RoRI RESEARCH ON RESEARCH INSTITUTE

RoRI Working Paper No. 10

Good practice in the use of machine learning & AI by research funding organisations:

insights from a workshop series

Edited by: Jon Holm, Ludo Waltman, Denis Newman-Griffis & James Wilsdon

December 2022

The Research Council of Norway

Published and developed in collaboration with the Research Council of Norway.

About the authors

Jon Holm is Special adviser at the Research Council of Norway. jon.holm@rcn.no
https://orcid.org/0000-0002-5345-2639

Ludo Waltman is Associate Director of RoRI and Professor of Quantitative Science Studies at the Centre for Science and Technology Studies (CWTS) at Leiden University. waltmanlr@cwts.leidenuniv.nl; https://orcid.org/00000001-8249-1752; @LudoWaltman

Denis Newman-Griffis is Lecturer in Data Science at the Information School of the University of Sheffield. d.r.newman-griffis@sheffield.ac.uk; https://orcid.org/0000-0002-0473-4226; @drgriffis; https://www.sheffield.ac.uk/is/people/academic/denis-newman-griffis

James Wilsdon is Director of the Research on Research Institute (RoRI) and (from January 2023) Professor of Research Policy in the Department of Science, Technology, Engineering and Public Policy (STEaPP) at University College London (UCL). @jameswilsdon; https://orcid.org/0000-0002-5395-5949; jw@researchonresearch.org

For citation: Holm, J., Waltman, L., Newman-Griffis, D. & Wilsdon, J. (2022). *Good practice in the use of machine learning & AI by research funding organisations: insights from a workshop series*. RoRI Working Paper No. 10. December 2022. DOI: 10.6084/m9.figshare.21710015

Acknowledgements.

This workshop report summarises recent experiences in applying ML/AI among funders involved in the pilot phase of the Research on Research Institute (RoRI). The workshop was organised by the Research Council of Norway (RCN) in January 2021 supported by a programme committee drawn from RoRI member organisations. RoRI and RCN would like to thank the contributors who took part in the workshop series for sharing their valuable insights. We also would like to make it clear that the perspectives presented in this summary are those of the authors.

http://researchonresearch.org

Contents

1. Introduction

Like any general purpose technology, machine learning and artificial intelligence (ML/AI) presents both opportunities and challenges. For research funders, possible areas of application cover many of their traditional areas of operation, from strategic analysis through project selection to the monitoring and evaluation of project outcomes and impacts. Through all of these processes research funders are collecting and analysing data. Some of this data is structured in a way that enables traditional statistical analysis, like administrative data on project grants and scientific publications from those projects. Other types of data like proposal texts and outcomes other than scientific publications are typically unstructured and thus not fit for systematic analysis without massive manual labour.

Machine learning technologies present new opportunities for funders to make use of unstructured data collected by themselves or through other providers. Many of the funders participating in the workshop reported on the use of Natural language processing (NLP) in analysing text from proposals or publications. In some cases, NLP was used to increase efficiency by providing decision support for case officers. In other cases, the aim was to gain more in-depth knowledge of the properties of project portfolios in terms of analysis of research themes, disciplines or other relevant features.

Most of the reported use cases were linked to the core processes of selecting and following up on research projects. The most frequently used categories of analysis were disciplines and research themes. Because these processes consume the better part of personnel resources of a research funder, this is also where the greatest potential for increased efficiency is found. The more explorative uses of ML/AI in strategic analysis and in assessments of project outcomes and impacts beyond scientific publications seems presently to be less developed among research funders. Still, the political expectations for research to tackle societal challenges could serve as a motivation for investing more resources in following-up on projects beyond the publication of immediate results.

Funders experience increased political pressure to document that their investments in research are actually delivering impact. This pressure will probably be a driver for more analysis of longer term outcomes and impacts of funded projects. ML/AI present significant potential for adding to this analysis, by leveraging the characteristics of projects that have had impact in the past to help with assessing likely success of new project proposals. This potential must however be balanced by an awareness of how those past indicators of success may reflect known structural inequities in the research landscape, and how these inequities might then be reflected in the ML/AI systems trained on these data. A key question is therefore how to best harness ML/AI for assessing prospective impact in *proactive* ways that look forward to new needs and support a changing, more inclusive research culture.

The workshop also made visible a set of common challenges for research funders wanting to use machine learning as a part of their analytical tools.

To enter the world of machine learning a funder will need to go beyond the comfort zone of standardised point-and-click software provided by Microsoft or other commercial providers. Machine learning algorithms are often written in open source programming languages like Python. Installing the programming language on your PC is just a first step. The actual machine learning tools are provided by custom made data code that is not a part of Python.

Making use of these tools in practice requires three things: the tools themselves, the skills to understand and use them, and the computational resources to run them. Fortunately, many development teams in machine learning willingly share their algorithms and software packages through open source platforms like GitHub, making it easier to source cutting-edge tools for ML/AI. However, using these tools effectively requires the methodological skills and knowledge to first choose an algorithm fit to purpose and then to adapt it to the specific needs and context of use. Running ML/AI systems can also be quite demanding in terms of hardware: computers with high calculation capacity (such as graphics processing units (GPUs)) are typically required, though cloud computing services such as Microsoft Azure and Amazon Web Services may be used instead of dedicated on-premise hardware.

To achieve the learning in machine learning there is a need for training data. And as with humans, both the quality and the amount of training counts. The availability of relevant training data is often the biggest challenge in developing machine learning algorithms. In the case of linguistic data, this means that world languages like English are well served, whereas the availability of already trained algorithms is limited in smaller languages like Norwegian. Training accounts for a large part of the costs of developing functioning algorithms, which means that there could be great interest for funders to cooperate on the training of algorithms for the analysis documents that are commonly found across funders, like proposals for funding and research publications. Because most NLP-based algorithms are language specific, the potential for cooperation is greater when world-languages like English are used. This is the case in most of the natural sciences and medicine, whereas national languages are still frequently in use in the social sciences and humanities.

And then there are important ethical considerations. When machines are trained on data produced by humans, they will replicate the inherent world view and values of the culture in which the training data was created. So if we plan to replace human experts with machine learning algorithms in the assessment process of candidates for a specific type of scholarship, we need to be aware that any gender bias or other preferences commonly shared by human assessors will be replicated by the machine. On the other hand, machine learning may help us detect and adjust for such biases if they are found unwarranted.

And then, as with any new technology, machine learning will create winners and losers. In line with the principles of Responsible Research and Innovation (RRI)[1], it is essential that funders consider the positive and negative consequences for the community of researchers and for wider society. The question of how to use ML/AI responsibly was an important part of the workshop. One take-home message is that ML/AI is that the assessment of potential consequences should be a part of the development process from the beginning. A best practice example is provided by Wellcome Trust who has stated clearly what it means for them to use ML/AI responsibly and included social science experts in its development teams to secure ethically acceptable, sustainable and socially desirable outcomes.

2. Context and aims for the seminar series

Across diverse sectors, many see applications of machine learning (ML) and artificial intelligence (AI) as the latest example of 'general purpose technologies', with the capacity to boost productivity and alter working practices.[2] Within the scientific community, there is growing excitement about how ML/AI may be applied in research – particularly by optimising or accelerating innovative computational methods.[3] To date, there has been less discussion of applications of ML/AI in the design and management of the research system itself, and to processes of peer review, evaluation, synthesis and assessment—although a handful of funders are starting to experiment with this in various ways.[4]

As with all uses of ML/AI, enthusiasm about technological possibilities is tempered with concern about inbuilt biases and blind spots, and unintended consequences. In early 2021, Research Council of Norway (RCN) brought together a select group of research funders, in cooperation with the Research on Research Institute (RoRI), to share insights and

[1] https://www.rri-practice.eu/about-rri-practice/what-is-rri/

[2] See eg. https://www.nber.org/system/files/working_papers/w24001/w24001.pdf and https://ec.europa.eu/jrc/communities/sites/jrccties/files/eedfee77-en.pdf
[5] https://www.datarobot.com/wiki/prediction-explanations/

[3] e.g. Royal Society/Turing Institute (2019) The AI revolution in scientific research. https://royalsociety.org/-/media/policy/projects/ai-and-society/AI-revolution-in-science.pdf; Procter, R., Glover, B. and Jones, E. (2020) Research 4.0 - Research in the Age of Automation. Demos, September 2020. https://demos.co.uk/wp-content/uploads/2020/09/Research-4.0-Report.pdf

[4] e.g. the Russian Science Foundation https://rscf.ru/en/news/en-57/no-jumps-to-the-kings-row-rsf-pushes-the-new-ai-based-system-of-finding-reviewers/; and National Natural Science Foundation of China https://www.nature.com/articles/d41586-019-01517-8

uncertainties, and explore a range of actual or potential applications of ML/AI technologies. The purpose of the workshop was to:

- Create an arena for funders to share evidence and experiences with ML/AI techniques;
- Discuss and disseminate ‘good practice’ in emerging uses of ML/AI among RoRI partners;
- Explore what responsible uses of ML/AI would look like in the context of research management and assessment;
- Identify an agenda for further work through RoRI on these issues, linked to our broader work-stream on randomisation and experimentation.

The workshop was divided in three acts with different perspectives, organised as 3 hour online meetings over three consecutive weeks. This paper provides a summary of the discussions around current applications of ML/AI in research funding within a broader frame of responsible use of ML/AI technologies.

3. Emerging tools: possibilities and pitfalls

In the first act of the workshop, we asked a handful of experts to provide some perspectives on possibilities and pitfalls in the use of machine learning techniques. The content of these contributions are briefly indicated in this chapter with reference to the full presentations in the appendix.

Daniel Hook took as a starting point the difference between augmented intelligence and generalised artificial intelligence. The development from one to the other could be illustrated by the history of man-machine duels in chess and other games: These duels have developed from a rules-based approach to a learning approach where machines learn efficient strategies through analysis of earlier games. The machine's performance thus depends on the amount of data available.

A general problem with automated processes is that they may reduce diversity while searching for increased efficiency. This relates to the use of ML/AI at the macro level of the research system. At a more micro level, there is great potential in the day-to-day use of digital assistants in science (Alexa). Still we might ask how the providers of these assistants might use the knowledge of the researcher's behaviour and preferences through what Shoshana

Zuboff refers to as 'data exhausts'[5]. Daniel Hook suggested investigating how such data exhausts may be used to open up the everyday workings of science to the greater public instead: Data socialism.

Marija Slavkovik discussed ethical concerns in the use of automated decision systems. A basic principle is to pay attention to the power balance between service provider and service receiver: The automation should not alter this balance! In particular, there should be an instance of appeal if a user wants to challenge the decision of the system.

It is useful to distinguish between two types of decision systems: 1) Automating tasks that require human cognition 2) Decision-making as a mathematical process. Slavkovik focused on the second type of automated systems. We retain a very clear advice formulated as a general rule: Do not automate decisions if you need to break norms to do a good job!

A panel of experts ended the day with the two previous speakers, joined by Kuansan Wang and Ruth Pickering. Dr Wang pointed out that Biases in Peer Review often are due to cognitive limitations. Machines offer the advantage of being able to process superhuman amounts of information, extract useful patterns, and make precise computations. These strengths can be effectively combined with human strengths in assessment and decision-making. We should thus look into how we could use machine learning to improve research assessment by applying the GOTO principles: Good and Open data - Transparent and Objective algorithm. Ms Pickering, presented how Yewno uses AI to synthesise text in 'Knowledge graphs'. These graphs can show relationships between concepts and how the prominence of different concepts change over time in a living corpus of texts representing public opinion. The exact section of a text corresponding to a specific concept is also available to the user.

Further, the panel discussed the complexities in calculating the impact of algorithms when several algorithms interact, and, on the other hand, when people adapt their behaviour to the algorithm (gaming).

[5] Shoshana Zuboff (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. Public Affairs 2019

4. Applications in use or under consideration by research funders

The second act of the workshop was reserved for the presentation of use cases from the funding organisations. The use of ML/AI techniques in research funding and evaluation are still in its beginning. RoRI partners see a great potential in the use of these techniques in their operations, but few have made use of them to date. The workshop served to document and discuss the experiences made in a handful of organisations in the context of the broader issues of responsible evaluation and research policy instruments. Cases were presented by the Research Council of Norway (RCN), Research England - UKRI, Swiss National Science Foundation (SNSF) and Wellcome Trust.

The areas of application included selection of proposals (process automation, panel selection), portfolio analysis (tagging of grants, alignment with policy goals) and assessment of research quality (see presentations in Annex B). A common characteristic of these applications being that ML/AI was used at specific stages of the work process to support human judgement or to automate routine operations. These stages typically include checking the eligibility of proposals (rules-based selection), matching of proposals to relevant experts, checking the quality of peer-review and analysing the portfolio of funded projects.

Methods generally depend on Natural Language Processing (NLP) and various algorithms that are used to analyse textual properties of written documents (TF/IDF, BERT, Support Vector Machine). In some cases, funders combine customised algorithms with commercial off-the-self services, like RCN who uses an in-house algorithm to distribute proposals on review panels and Elsevier's Expert Lookup to identify relevant experts to serve on the panels. Others, like SNSF, have developed their own algorithm to match individual proposals to reviewers based on machine reading and analysis of the proposal text on the one hand, and data on the publications of potential reviewers from Elsevier's abstract and citation database on the other hand (Titles, Keywords, Abstracts, Journal titles Subject areas and Publication years).

Limitations of the various methods were discussed. A number of applications were based on methods that identify characteristics of documents based on word frequency (TF/IDF). These methods are not able to distinguish between different meanings of the same word, which is a particular challenge for parts of the social sciences and humanities (SSH). While a term-based approach may work well to distinguish various areas and directions of research in the biomedical sciences because it uses specified concepts like ‘Covid19’ and ‘SARS-CoV-2’, it may miss such distinctions within SSH where researchers tend to use more common language with fewer specific concepts.

To teach a computer to distinguish between different meanings of the same word, we need a language model that includes the context in which the word occurs. While word frequencies can be calculated by quite simple statistical methods, machine learning helps us develop more sophisticated language models that take into account more complex relations within a given text. Google introduced such a model in 2018, the neural network model BERT. Wellcome has used this model in combination with a Support Vector Machine (SVM) in its portfolio analysis with good results. Interdisciplinary research remains a particular challenge, often perceived as noise in the results. Still, Wellcome found that this 'noise' was in fact useful as an indicator of interesting areas for further investigation by other methods or human expertise.

The possible use of ML/AI to assess different aspects of research quality was also explored. SNSF has supported the development of an algorithm to assess the quality of journal peer review that may in the future be applied to grant proposal reviews. The algorithm was trained on a dataset of reviews that was rated for different aspects of quality (thoroughness and helpfulness of comments). Even more ambitious in terms of conquering the holy grail of peer review, Research England is planning an experimental pilot study on automated assessment of scientific publications, based on REF2021 data. This study will provide valuable evidence on opportunities and challenges for integration of ML/AI into peer review processes.

Data

The availability of relevant data is crucial in machine learning. In general, more data will allow for better learning and thus more precise results. In the cases presented by RoRI partners, both supervised and unsupervised learning was used. In supervised learning, the algorithm is trained to replicate some known outcomes that are determined beforehand. The experiment planned by Research England on replicating REF2021 publication reviews and scoring is one such example. Another was presented by RCN who has trained algorithms to replicate the tagging of funded projects based on historical data of projects tagged by RCN staff.

Algorithms may also be used to identify patterns within a dataset without any predefined specification of outcomes. Such ‘unsupervised learning’ can be valuable for exploratory data analysis to discover new and unexpected patterns, in contrast to supervised learning with a limited set of expected outcomes and a predefined ‘ground truth’ of human assessment.

However, unsupervised learning often requires large volumes of data in order to identify useful, novel patterns. Few funders have sufficient data and computational capacity to develop such algorithms. One option is to use algorithms trained on more general datasets and adapt them for use in a research funding context. The experience of Wellcome Trust in adapting BERT to analyse research within specific disciplinary domains appeared as promising. This adaptation is done by training a pre-trained algorithm on a dataset from the relevant scientific discipline, so-called transfer learning.

Disciplinary differences and interdisciplinarity

Fields that use common terms in a non-specified way, like in many of the social sciences and humanities, are more difficult to analyse and classify by traditional NLP (TF-IDF) techniques. BERT performs better because of context aware embeddings. BERT can also be quickly adapted to new fields by a process of pre-training on mixed-domain data and transfer learning to domain-specific data. Interdisciplinary projects constitute a specific challenge for algorithmic identification. When working on topic modelling, Wellcome has experienced that the BERT perceived interdisciplinarity as noise, but this 'noise' was actually the most interesting area to investigate.

Organisation

How to best use ML/AI techniques to enhance operations is also a question of organisation. Solutions vary across funders from building designated data science departments, like the Data Lab at Wellcome, to depending largely on external consultants in the case of RCN. Regardless of where data science sits, a common challenge is how to guide technological development by organisational goals and ethical concerns. As in all interdisciplinary work, it takes time to build a shared understanding of technological possibilities and organisational and societal purposes. The Data Lab at Wellcome provides an interesting example on how to build this shared understanding by setting up interdisciplinary teams including both data scientists and staff trained in social science. Another challenge is computational capacity. To fully exploit the potential of ML/AI, funders will need access to appropriate data infrastructures.

5. Ethical and responsible uses of ML/AI in research funding and evaluation

The third and final workshop was dedicated to a discussion on responsible use of ML/AI and principles that could be developed to guide such use. The topic was approached from different points of view.

Alasdair Cowie-Fraser from the data team at Wellcome Trust provided an organisational perspective on how to bring ethical concerns into the data analysis process. As previously mentioned, social scientists work together with data scientists and software developers in interdisciplinary teams at Wellcome. One role of social scientists is to identify negative unintentional consequences of the algorithms being developed, in a way that can inform the development process. This means that the assessment of potential negative consequences has to be repeated for each iteration of the design. The full integration of the social scientist in the team is necessary to be able to work at the same speed and to the same rhythm as the software developers and data scientists. Further, there are many other potential contributions of integrating social scientists into development teams that can help funders to achieve their desired outcomes.

On the other hand, this integration may create a tension between their role of the impartial observer and the alternative of being an active participant. To secure the professional detachment and objectivity of the impact assessment, Wellcome suggests pairing up the embedded researcher with another social scientist outside the team to provide a critical view and the necessary checks and balances on their analysis. See the Medium page of Wellcome Data Labs for more information.

Jeroen van den Hoven, professor of ethics & technology, Delft University of Technology, presented a more general framework for value-sensitive design, which aims to take human values into account throughout the whole design process. A central principle is to translate abstract concepts, such as fairness, into specific requirements to be observed by the algorithms. This process of translation could go through several phases including conceptual investigations aiming at understanding and articulating the various stakeholders of the technology, empirical studies to inform the designers' understanding of the users' values, needs, and practices, and finally the design of systems to support values identified in the conceptual and empirical investigations[6]. By linking design choices to stakeholder values, this

[6] see https://en.wikipedia.org/wiki/Value_sensitive_design

process creates transparency. To build trust we also need to look beyond the design of the specific algorithms to the people and institutions using them, and the mechanisms to hold those instances to account for the conclusions they draw from data and algorithms.

Two presenters, **Josh Nicholson, CEO of scite.ai and David Pride, research associate at the Open University Knowledge Media Institute (KMI),** presented examples of how machine learning can provide a more nuanced and complete understanding of the meaning of citations in the scientific literature. Scite.ai provides a service called Smart citations that display the context of the citation and describe whether the article provides supporting or contrasting evidence. It uses ML and deep learning analysing the full text of scientific publications. David Pride discussed how we could overcome the limitations of citation metrics, treating all citations as equal, even though we know that citations may have different meanings. Pointing out that it is far more interesting and useful to understand why a paper is cited than just that something is cited, he suggested several ways to move forward in our understanding of this question. Plans at KMI include making a survey among authors about their motivations for giving citations and potentially use such data to train algorithms to identify these different motivations. Data sets for training of algorithms are small at KMI and focus on just a few scientific domains. The variations of citation practices across disciplines remains a challenge. It is uncertain if algorithms trained in one specific discipline could be used in other disciplines.

Discussion

Three topics were prominent in the discussion: How to identify bias? How to involve relevant stakeholders in the development of analytical tools? How to achieve transparency and trust?

Regarding biases and fairness of algorithms a main concern was how to avoid that analysis based on machine learning reproduce known biases in the system, such as the Matthew effect, gender biases, prestige of journals and institutions etc. Ultimately what should count as fair must be determined by humans. Algorithms trained on historical data may actually be useful in making visible the biases inherent in expert assessment. Mechanisms need to be put in place to make adjustments to the algorithms that correspond to relevant stakeholders' conceptions of fairness. We should still be mindful that concepts of fairness may differ between different spheres of justice and contexts (hiring, funding, publishing).

A main concern when involving stakeholders is to make the questions to be discussed understandable also by people with limited data literacy. It may help to focus on the impact of algorithms rather than inputs. One problem is that the developers of analytical tools have limited knowledge and control of their potential uses. Still, developers may expand their insights into intended and unintended consequences of their analytical tools by involving a larger set of stakeholders, not only the direct ones, but also the indirect ones.

As reported above, transparency and trust was discussed as a quality of the design process itself in its ability to translate values held by relevant stakeholders into requirements for the technological solution (value-sensitive design). The issue of transparency is also related to openness of data and code. Not all providers of research analysis make their data and programming code publicly available. This makes it difficult to have an open discussion on the fairness of the algorithms. Open access to relevant training data is also important to give researchers, funders and various other stakeholders the possibility to scrutinise existing algorithms and develop their own.

6. Conclusions: priorities, possibilities & avenues for further research

The motivations and aims for current experiments in the use of ML/AI vary across funders. At the moment, the need for more efficient operation of selection processes and grant management appears as the most prominent drivers for the adoption of new technologies. On the other hand, several funders also look at how ML/AI can increase the effectiveness of their funding, helping to assure that it actually meets the goal set by boards and governments. The algorithm developed to check the quality of peer review at SNSF is one such example, the use of algorithms to secure more consistent tagging of grants at RCN is another. ML/AI may also help discover the inherent biases in peer review so that these can be discussed and corrective measures applied if deemed necessary

RoRI may offer a context for broader reflection on how to use ML/AI technologies responsibly in this context, and how to choose the most suitable methods for various purposes. Cooperation may span from technical issues – as choice of methods and training of algorithms – to a more general discussion on ethical, legal and societal concerns. There might also be scope for joint research on the impact of ML/AI on the research system. The following list of possible themes for further cooperation were identified in the workshop:

Technical issues

1. ***Availability of data for training and testing algorithms is a big issue.*** Pooling of data from several funders could be a way of creating larger datasets. This would require *standardization* and a sufficient understanding of the local context to secure comparability. Partners could plan data collection and curation so that the data can be shared and explored by AI methods.

2. ***There is a need to share experience with various methods*** e.g. use of more advanced algorithms like BERT may provide a more context sensitive interpretation of linguistic representation of a discipline / field of research.

3. ***How to share code?*** Wellcome's team shares open source code at https://github.com/wellcometrust/wellcomeml, and a blog at https://medium.com/wellcome-data-labs

4. ***How to implement ML/AI systems?*** ML/AI requires high levels of computation and technical expertise. This could be provided through in-house competence and consultants, and through some combination of on-premise hardware and negotiated cloud computing services. Maintenance of the systems over time also needs to be considered.

Ethical, legal and societal issues

1. ***Develop protocols to assess ethical consequences*** of the use of ML/AI.

2. ***Problems of ground truth:*** When training datasets are based on previous human judgement, these may contain biases or misconceptions that we do not want to be learned by the algorithms. Thus, when assessing algorithmic results for biases, the reference cannot be another subset of the same dataset. Rather, we need to identify the types of (mis)conceptions underlying earlier assessments that we would like to adjust.

3. ***Questions regarding the combination of machine learning and human judgement***
    a. When should we use ML/AI to develop automated systems, and when should we use it as support for human expert decision–recognising that these are not mutually exclusive?
    b. Do we expect algorithms to replicate human judgement, just more efficiently, or would we like to make improvements in terms of outcomes? Such improvements could include more consistent use of assessment criteria and data, compliance with formal requirements for assessment statements and identification and correction of biases.
    c. Use of ML/AI to provide decision support to the panels in the form of metrics or other 'objective' data.
    d. Monitor the development of the portfolio to look for systemic bias.

4. ***Transparency: How could we explain the process of algorithmic prediction in a way that is understandable – at least in principle – for all interested parties.*** There are barriers of two kinds to such transparency. Traditionally, ML models have not included insight into why or how they arrived at an outcome. This makes it difficult to objectively explain the decisions made and actions taken based on these models.[5] On the other hand, there are commercial tools available for research analysis where access to algorithms and/or relevant data is hindered by

intellectual property rights. What standards should we set for transparency, e.g., through documentation and reporting at all stages of developing and using ML/AI? What are appropriate governance structures and processes to ensure that the implementation of technology serves the needs and aims of the institution? Some countries have developed guiding principles for Open Research Information, such as the Netherlands.[7]

Proposed directions for future work

Some possible strands of future work through the RoRI consortium were suggested in the group discussions of the workshop:

A. ***Develop a framework for responsible AI/ML, building on responsible assessment.***

B. ***Facilitate a discussion on how to discover and adjust for well-known biases in human expert assessment relative to age, gender or other non-relevant properties of the researcher.***

C. ***Guide funders in their choice of specific methods and in how to implement AI/ML techniques in their business processes: use off-the-shelf solutions or building tools themselves.***

D. ***Develop standards for evaluating algorithms.***

E. ***Set up a repository for training data for AI/ML algorithms, following unified standards for sharing of data sets. One benefit of such a repository could be to learn more about how different methodological approaches work out on the same data.***

F. ***Set up a community of interest around AI/ML in research funding where practitioners and researchers can share experiences.***

G. ***Organise joint projects on how to use AI/ML in research systems analysis more generally.***

H. ***Bring together people in funding organisations who do research analysis and people in funding organisations that do open science; openness will promote data sharing and ethical approaches to the use of AI/ML.***

[7] See https://www.leidenmadtrics.nl/articles/seven-guiding-principles-for-open-research-information

A vision for moving forward

This section was developed by RoRI during the fall of 2022 to identify key directions and next steps for building on the discussions that emerged in the January 2021 workshop.

As research funders continue to explore and experiment with the use of AI and ML techniques in their work, there is a need for tools to facilitate a shift in paradigm from a project-oriented perspective to a more organisational perspective, in which the design, implementation, and management of AI and ML tools are situated in the broader context and goals of the organisation.

We propose to support this paradigm shift through developing a suite of theoretical and practice-oriented tools, aimed at facilitating a holistic and interdisciplinary view towards use of AI and ML as situated tools in the funding context. Based on the themes and issues identified through the workshop series, we suggest specific interventions at each stage of the AI technology lifecycle, including initial goal-setting and design, technical implementation, and lifecycle management.

Strategic and organisational issues

One key aspect of this work, which is under-studied in the AI/ML literature, is to Identify strategic and organisational challenges for the use of AI in research funders and research policy agencies. These issues must be explored in partnership with funders, e.g. through discussions and semi-structured interviews with key staff at RoRI-partners. Questions of interest could include:

- What types of digitalisation strategies exist? How to find the right balance between use of consultants and building in-house competence? What is the role of training programmes vs hiring new personnel with relevant competence. How are ethical considerations built into the development and use of ML/AI technologies?
- How is co-production between different competencies achieved, for instance in the interpretation of results, evaluation of algorithms, and the management of bias? Are analyses performed centrally or decentralised within the organisations?
- How is communication of results and uptake within the organisations achieved? How are methods and results communicated outside of the organisation?

Design

Using AI and ML methodologies for practical problems within organisations requires a targeted design process, to identify the characteristics of the problem, the data to analyse, and how the output of AI/ML systems will be used. Key steps in addressing AI/ML design considerations include:

- Develop a framework to guide the process of translating a proposed application of AI/ML into a concrete problem formulation with a specific focus on how to deal with the issue of construct validity.
- Identify key strategies for taking relevant public policies and the specific organisational contexts of deployment into consideration throughout design.
- Pilot the use of this framework to facilitate interdisciplinary co design of AI/ML technologies, integrating technical and social science perspectives on the data used, the role of the technology within organisational processes, and the appropriateness of the design.
- Convene a community of funders and AI/ML researchers to discuss key considerations and best practices in designing AI/ML tools along the spectrum of decision support and process automation.

Implementation

Once a potential problem to tackle with AI/ML approaches has been identified and an initial design defined, funders are faced with a distinct set of concerns in how that design can best be implemented both from a technical and organisational perspective. We highlight the following actions as valuable steps towards supporting more formalised implementation processes for AI and ML in the funder context:

- Guide funders on approaches to break down complex needs for AI/ML tools into well-defined subproblems, and selecting appropriate models and methodologies for these subproblems.
- Develop best practices and recommendations to guide funders in the selection of appropriate tools (e.g., software packages, computing environments, deployment infrastructure) for AI/ML solutions, and identifying robust strategies to adapt off-the-shelf tools to particular funder contexts.

- Develop rubrics for selecting and assessing appropriate data for AI/ML development and evaluation, particularly tackling the tension between data representativeness and inequity (i.e., "rich get richer" patterns).

Management

Finally, like any other technology AI/ML tools are not one-off solutions but must be actively managed as part of dynamic organisational processes over time. The practical concerns and principles of managing AI/ML systems in living organisations have been the focus of relatively little research, and AI and ML pose unique challenges for responsible management. Valuable first steps in better understanding management needs include:

- Work with funders to identify strengths and challenge points in current human processes and AI/ML tools, to develop a better understanding of how AI/ML tools can best be used to complement (rather than replace) human decision making.

- Investigate strategies for assessing AI/ML tools beyond technical performance metrics, to evaluate their impact on process quality and efficiency.

- Identify key considerations and decision points for managing AI/ML tools over time, including periodic assessment, re-training ML models, and replacement or decommissioning.

Each of these targets contributes in different ways to the themes that emerged from the workshop discussion, of:

1. Responsible and ethical use of AI/ML;
2. Transparency in AI/ML design and communication; and
3. Reusable, practical pathways for developing, deploying, and sharing AI/ML solutions.

These targets will be best addressed through a co productive approach, integrating organisational, social, and technical perspectives from the beginning. We propose two primary methods of engagement around these efforts moving forward:

Community-building seminar series

A regular seminar series will provide an opportunity to bring together stakeholders across different funders and components of the AI/ML process, to facilitate ongoing discussion, knowledge exchange, and rapid feedback. Such a seminar series can serve several key functions in producing organisationally-oriented strategies for using AI/ML in the research funding context, including:

- Early and often opportunities for “taking the temperature” on ideas for addressing different parts of the AI/ML process, and for sharing experiences, challenges, and strategies across the community of funders;
- Mini-workshops to tackle individual aspects of design, implementation, or management from multiple perspectives;
- Practising strategies for clear and transparent communication of the AI/ML process, including practising using the conceptual frameworks and rubrics under development;
- Peer-to-peer learning to share knowledge and findings across projects and pilot studies.

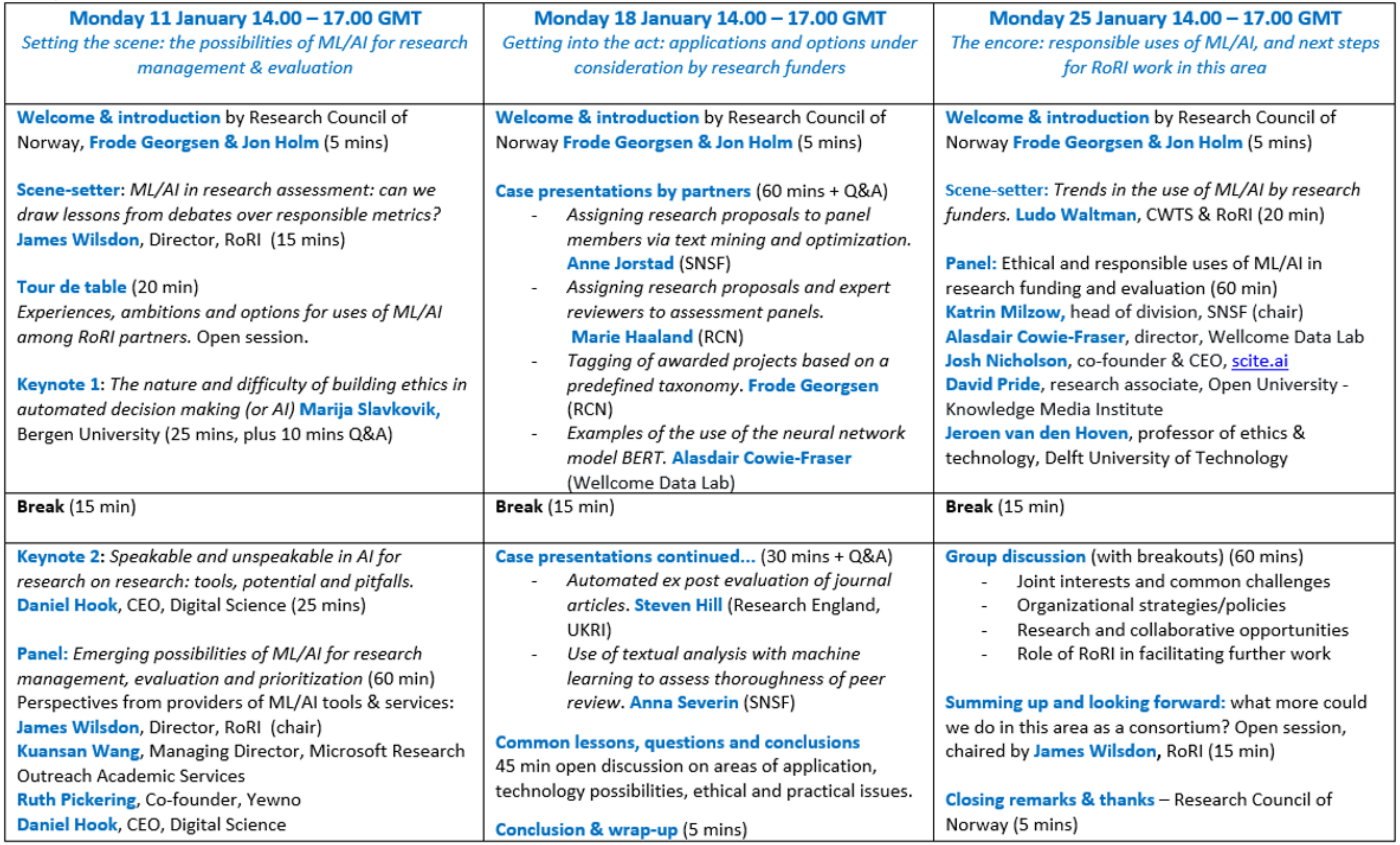
The discussions facilitated by a seminar series focused on knowledge and practice exchange among RoRI partners will be the cornerstone of the co-productive process for addressing these various targets and maintaining cross-partner engagement and focus throughout the project.

Pilot studies with individual funders

The targets proposed above do not need to be addressed in sequence or in a single setting. We suggest that individual targets can be the subject of pilot studies with individual funders, with well-defined scope for feasibility and rapid iteration. For example, the proposed translational design framework could be piloted in partnership with the Wellcome Trust, to build on their existing processes for integrating social scientists into the design and evaluation of AI/ML solutions. The monthly seminar series can provide a venue for building these partnerships and setting the scope of pilot studies.

Addressing the issues highlighted through the workshop series in a robust and forward-looking way will require significant effort and change, and it will not happen overnight. We have proposed a variety of potential targets for making progress along this path, with an eye firmly fixed on the overall organisational paradigm in which AI and ML tools will be situated. By taking small, manageable steps towards selected targets, and with ongoing engagement across the diverse body of stakeholders involved in using AI/ML in the research funding context, we can move towards well-defined and transparent use of AI and ML as tools in the funding management toolbox.

Appendix A: Workshop series agenda



Monday 11 January 14.00 – 17.00 GMT *Setting the scene: the possibilities of ML/AI for research management & evaluation*	**Monday 18 January 14.00 – 17.00 GMT** *Getting into the act: applications and options under consideration by research funders*	**Monday 25 January 14.00 – 17.00 GMT** *The encore: responsible uses of ML/AI, and next steps for RoRI work in this area*
Welcome & introduction by Research Council of Norway, **Frode Georgsen & Jon Holm** (5 mins) **Scene-setter**: *ML/AI in research assessment: can we draw lessons from debates over responsible metrics?* **James Wilsdon**, Director, RoRI (15 mins) **Tour de table** (20 min) *Experiences, ambitions and options for uses of ML/AI among RoRI partners.* Open session. **Keynote 1**: *The nature and difficulty of building ethics in automated decision making (or AI)* **Marija Slavkovik,** Bergen University (25 mins, plus 10 mins Q&A)	**Welcome & introduction** by Research Council of Norway **Frode Georgsen & Jon Holm** (5 mins) **Case presentations by partners** (60 mins + Q&A) - *Assigning research proposals to panel members via text mining and optimization.* **Anne Jorstad** (SNSF) - *Assigning research proposals and expert reviewers to assessment panels.* **Marie Haaland** (RCN) - *Tagging of awarded projects based on a predefined taxonomy.* **Frode Georgsen** (RCN) - *Examples of the use of the neural network model BERT.* **Alasdair Cowie-Fraser** (Wellcome Data Lab)	**Welcome & introduction** by Research Council of Norway **Frode Georgsen & Jon Holm** (5 mins) **Scene-setter:** *Trends in the use of ML/AI by research funders.* **Ludo Waltman**, CWTS & RoRI (20 min) **Panel:** Ethical and responsible uses of ML/AI in research funding and evaluation (60 min) **Katrin Milzow,** head of division, SNSF (chair) **Alasdair Cowie-Fraser**, director, Wellcome Data Lab **Josh Nicholson**, co-founder & CEO, scite.ai **David Pride**, research associate, Open University - Knowledge Media Institute **Jeroen van den Hoven**, professor of ethics & technology, Delft University of Technology
Break (15 min)	**Break** (15 min)	**Break** (15 min)
Keynote 2: *Speakable and unspeakable in AI for research on research: tools, potential and pitfalls.* **Daniel Hook**, CEO, Digital Science (25 mins) **Panel:** *Emerging possibilities of ML/AI for research management, evaluation and prioritization* (60 min) Perspectives from providers of ML/AI tools & services: **James Wilsdon**, Director, RoRI (chair) **Kuansan Wang**, Managing Director, Microsoft Research Outreach Academic Services **Ruth Pickering**, Co-founder, Yewno **Daniel Hook**, CEO, Digital Science	**Case presentations continued…** (30 mins + Q&A) - *Automated ex post evaluation of journal articles.* **Steven Hill** (Research England, UKRI) - *Use of textual analysis with machine learning to assess thoroughness of peer review.* **Anna Severin** (SNSF) **Common lessons, questions and conclusions** 45 min open discussion on areas of application, technology possibilities, ethical and practical issues. **Conclusion & wrap-up** (5 mins)	**Group discussion** (with breakouts) (60 mins) - Joint interests and common challenges - Organizational strategies/policies - Research and collaborative opportunities - Role of RoRI in facilitating further work **Summing up and looking forward:** what more could we do in this area as a consortium? Open session, chaired by **James Wilsdon**, RoRI (15 min) **Closing remarks & thanks** – Research Council of Norway (5 mins)

Appendix B. Organisers, chairs and speakers

Programme committee

James Wilsdon, RoRI (chair)
Jon Holm, Research council of Norway (local organiser)
Ludo Waltman, CWTS/Leiden
Vibecke Helene Ahmed Viul, Research council of Norway
Marie Haaland, Research council of Norway
Marianne Stephanides, Austrian Science Fund

Speakers and panel chairs

Alasdair Cowie-Fraser, director, Wellcome Data Lab
Frode Georgsen, Department director, Research council of Norway
Marie Haaland, Research council of Norway
Daniel Hook, CEO, Digital Science
Jeroen van den Hoven, professor of ethics & technology, Delft University of Technology
Anne Jorstad, Swiss National Science Foundation
Katrin Milzow, head of division, Swiss National Science Foundation (SNSF)
Josh Nicholson, co-founder & CEO, scite.ai
David Pride, research associate, Open University - Knowledge Media Institute
Ruth Pickering, Co-founder, Yewno
Marija Slavkovik, Bergen University
Ludo Waltman (CWTS, Leiden)
Kuansan Wang, Managing Director, Microsoft Research Outreach Academic Services
James Wilsdon, University of Sheffield and RoRI director

Speaker slides will be available as supplementary material on the Figshare page for this report at DOI: 10.6084/m9.figshare.21710015

The Research Council of Norway

RoRI RESEARCH ON RESEARCH INSTITUTE

https://researchonresearch.org/