: analysis of its evolution between 2002-2023, Publications Office of the European Union. dx.doi.org/10.2139/ssrn.4581061
European Education and Culture Executive Agency. (2024). Launching EHESO – The European higher education sector observatory. Available at: www.eacea.ec.europa.eu/news-events/news/launching-eheso-european-higher-education-sector-observatory-2024-05-13_en. [Accessed: 16 April 2025]

Fader, A., Zettlemoyer, L., & Etzioni, O. (2014). Open question answering over curated and extracted knowledge bases. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 1156-1165. doi.org/10.1145/2623330.2623677

Future Research Assessment Programme. (n.d.) Future research assessment programme. Available at: www.jisc.ac.uk/future-research-assessment-programme

GO FAIR (n.d.). Fair principles. Available at: www.go-fair.org/fair-principles. [Accessed 17 April 2025]

Government of Canada. (18 November 2024). Guidance on the use of artificial intelligence in the development and review of research grant proposals. Available at: science.gc.ca/site/science/en/interagency-research-funding/policies-and-guidelines/use-generative-artificial-intelligence-development-and-review-research-proposals/guidance-use-artificial-intelligence-

development-and-review-research-grant-proposals

Holm, J., Waltman, L., Newman-Griffis, D., & Wilsdon, J. (2022). Good practice in the use of machine learning and AI by research funding organisations: insights from a workshop series (Version 1). Research on Research Institute. doi.org/10.6084/m9.figshare.21710015.v1

Holm, J., Newman-Griffis, D., & Petersson, G. J. (2024). Big data for big investments: Making responsible and effective use of data science and AI in research councils. Artificial Intelligence and Evaluation: Emerging Technologies and Their Implications for Evaluation, 120. doi.org/10.4324/9781003512493-7

Iansiti, M., & Lakhani, K. R. (2020). Competing in the age of AI: How machine intelligence changes the rules of business. Harvard Business Review, 98(1), 60-67.

Kitchin, R. (2019). Thinking critically about and researching algorithms. In The social power of algorithms. Routledge. 14-29.

Liu, L., Jones, B. F., Uzzi, B., & Wang, D. (2023). Data, measurement and empirical methods in the science of science. Nature human behaviour, 7(7), 1046-1058. doi.org/10.1038/s41562-023-01562-4

McCarthy, J. (n.d.) What is AI? / Basic Questions. Available at: jmc.stanford.edu/artificial-intelligence/what-is-ai. [Accessed 10 April 2025]

Miettinen, R., Tuunainen, J., & Esko, T. (2015). Epistemological, artefactual and interactional–institutional foundations of social impact of academic research. Minerva, 53, 257-277. doi.org/10.1007/s11024-015-9278-1

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. doi.org/10.48550/arXiv.1301.3781

MIT (n.d.). MIT AI Risk Repository. Available at: airisk.mit.edu. [Accessed 17 April 2025]

Newman-Griffis, D. (2025). AI Thinking: a framework for rethinking artificial intelligence in practice. Royal Society Open Science, 12(1), 241482. doi.org/10.1098/rsos.241482

NWO. (16 January 2024). NWO's preliminary position on generative AI in the application and review process. Available at: www.nwo.nl/en/nwos-preliminary-position-on-generative-ai-in-the-application-and-review-process

OECD (2024), "Governing with Artificial Intelligence: Are governments ready?", OECD Artificial Intelligence Papers, No. 20, OECD Publishing, Paris, doi.org/10.1787/26324bc2-en

OECD. (n.d.). Research and Development Statistics. Available at: www.oecd.org/en/data/datasets/research-and-development-statistics.html

Okasa, G., de León, A., Strinzel, M., Jorstad, A., Milzow, K., Egger, M., & Müller, S. (2024). A supervised machine learning approach for assessing grant peer review reports. arXiv preprint arXiv:2411.16662. doi.org/10.48550/arXiv.2411.16662

Okasa, G. & Jorstad, A. (2024). The value of pre-training for scientific text similarity: Evidence from matching grant proposals to reviewers. In Proceedings of the 9th edition of the Swiss Text Analytics Conference, Association for Computational Linguistics, 89–101. aclanthology.org/2024.swisstext-1.8

OpenAlex. (n.d.). OpenAlex. Available at: openalex.org

Oxley, K. & Gulbrandsen, M. (2025) Variability and negligence: grant peer review panels evaluating impact ex ante, Science and Public Policy, 52(2), 254–268. doi.org/10.1093/scipol/scae081

Price, S., & Flach, P. A. (2017). Computational support for academic peer review: A perspective from artificial intelligence. Communications of the ACM, 60(3), 70-79. doi.org/10.1145/2979672

REF. (n.d.). Research Excellence Framework. Available at: www.ref.ac.uk

European Parliament and the Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Official Journal, L 119, 1-88. data.europa.eu/eli/reg/2016/679/oj

Research Council of Norway. (2022). Societal impact of research financed by RCN: What can altmetrics tell us? Oslo: Research Council of Norway. www.forskningsradet.no/siteassets/publikasjoner/2022/altmetric-analysis_rcn-annual-report2021_report_22-1.pdf

Rushforth, A., Kvarven, A., Fraser, C., Newman-Griffis, D., Wilsdon, J., Alshamsi, M.A., Gogadze, N., Kolarz, P., McGuirk, S. (2025). Transforming Assessment: the 2025 Global Research Council survey of funder approaches to responsible research assessment. Research on Research Institute. Report. doi.org/10.6084/m9.figshare.28856480.v4

Shah, C., Anderson, T., Hagen, L., & Zhang, Y. (2021). An iSchool approach to data science: Human-centered, socially responsible, and context-driven. Journal of the Association for Information Science and Technology, 72(6), 793-796. doi.org/10.1002/asi.24444

Sivertsen, G. (2022). Publishing in the social sciences and its representation in research evaluation and funding systems. In Handbook on research assessment in the social sciences (pp. 238-261). Edward Elgar Publishing. doi.org/10.4337/9781800372559.00024

Spengeman, A., Mikhailov, D., Sorros, N., & Mankoo, A. (2021). Wellcome Data Labs: An Ethics Approach to Data Science Product Development-a Intentional Work in Progress. Available at: SSRN 3974121. dx.doi.org/10.2139/ssrn.3974121

Stevenson, C., Grant, J., Szomszor, M., Ang, C., Kapoor, D., Gunashekar, S., & Guthrie, S. (2023). Data enhancement and analysis of the REF 2021 Impact Case Studies. Santa Monica, CA: RAND Corporation. www.rand.org/pubs/research_reports/RRA2162-1.html

Thelwall, M., Kousha, K., Wilson, P. Makita, M., Abdoli, M., Stuart, E., Levitt, J., Knoth, P., & Cancellieri, M. (2023). Predicting article quality scores with machine learning: The UK Research Excellence Framework. Quantitative Science Studies, 4(2), 547–573. doi.org/10.1162/qss_a_00258

Thelwall, M., Kousha, K., Wilson, P. Makita, M., Abdoli, M., Stuart, E., Levitt, J., Knoth, P., & Cancellieri, M. (2022). Can REF output quality scores be assigned by AI? Experimental evidence. arXiv preprint arXiv:2212.08041. doi.org/10.48550/arXiv.2212.08041

Thelwall, M. (2024). Evaluating research quality with large language models: an analysis of ChatGPT's effectiveness with different settings and inputs. Journal of Data and Information Science, 241218-241218. doi.org/10.2478/jdis-2025-0011

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Williams, K., Michalska, S., Cohen, E., Szomszor, M., & Grant, J. (2023). Exploring the application of machine learning to expert evaluation of research impact. Plos one, 18(8), e0288469. doi.org/10.1371/journal.pone.0288469

Xu, Z., Cruz, M. J., Guevara, M., Wang, T., Deshpande, M., Wang, X., & Li, Z. (2024). Retrieval-augmented generation with knowledge graphs for customer service question answering. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval 2905-2909. doi.org/10.1145/3626772.3661370

Yuan, W., Liu, P., & Neubig, G. (2022). Can we automate scientific reviewing?. Journal of Artificial Intelligence Research, 75, 171-212. doi.org/10.1613/jair.1.12862

About the authors

Denis Newman-Griffis is Research Fellow at RoRI and Senior Lecturer and AI-Enabled Research Lead at the University of Sheffield, Centre for Machine Intelligence.

www.sheffield.ac.uk/cs/people/academic/denis-newman-griffis

d.r.newman-griffis@sheffield.ac.uk

orcid.org/0000-0002-0473-4226

@drgriffis

Helen Buckley Woods is Senior Research Fellow in Metascience at RoRI and University College London.

researchonresearch.org/team/helen-buckley-woods

h.woods@ucl.ac.uk

orcid.org/0000-0002-0653-9803

@HelenBWoods

Youyou Wu is Associate Professor in Psychology at University College London, IOE, UCL's Faculty of Education and Society.

profiles.ucl.ac.uk/85867-youyou-wu

youyou.wu@ucl.ac.uk

orcid.org/0000-0002-8081-0817

@youyouwooo

Mike Thelwall is Professor of Data Science in the Information School, University of Sheffield, UK.

www.sheffield.ac.uk/ijc/people/mike-thelwall

m.a.thelwall@sheffield.ac.uk

orcid.org/0000-0001-6065-205X

Jon Holm is Special Adviser and research evaluation expert at The Research Council of Norway.

jon.holm@rcn.no

orcid.org/0000-0002-5345-2639

Who we are

Research on research – also known as meta-research, metascience, or the science of science – uses a rich blend of traditional and emerging disciplinary and methodological approaches to test, evaluate and experiment with different aspects of research systems, cultures and decision-making.

/company/research-on-research

bsky.app/profile/rorinstitute.bsky.social

@researchonresearchinstitut6175

indieweb.social/@RoRInstitute

The Research on Research Institute (RoRI) was founded in 2019 with a mission to accelerate transformational and translational research on how research is organised, supported and assessed. We bring together people and organisations who are committed to informing and improving how research is funded, practised, communicated and evaluated.

From the start, RoRI's emphasis has been on applying and accelerating metascience through partnerships. Our greatest asset is our consortium, which now spans 20 research funders in 15 countries, between them investing almost US$30bn each year in research and innovation. To support these partnerships, and the projects that flow from them, RoRI has a 30-strong research and operational team, networked across eight universities in the UK, the Netherlands, Denmark, India, Argentina and Australia.

RoRI can connect you with people, projects and organisations that care about making research systems work better for everyone.