Enhancing Trust, Integrity, and Efficiency in Research through Next-Level Reproducibility Impact Pathways

Deliverable D4.1 – The Future(s) of Reproducibility in Research

30/11/2023

Lead Beneficiary: Vrije University Medical Centre

Authors: Joeri Tijdink, Barbara Leitner, Nicki Lisa Cole, Serge P.J.M. Horbach, Simone Kopeinik, and Tony Ross-Hellauer

Reviewers: Thanasis Vergoulis, Christopher Osborne

Funded by the European Union

Prepared under contract from the European Commission
Grant agreement No. 101094817

EU Horizon Europe Research and Innovation action

Project acronym:	**TIER2**
Project full title:	**Enhancing Trust, Integrity, and Efficiency in Research through Next-Level Reproducibility Impact Pathways**
Start of the project:	January 2023
Duration:	36 months
Project coordinator:	Dr. Tony Ross-Hellauer

Deliverable title:	The Future(s) of Reproducibility in Research
Deliverable n°:	D4.1
Version n°:	1.0
Nature of the deliverable:	Report
Dissemination level:	Public

WP responsible:	WP 4
Lead beneficiary:	Stichting VUmc

TIER2 Project, Grant agreement No. 101094817

Due date of deliverable:	Month n° 11
Actual submission date:	Month n° 11

Deliverable status:

Version	Status	Date	Author(s)
0.7	Draft/Final	13 November 2023	BL, JT, SH, TRH, NLC, SK VUmc, AU, KC
0.8	Review	20 November 2023	TV, CO ARC, UOXF
0.9	Comments addressed	27 November 2023	JT, BL, NLC, SH
1.0	Final editing	29 November 2023	JT, BL, TRH

Table of contents

Executive Summary

This report presents the main findings from the work undertaken to in TIER2 Task 4.1 “Future studies to identify priorities from the stakeholder community to predict future of reproducibility and identify actionable steps”. The study aims to determine the future of reproducibility. We do this by exploring potential scenarios of the future, examining what steps are needed to make these scenarios workable in the future and detect barriers and facilitators for reproducibility in the current research system and the future one. In addition, we also aim to identify the key factors and stakeholders that may have the ‘power’ to influence these scenarios. To address these aims, we formulated four research questions:

1. What are the preferred futures of reproducibility?
2. What are the alternative futures: dystopian, realistic, outlier?
3. What are the enablers and barriers on the way to the preferred future or reproducibility more generally?
4. Which stakeholder groups bear responsibility for addressing concerns related to reproducibility and have the ability to trigger change for reproducibility in the future?

To answer the questions above, and as presented in the Methodology section of this document, we have conducted four workshops with different stakeholders (publishers, funders, and researchers) from different domains (machine learning and qualitative research). The study employed the method of futures studies, which aims to explore and predict the future by envisioning alternative future scenarios, to inform strategies for action to eventually shape the desired future at different levels.

The different components of the workshops drew light on the most important stakeholders who have the greatest influence in fostering change for reproducibility. The themes that emerged from exploring the preferred futures for reproducibility in different stakeholder categories are insightful. Furthermore, we identified enablers that may lead to these futures and the barriers that need to be overcome to achieve these futures. Furthermore, we describe the alternative futures: a dystopian, a realistic and some outlier-scenarios. Participants emphasized the need for a cultural shift towards quality over quantity, with reproducibility becoming the norm. This involves developing common understandings of definitions across disciplines, implementing incentives for reproducible practices, and developing clear guidelines and infrastructures to foster reproducibility. Lack of consensus on terminology, poor levels of awareness, dysfunctional incentives, challenges of epistemic diversity, and the risk of marginalization were identified as the main barriers for cultural change.

The outcome of our work is aimed at furthering the discussion on reproducibility. Specifically, the next steps are to use the results in the next stages of the TIER2 project to help develop tools that can enable and facilitate the preferred future. Furthermore, we can examine whether the tools are effective for reaching the preferred futures for the stakeholders. The enablers and barriers identified during the workshops can have a significant influence on the implementation of the pilot tests in the next phase of the project, particularly for helping to successfully implement the tools for each stakeholder group. This can help in steering towards the right direction and being prepared for certain barriers and to help use certain enablers.

1. Introduction

1.1.The reproducibility crisis

Reproducibility is often claimed as a central principle of the scientific method (Popper, 2002). Definitions for reproducibility have varied, but ultimately reproducibility refers to the possibility for members of the scientific community to obtain the same results and conclusions from the same data by following the documentation shared by the original investigators (Nichols et al., 2021). As several terms that have similar meanings are being used, it is necessary to highlight the difference between them. We use replicability to refer to the phenomena where independent researchers obtain consistent results across studies aimed at answering the same scientific question, where each study has obtained its own data (Nichols et al., 2021).

Recent years have seen perceptions of a "reproducibility crisis" grow in a variety of disciplines. The failure to reproduce results was found in studies in biomedicine with high rates as much as 50-90% in oncology research (Walters et al., 2019) and psychology failing to reproduce 61% of a set of prominent experiments (Open Science Collaboration, 2015). And most recently, Cobey and colleagues calculated an average of 54% of research results that could not be replicated across disciplines (Cobeyet al., 2023). These results launched discussions, with some demanding reproducibility issues to be put firmly on the agenda of all disciplines, including the humanities. Consequently, the framing of a crisis was not adopted by all disciplines.

This triggered a response from other research communities, which oppose the universal requirement for studies to adhere to strict reproducibility standards. Instead, they argue that in some epistemic contexts, reproducibility should only be cautiously considered as the pillar of academic scrutiny (Leonelli, 2018). They argue that reproducibility should be considered more important for data intensive disciplines, but that understandings of reproducibility and the implications for methodological and reporting purposes should be different for other methodological and disciplinary contexts.

We reside in the latter camp, believing that understandings of reproducibility should be sensitive to epistemic contexts and respect disciplinary diversity in methods and research, and reporting practices. This is well reflected in the work of Sabina Leonelli (2019B). She highlights that reproducibility should be reconsidered as a pillar of research quality and distinguishes six methodological contexts that interpret reproducibility differently by differentiating between the interpretation of reproducibility, the degree of control researchers have over their environment, and the extent to which they rely on statistical interference (Leonelli, 2019). Leonelli henceforth distinguishes between computational reproducibility, direct experimental reproducibility, scoping, indirect and hypothetical reproducibility, reproducible expertise, reproducible observations and irreproducible research (e.g., participant observations). This resonates well with the reproducibility spectrum proposed by Peng (2011). The reproducibility spectrum aims to fill the gap in the scientific evidence-generating process between full replication of a study and no replication by describing a spectrum of possibilities for a study to be more or less reproducible than another one depending on whether the data and code made available by the original researchers (Peng, 2011).

1.1.1. How to improve reproducibility?

Improving reproducibility is a multifaceted challenge requiring both behavioural and cultural change. The adoption of reproducibility practices has been sparked and embraced by the Open Science movement. However, a lot of researchers are not fully aware of the implications of reproducibility and how Open Science and reproducibility are connected and intertwined (Haven et al., 2022). To increase awareness and change research practices several steps should be taken

(Nosek, 2019). First, the infrastructure for the desired behaviour should be provided to make it possible. Second, the user interface and experience of the infrastructure should be improved to make the behaviour easy. Third, communities of practice should be fostered to make the behaviour visible and so increasingly normative. Fourth, incentives to enact the behaviour should be provided to make it rewarding. Last, policies should be enacted to make the behaviour required (Nosek, 2019). To further this work, we sought to explore the future of reproducibility for different stakeholders and question what should be the next steps for reproducibility and how diverse epistemic contexts can adopt reproducibility in different forms. In this deliverable, we aim to add nuance to the reproducibility debate through flexible investigation of diverse epistemic contexts (researchers from the field of machine learning and researchers working with qualitative methods), exploring the future of reproducibility through the lens of diverse research stakeholders – researchers, funders, and publishers.

In this context, we look to the future of reproducibility by exploring the preferred scenarios for multiple stakeholders, including how these scenarios can be realized. We reflect on the steps that are necessary for adherence to reproducibility-enabling practices and what different epistemic contexts need to make reproducibility a priority. Lastly, we reflect on what are the new problems that we may face when aiming to improve reproducibility. We believe exploring the possible futures for reproducibility is essential to discover the next steps for different members of the scientific community to take to realize the preferred future and the actions to avoid steering away from the dystopian futures.

We aim to highlight the essential role of institutions, funders and publishers in this endeavour to make reproducibility a priority by recognizing, rewarding, evaluating and monitoring reproducibility. Ultimately, we hope to steer and move forward the debate on reproducibility in the research community by addressing a set of core research questions related to how key stakeholders in the academic community envision the way in which matters of reproducibility should be addressed in the future. More specifically, it asks representatives from research funders, scholarly publishers, and researchers from diverse disciplinary backgrounds:

1. What are the preferred futures of reproducibility?
2. What are the alternative futures: dystopian, realistic, outlier?
3. What are the enablers and barriers on the way to the preferred future or reproducibility more generally?
4. Which stakeholder groups bear responsibility for addressing concerns related to reproducibility and have the ability (capability, motivation) to trigger change for reproducibility in the future?

2. Methods

In this study we employ futures studies. Futures studies is the systematic study of possible, probable and preferable futures (Inayatullah, 2013). It aims to explore and predict the future by envisioning alternative future scenarios, which then inform strategies for action to eventually shape the desired future at different levels. Futures studies can challenge current frameworks and commonplace assumptions by creating scenarios of the future (both desired and not) that can inform fundamental rethinking of the necessary steps required for their realization (in the case of desired futures) or avoidance (where not desired). Creating and testing potential scenarios of how reproducibility will evolve helps to prepare and respond to the challenges that will be faced in these futures.

2.1.Ethics and research integrity

2.1.1. Ethical approval

Ethical approval was obtained from the Ethical Review Board of the Faculty of Humanities at the Vrije Universiteit Amsterdam (2023.0351[1]). Prior to participation, participants were provided with an information letter and privacy policy statement. All participants provided consent via an online form.

2.1.2. Pre-registration

This study was pre-registered[2]. The aim of the pre-registration was to increase methodological rigour and integrity by highlighting strategies that can mitigate potential biases and methodological shortcomings and improve credibility of the findings. These strategies include: member checking of the results of the workshops, including different perspectives from different stakeholders, and reflexivity from participants and the research team.

2.1.3. Participant recruitment

For the workshop we recruited participants through a purposive sampling strategy based on the following inclusion criteria: people should be working in the specific stakeholder group targeted for the specific workshop (i.e., working in Machine Learning, qualitative social science, working as a funder, working at a publisher) and have demonstrable expertise and experience with issues of reproducibility in that specific stakeholder group. The latter refers to either studying reproducibility-related issues or contributing to projects or procedures that aim to improve reproducibility or to tackle related issues.

Participants were recruited through a three-fold recruitment strategy: (i) through existing connections and networks of the TIER2 consortium, (ii) through snowballing, i.e. by asking initially identified potential participants to identify others within their targeted stakeholder group that match the inclusion criteria, and (iii) through internet and literature searches to identify authors of papers, presenters of talks, or contributors to projects related to reproducibility issues within the four target communities. The latter builds on previous work within the TIER2 project. Participants were recruited for their expertise and knowledge on reproducibility and reproducible research.

2.1.4. Sample size and characteristics

Across the four workshops there were a total of nineteen participants. There were four representatives of funders, five qualitative researchers, five representative of publishers, and five Machine Learning researchers that participated. We had four participants representing organizations from England, three from the United States, three from the Netherlands, two from Norway, and one from Denmark, Ireland, Spain, Germany, Switzerland, Luxembourg, and Belgium.

2.2.Procedure

2.2.1. Pre-workshop survey

Prior to the workshops, all recruited participants were presented with information about our study and were asked to complete a pre-workshop survey. In this information package, we let the participants familiarize themselves with the topic of reproducibility, the goals of the workshop, and

[1] https://osf.io/b6ac2

[2] The pre-registration for the study can be found here: https://osf.io/87z29.

the exercises, and gave them a glimpse of the Miro board that served as the collaborative platform for the workshops. Through the survey, we collected demographic data (gender, academic rank, disciplinary field, main research methods used, age, expertise, institution, and country) and requested that participants started thinking about the present and potential future states of reproducibility.

The three main survey questions asked were:

1. How would you describe the current situation in relation to reproducibility within your field/professional context?

2. What in your view are the largest potential barriers in improving the situation of reproducibility?
3. What are 3/4 main topics that come to mind when you think of the future of reproducibility?

In this pre-workshop interaction with our participants, we also provided them with an informed consent form and information about our project.

2.2.2. Workshops

We conducted four online workshops. Each workshop took three and a half hours and was organized using Miro and hosted on Zoom. Each workshop had a short introduction to the TIER2 project and the goal of the workshop. Thereafter, participants were asked to do a short round of introductions to familiarise themselves with other participants. Next, participants were provided with a short summary of their pre-workshop survey answers and asked to validate and reflect on the four main topics that emerged from their pre-workshop answers. Three exercises were then undertaken by participants, which we will now detail[3].

2.2.2.1. Stakeholder mapping

The first exercise was a stakeholder mapping exercise (Appendix A1). On the Miro board, there was a target board and a list of pre-identified stakeholders: researchers, funders, institutions, publishers, institutional policy makers, governments, and the public. Participants were asked to identify the most important stakeholders in regards who has the most impact on reproducibility within research. The middle of the target board represented the stakeholder who has the most impact, and with each outer ring the potential impact of the identified stakeholder decreased. Participants used sticky notes on the Miro board to complete each task, which each having a colour assigned to their contributions.

2.2.2.2. Scenario planning

The second exercise was called scenario planning (Appendix A2), here participants created different scenarios for each of the main topics that were identified in the pre-workshop survey. The first scenario is the ideal. This describes the reproducibility world we want. The second scenario is the disowned. This is the world we reject or do not want to negotiate. The third is the integrated scenario where both the preferred and disowned are combined and gives a perspective that may be the most likely scenario. As a final scenario, participants also outlined the outlier scenario. This

[3] https://osf.io/wf2br

is a separate scenario that describes out-of-the-box ideas. Participants were stimulated to give as much details as possible for each theme on these scenarios on post-it notes on the Miro board.

2.2.2.3. Backcasting

The last exercise aimed to back cast the preferred scenario (Appendix A3). Working in pairs in breakout rooms, participants were asked to focus on the preferred future they defined in the previous exercise. Participants were instructed to describe a specific goal for the state of reproducibility in 2033. Keeping this goal in mind, participants had to identify potential enablers and barriers that may affect the realization of the goal. Next, enablers and barriers had to be attached to a timeline in the order participants believed they would occur. Participants were instructed to work backwards from the future goal towards the present. At the end of the exercise, each pair was asked to describe their goal and back cast the main enablers for and barriers to it.

At the end of each workshop, we asked participants to reflect on and evaluate the workshop and the insights that emerged from the exercises. First, participants were asked to evaluate the preferred scenario and backcasting timeline. Last, participants were asked about their opinion on the workshop, particularly what they found most useful from the discussion and any suggestions they have for future workshops.

2.2.3. Validation workshop

Following the initial four workshops, we conducted a validation workshop which was conducted after preliminary analysis as described below in 2.4. We invited all participants from the four stakeholder groups. There were nine attendees: two qualitative researchers, one Machine Learning researcher, one representative from funders, and five representatives from publishers. During the validation workshop, we invited the participants to reflect on the combined preliminary findings from all the workshops. Using Miro, we presented the results for every workshop exercise. Following brief presentations of these results, participants were asked to reflect on them in terms of the degree to which they recognised the results from their own workshop, whether they had expected these results and whether there was anything they felt to be missing from them. We also asked them whether we correctly understood the results and asked them to reflect on implications for the future of reproducibility. This information was used to validate our findings, reflect on the results and address potential missing points. Finally, we gave the participants the chance to reflect on the study in its entirety and solicited any feedback they might have for future iterations of our study.

2.3.Qualitative data analysis

For data analysis, we included all the materials produced by participants during the workshops (i.e., written content on post-it notes and any other contributions to the Miro board), their responses to the pre-workshop survey, field notes of the co-facilitators, and the anonymized transcripts of the workshop recordings. Using NVivo, we first engaged in flexible coding as described by Deterding and Waters (2018) to organize our data for collaborative coding. Then, we created a deductive code structure (Appendix B) based on the research questions, topics of the focus groups, and a broad understanding of the themes that arose during the workshops. Each member of the team was then responsible for coding an assigned section of the data (portion of a workshop). Team members kept detailed notes to document their data analysis process and were able to create additional codes as they worked. Each member of the team worked on an

independent copy of the project offline to enable asynchronous collaboration, after which the separate projects were merged into a master project with all coded data present.

Following this process, we held an internal meeting to triangulate our analytic process and then moved onto an analysis phase focused on answering the research questions. Each team member was assigned a portion of data to analyse and write up. Through the process of analysis and writing, we continued to triangulate through group meetings and collaborative drafting of this document.

3. Results

In the following section, we aim to answer our research questions as we have posed them in the introduction and methods section. We discuss the most influential stakeholders in shaping the future of reproducibility, following the most important themes that describe the preferred, dystopian, realistic, and outlier future. Lastly, we delve into the enabler and barriers towards the described preferred future.

3.1. Stakeholder mapping

Almost all participants reported that researchers are the most influential stakeholders shaping the future of reproducibility as evidenced by them placing themselves at the centre in our stakeholder mapping exercise. They agreed that what happens in terms of reproducibility is ultimately up to researchers, because it is they who manifest cultures of research (that either do or do not support values and practices aligned with reproducibility). One participant pointed out that researchers are central because they play a variety of roles that can influence the development of reproducibility, including reviewers and collaborators. Other participants added that professional societies, domain-specific committees and international committees are relevant stakeholders, which are also typically composed of researchers. As one participant from the funders workshop put it (P3), "if researchers don't engage with reproducibility, then then we will achieve nothing."

Almost all participants agreed that funders are the second most influential stakeholder group, placing them with near unanimity in the second layer of influence (except for our Machine Learning (ML) researchers, who placed them in the centre, alongside themselves, and for funders, who placed themselves slightly outside of other secondary stakeholder groups), and they mostly agreed that publishers have the same level of influence. There was consensus among them that these are influential because it is their policies, standards, and practices (of assessing, rewarding and recognising research) that have the power to influence the development of research cultures that include reproducibility (we provide more details on this in our findings on enablers of reproducibility). Participants believed that funders have a lot of influence on researchers, because researchers will comply with their demands to get funded. As P3 said in the funders workshop, funders (and publishers) "can do a lot in changing the culture to foster reproducibility". And, while funders were considered less influential than researchers during this exercise, we observed during our validation workshop, when presented the results of this exercise, that some participants asserted that funders should be considered equally influential (to researchers) because of their power to determine what researchers due through funding.

Other stakeholder groups which were largely considered to be secondarily influential, alongside funders, include institutions and institutional policymakers. Participants recognised that (research) institutions are influential in the same way as funders and publishers, due to their role as assessors of research value, as well as their role in providing education and training. Additionally, some participants added stakeholder groups to this level that we had not provided. These included infrastructure and service providers, libraries, research integrity offices, multi-stakeholder organisations, students and educators.

Participants viewed governments as less influential, mostly placing them in the second layer removed from the centre (though a few placed them in the first layer and a few in the third). Yet, as a participant from the funder workshop pointed out (P1), governments (when making science policy) set key performance indicators (KPIs) for funders and other stakeholders, and therefore they play an important role in the pipeline of creating incentives for reproducible research. In their own words, "If the government say institutions, funders, we want you to focus on reproducibility and this is what we value, then people will do it."

Our participants agreed that the public is the least influential stakeholder, with nearly all placing them in the third layer removed from the centre. As one participant from our workshop with qualitative researchers and experts put it (P5), it's not that the public don't matter, but that, in an ideal world of reproducible research, they shouldn't have to even consider it.

These findings from our stakeholder mapping exercises are reflected in the discussions that took place across all four workshops. Our results showed that researchers are the stakeholder group that were mentioned most frequently in discussions (116 coded mentions), along with funders (86 coded mentions) and publishers (79 coded mentions). Others, considered less influential in the stakeholder mapping process, received far fewer mentions.

Finally, we note that our participants shared additional stakeholders of relevance during our validation workshop. When asked whether they believed any were missing from the results of the stakeholder mapping exercise, they answered that they would also include private companies (that both use and fund research), NGOs/agencies (as users of research), and professionals like medical doctors (as users of research), and research support staff as relevant stakeholders.

3.2.The preferred futures of reproducibility

The four tables below represent the four preferred futures for the four different stakeholders that were discussed in the four workshops. Each table has four themes for which stakeholders described specific elements of their preferred futures as related to the themes they agreed on the beginning of the workshop. Each theme has excerpts from participants discussing the specifics of their preferred futures, these can be found in Tables 1 through 4.

When discussing their preferred futures participants highlighted that they see them as dynamic and forward-looking. They believe the themes identified below are the core elements that will shape the future of reproducibility. However, once these are realised, new issues will arise and will need to be addressed, as it is a dynamic process. Additionally, it is essential to know and imagine the discussed futures as a cycle (this was also highlighted in the validation workshop by

several participants) because the themes are dependent on each other when one comes to fruition then we can see what is missing, it creates a network effect.

Table 1. A summary of the preferred scenarios for the publishers for each domain identified in the workshop.

Incentives and Costs	• *"A key aspect of every grant application is evidence that the research will be reproducible and builds on prior results"* • *"There are scalable practices for checking at a reasonable cost"* • *"Articles published based on quality of research rather than impact"* • *"Funding is contingent on adherence to standards and processes"* • *"Preregistration mandated for all empirical research"*
Clearer Definitions and Expectations	• *"Journals adopt community-based standards"* • *"Definitions are understood and adopted by research disciplines"* • *"Readers are clear on exactly the standards and practices the researcher has followed, and what the publisher has checked"* • *"Shared standards and definitions across domains and stakeholders of what improves reproducible research."* • *"Professional societies set domain-specific expectations and standards"*
Guidance and Training	• *"Publication reproducibility requirements part of mentoring"* • *"Making research reproducible is baked into undergraduate and graduate research training programs"* • *"Key messages and practices based on community endorsed standards can be amplified by publishers"* • *"Institutions emphasise the importance of training - mandatory rather than expected/encouraged"*
Trust in Science	• *"Conclusions and policies are seen to be based solidly, solely and reliably on data and evidence"* • *"Clear standards to certify that research has been done by evaluated/validated researchers (to prevent paper mills and fraud)"* • *"All scientific claims can be traced to evidence"* • *"Governments reinforce trust through evidence-based decision-making and direction of policy, and communicate this clearly to the public"* • *"Provenance of material and data are clear"*

Table 2. A summary of the preferred scenarios for the funders for each domain identified in the workshop.

Recognition and rewards	• *"Reproducibility is recognized and rewarded as one of a broader set of OS practices in grant awarding procedures by Funders"* • *"Rewarding alternative research outputs"*

	• *"Calls for replication studies"* • *"Reproducibility (contextualised within different disciplines) is globally recognized as a core element of scientific quality/excellence"*
Training and mentorship	• *"Embedded in graduate programs"* • *"Reproducibility courses, and/or integrated into other training"* • *"Special calls or funds for reproducibility training"* • *"Reproducibility as collaboration opportunity"*
Awareness and communication	• *"There are vibrant national reproducibility networks in all countries across Europe, working together closely"* • *"Conferences on reproducibility, and what can't be reproduced"* • *"Institutions and funders integrate reproducibility awareness and needs in all documentation"* • *"Reproducibility is the 'new normal'"* • *"Awards or showcasing for/of best practices"*
Reproducibility practices	• *"Reproducible-by-design is mainstreamed"* • *"Clear guidelines and policies from funders"* • *"Groups just focused on reproducing other's work"*

Table 3. A summary of the preferred scenarios for the Machine Learning researchers for each domain identified in the workshop.

Incentives	• *"Probability of being accepted is much lower (or zero) when research is not reproducible unless properly justified, public funders also set high levels of reproducible research to ensure its quality and future validation"* • *"Clear transparent visibility of if something is indeed reproducible"* • *"Journals and conferences require that all resources (data, code, etc) used to create research are submitted alongside papers and assess completeness as part of reviewing process"* • *"Funders and publishers require rigor"*
Guidelines and data standards	• *"It is clear what reproducibility means and what is expected from a paper to report, both in terms of data processing and algorithm tuning"* • *"There is a body of standard resources (guidance, scripts, vocabulary, training, etc) available that are used widely"* • *"We are clear on whether research is exploratory or confirmatory"* • *"Methodology is standardized"*
Resources and costs	• *"Hosting code or datasets should not rely on researchers or projects, but have a long-term viability, perhaps based on public funding / solutions"* • *"Funders support reproducibility work via funding earmarked for reproducibility assurance in research projects - an additional bit of funding that can be used only for reproducibility work in the project (carrot, rather than stick)"* • *"Standard shared resources (guidance, scripts, vocabulary, training, etc) offset costs"*

Culture	• *"Reproducibility, best practice software engineering (including code review), etc are perceived as part and parcel of researcher training"* • *"Rigorous research is expected"* • *"Include engineer or developer in every research lab"* • *"Enforce collaborations within group and with external groups"* • *"As a conference, running a reproducibility session should make sure that accepted papers are reproducible"*

Table 4. A summary of the preferred scenarios for the qualitative researchers for each domain identified in the workshop.

Costs and Incentives	• *"Storage of large transcript files becomes free of charge"* • *"Costs are lowered by having very few go-to repositories"* • *"Main incentives for reproducibility are intrinsic to research community, funders & publisher support & re-enforce"* • *"Sharing and re-use of data is actually rewarded"* • *"Clear incentives for sharing different outputs (data, derived data, analysis workflows)"*
Terminological plurality and diversity	• *"Different schools of qual have well-established (diverse & suitable) standards for repro/transparency"* • *"Reproducibility is no longer a daunting term but something that is constructively engaged with"* • *"Reproducibility is defined in ways that are meaningful to qualitative research. It does not become a barrier to progression with qual methods"* • *"Reproducibility is considered as a norm related to scientific progress rather than pointing out who is wrong"*
Data and Information Sharing	• *"Easy to understand metadata standards are implemented widely"* • *"Data are widely shared and re/used, using access controls where needed"* • *"Clear guidelines for anonymized and pseudo anonymized qualitative data sharing"* • *"University teachers actively have students engage with existing qualitative data (e.g., interviews) to promote secondary use and searching"*
Infrastructure and methodological rigor for qualitative research	• *"Qualitative methodologists and helpdesks are available everywhere"* • *"Qualitative data analysis tools become open source; allowing for easier exchange"* • *"Workflows are not rigid"* • *"Infrastructure (tools, formats, repositories) are widely known, used, open and affordable"* • *"Standards on how to share raw data (format)"*

Below is a summary of the main overlapping themes that have emerged from all four stakeholders' preferred futures.

3.2.1. Culture

Quality over quantity: Across the research community quality for research is emphasized over quantity. Thus, publications are assessed for the quality of the procedure, Open Science practices, and reproducible practices not only whether the results are significant. In relation to a shift in understanding of quality is shown where if research is not reproducible it does not reduce the quality of it because the procedure and knowledge gathered leads to further explorations and adds on to previous literature and knowledge.

Reproducibility is the new normal: Reproducible practices are ingrained and incorporated into the daily life of researchers. Research becomes reproducible by design and streamlined into workflows. This shift illustrates an overall acceptance of reproducibility's importance for research.

New researchers: Researchers graduating for their master's or PhD candidates need to complete training in reproducibility as part of their studies.

> *In the future, I think research assessment should be based on, you know, the quality of the research rather than the impact. And, you know, it should be judged on, you know, is the methodology good? Has there been pre-registration? Things that actually help improve trust in science. (Publisher, Scenario Planning)*

> *I was thinking, in the future, the perfect future is all research reproducible and I don't think so that is not a good future. I think it's a terrible risk-averse future. And as an exploratory agent, you need to explore and you need to throw out ideas. And often what seems like you have an idea that it's poor, it showed poor results. But then someone reads this idea and they build on it and then they learn from it because it has some qualities. It's just your implementation of that idea that didn't really work, but someone else took it and built on it. (Machine Learning Researcher, Scenario Planning)*

> *We want to, to see reproducibility being recognized as a core element of scientific quality. (Funder, Scenario Planning)*

> *And yeah, my idea would be that reproducibility, as a research practice should be embedded in, in graduate programs." (Funder, Scenario Planning)*

3.2.2. Definitions and standardization

Solid definition: There is a solid definition for reproducibility and reproducible research that is accepted and utilized by reviewers, researchers, publishers, and funders. There is a clear difference in the definition of replicability and reproducibility. However, there is enough adaptability in the definition for different disciplines and methodologies to adapt.

Minimum requirements: Publishers set a minimum standard for reproducibility.

Standardized methods: Standardized and shareable methods and statistical software are available. Detailed instructions on how to use the programs are shared with researchers.

Furthermore, standardisation does not refer to checklists but rather clear guidelines to account for the flexibility that may be needed for different methodologies and disciplines.

> *The first thing we will do is to make sure we know what the definition of reproducibility is. (Machine Learning Researcher)*

> *I explicitly mentioned that both reviewers and authors should agree on when a paper is reproducible. And yeah, the conferences and journals should also expect the same standards for, for everyone. (Machine Learning Researchers, Scenario Planning)*

> *I imagine that there will be standardized methods, workflows that are shareable and standardized. For example, instructions on how to share step by step qualitative analysis, because that's also really important for reproducibility and qualitative researchers just don't know how to share analysis. (Qualitative Researcher, Scenario Planning)*

> *But coming back to the definition of reproducibility, what we personally or what we use here is like computational reproducibility. That means there is a depth to the reproducibility. And the question is like what is covered by reproducibility? And my feeling is that reproducibility basically refers to deterministic reproducibility of the data analysis. So that means that in the most naive case there should be same data, same method. The same implementation gives the same result and I think that has to be decoupled from other types of reproducibility. And that seems to be relevant for at least our fields. And that's like how we typically present it. (Machine Learning Researcher, Scenario Planning)*

3.2.3. Incentives

Recognition and visibility: Researchers reproducible actions and practices are made visible.
Alternative outputs: Researchers are rewarded and recognized for alternative research outputs, for example datasets or software, not just for scientific publications.

Opportunity for collaboration: Increased possibility for collaboration between researchers. Additionally, reproducibility networks start to form when researchers engage in reproducibility and Open Science practices.

Funding schemas: Extra funding is allocated for reproducible and Open Science practices; researchers no longer need to apply for extra funding. Extra funding streams continue to encourage researchers to maintain the momentum for reproducibility in research.

> *Reproducibility can be seen as collaboration, right? Because it's an opportunity for you to go to another group somewhere else in the world and say, Hey, let's work together. Had these findings. Can you check this out for us? Right so pushing collaboration as one of these values and, you know, and saying reproducibility is a way to foster that. (Funder, Scenario Planning)*

> *Having it recognizable so that reproducible, reproducible, reproducible studies or activities in reproducibility can just be put on the CV. (Funder, Scenario Planning)*

3.2.4. Infrastructure

Guidelines: Funders will have clearer guidelines and policies for what researchers need to comply with.

Tools: Various tools will be available for researchers to aid reproducibility efforts. These tools will be available in the long-term and open source, facilitating the exchange of data.

> *So, there's lots of Open Science policies out there and some of them are detailed and, and yeah, just have a lot of others, a lot of hot air. So, it's really like it should be a bit more clear. And really what, what should the researcher then do to kind of comply with these Open Science policies or reproducibility policies that the funder writes down? (Funder, Scenario Planning)*

Whilst there were significant similarities between the stakeholders' preferred futures, some differences arose. These differences pertained to the variances in the definition of reproducibility. The most notable difference was between qualitative researcher and the other stakeholders. Qualitative researchers want to emphasize and assure there is enough variability for qualitative research methods. This does not only pertain to the definition of reproducibility but also to requirements, as qualitative researchers may not be able to share data in the same ways quantitative researchers can. They advocated for the importance of understanding that “one size fits all” is not part of their preferred futures. Machine learning researchers discussed similar concerns for computational reproducibility.

4. Alternative futures of reproducibility

4.1.Dystopian future

Interestingly, participants while discussing their undesired and dystopian futures of reproducibility indicated that their thought process behind it was that they thought of the opposite of their preferred or ideal futures. A funder mentioned that their dystopian future is reflective of the current state of reproducibility. We next discuss themes that participants brought that would be their dystopian scenario.

4.1.1. Culture

Systematic culture: Reproducibility and reproducibility practices are completely ignored because it becomes too difficult and rigorous for researchers to engage with it.

Incentive driven research: Researchers conduct research for their own gain, losing motivation to contribute to the wider scientific community with knowledge, but only do so to get as many papers as possible to get published. In these circumstances if researchers do follow reproducibility guidelines, they do so only to follow a trend. Furthermore, research will be assessed according to old incentives discouraging researchers to learn and keep engaging with Open Science and reproducibility.

Quantitative and qualitative differences: Quantitative Open Science advocates expect the same reproducible practices from qualitative researchers, resulting in irresponsible data sharing putting participants privacy and trust at risk.

Training and mentorship: Supervisors and institutions push and encourage their students to engage in fraud and questionable research practices to get the results that are expected to ensure future funding.

Publishing: Less papers are being published due to fear of new reproducibility guidelines and requirements.

> *We call chasing ambulances when someone is publishing simply to follow a trend without really being, you know, very careful about doing a robust analysis. And for me, this is the main worry. (Funder, Scenario Planning)*

> *My worry would be that people are doing it because they have to and not because they're really there's any meaning for it, you know that it's not meaningful. Um, that feels like just a shallow, empty future that people are doing something they have to do to jump through a hoop, to tick a box, and not because they necessarily ethically wouldn't be able to buy into it if it develops in a way that, that it could. Because I think ethically, there's so much about open research, Open Science broadly that people really buy into. But if it loses its meaning, it loses its purpose and it just becomes a, you know, a stick to beat people with that. Yeah, they may do it because they have to. And then what will happen is they'll use very limited methods, they'll use very limited approaches and perspectives and just churn out stuff that doesn't really add anything meaningful to society, but they'll check the right boxes. So that's the stuff they'll do. So that would be my that would be my kind of dystopian future. (Qualitative Researcher, Scenario Planning)*

4.1.2. Loss of trust in science

Failure to produce results: Due to promises for bigger and flashier findings by the research community and a lack of results, particularly reproducible results, the public loses trust in science.

Geopolitics: The Open Science agenda is put into question as a result of political debates.

Fear of AI: The influence of AI on research is significant, with increased production of fraudulent content it becomes impossible to make any progress in research.

Decreased participation in research: Members of the public fear for their privacy being undermined due to mishandled data and are less likely to engage and participate in research.

> *A scenario would be that the broader public would lose its trust in science. Of course, it would be most disastrous outcome of all this. (Funder, Scenario Planning)*

> *It's all just shiny, flashy, new things which causes a loss of trust and loss of faith in science. And then that kind of destroys the culture and destroys the, the, the system kind of because if nothing is if nothing that we do gets anywhere, you know, then people the public will lose faith in us. (Funder, Scenario Planning)*

4.1.3. Definitions and standardization

Lack of definition: Dystopian futures describe a lack of a clear and solid definition for reproducibility and reproducible research. Furthermore, there is no movement towards the direction of a definition and the research community will be having the same conversations.

Lack of standardization: Next to the fear of a lack of definitions, the fear of too many standards represents the dystopian futures. Too many standards will make it impossible for researchers to fulfil them, particularly for groups who have less funding or less researchers.

> *Like we could still be trying to define reproducibility, we could still be trying to find what works. So yeah, I think that's worse than having a position and it not working is just there's still no position, there's still people having projects and meetings trying to make some definitions to take some steps forward. Yeah, I think that would be the worst thing. (Publisher, Scenario Planning)*

> *So one is that we may have too many standards, right? So that nobody knows what is a standard anymore. And in particular, we may have some guidelines or standards or whatever we want to call it, that are biased. (Machine Learning Researcher, Scenario Planning)*

4.1.4. Infrastructure

Guidelines and tools: The introduction of new tools increase the bias for researchers as they may only be available to institutions with financial resources. On top of the limited accessibility of tools the tools are misused because researchers dump their data on online repositories without properly handling and managing the data. Furthermore, the data will not be accessible for other researchers particularly to reuse that data.

> *Barriers start to be so sort of difficult to overcome that we'll stop sharing data from qualitative research or data will be shared carelessly, right, without, without thinking about it. So, this sort of dystopian scenario can go either way. (Qualitative Researcher, Scenario Planning)*

4.2.Realistic future

We also asked the participants to picture the most likely/realistic scenario for the future of reproducibility. When discussing their likely future of reproducibility, some participants showed quite some cynicism towards the future if nothing would change. This was a slightly different

interpretation of the exercise because likely is not something static and can change for the better (or worse) over time.

In the discussions, a couple of possible scenarios were identified and highlighted. The following sections were presented in detail; culture, infrastructures, incentives and recognition, funding reproducibility.

4.2.1. Culture

Awareness and Communication: There is a growing awareness of the need for reproducibility in research. Efforts should focus on framing reproducibility in a positive context and making it a norm rather than an exception. Reproducibility is considered crucial in areas where it is perceived as a significant problem.

Leadership and Culture Change: Leadership is essential for driving the changes required to improve research reproducibility. It may take a generation to change the scientific culture fully.

Challenges and Acknowledgment: Acknowledging that some research may be impossible to reproduce due to the nature of science. Encouraging researchers to report their efforts in reproducing research.

Incorporating Reproducibility: Reproducibility is increasingly becoming part of the Open Science and Research Integrity discourse. Funders can play a role in highlighting reproduced research on their websites.

Normalization of Reproducibility: The emphasis is on normalizing the term "reproducibility" in policies, documents, and guidance. Integrating the concept of reproducibility into research practices and quality assessments.

Collaboration and Future Trends: The importance of reviewing articles accurately and including solid science from diverse sources for reproducibility and knowledge access.

> *And culture is really hard to change. Really, really hard to change. It's the reason that you have this, uh, uh, this saying about scientists dying day by day. And so, in a generation you might have sort of change the culture properly and fully. (Machine Learning Researcher, Scenario Planning)*

> *And so, the awareness and critical mass around reproducibility will hopefully continue and increase until something bigger happens. (Funder, Scenario Planning)*

> *So, people see that it's important that it becomes kind of normalized in terms of policies and documents and guidance and all that sort of stuff. So, people like maybe aren't thinking, I need to focus on reproducibility, but they have it in the back of their head as, oh, this is not bad. This is something we should think about. (Funder, Scenario Planning)*

4.2.2. Infrastructures (including training, guidelines, tools)

Participants highlighted that in 10 years' time it is realistic an infrastructure that better supports more reproducible methods and results.

Training and Network Building: Building a network of data stewards is essential for better support infrastructure at institutions. Conferences and webinars on reproducible research are being co-funded by reproducibility networks. The idea of launching a journal focused on realistic scenarios is considered feasible. There is a need for institutional training to enhance reproducibility. Growing interest exists in providing specific training for PhDs and early career researchers. Embedding the importance of reproducibility in training programs, including PhD training, is essential.

Impact of AI. AI is expected to provide automated transcripts, making research more cost-effective but also potentially enabling the fabrication of qualitative data. The future of AI and automation is expected to impact research by automating tasks such as transcriptions and coding. Automated coding through AI could lead to a decline in the quality of qualitative research. There is a need to consider guidelines for anonymization due to AI's potential to easily pseudo-anonymize data.

Reproducibility Guidelines. Researchers often find existing research to be "messy" and lacking in transparency. Clear policies on open science and reproducibility are needed, including specifying what funders expect to see in the research they fund. Reproducibility is not accessible to everyone due to resource and cost limitations.

Role of Publishers and Institutions: Publishers should reinforce standards, journal policies, and workflows to improve reporting and verifiability. Professional societies, governments, and industry should play a role in setting norms and standards.

> *We will have very good automated transcripts through AI, will make things cheaper and more feasible – it will become easier to fake qualitative data, especially as it becomes more valued. (Qualitative Researcher, Scenario Planning)*

> *Put that reproducibility for now is not for everyone. Not everyone is doing reproducibility and not everyone can do it because. Of resources and costs. (Machine Learning Researcher, Scenario Planning)*

> *I think we can have better education of trainees around the idea that this is what is expected. If you work in this field, this is what you have to do. This is the accepted norm. (Publisher, Scenario Planning)*

> *And training and mentorship, just embedding of, of, you know, importance of reproducibility across training. So definitely PhD trainings, but also potentially in other levels. (Funder, Scenario Planning)*

4.2.3. Incentives and recognition

Research Assessment Reform: Research assessment is a crucial area for reform to safeguard reproducibility. The current direction of reform in research assessment does not explicitly prioritize reproducibility, although it is not excluded. Incentives for reproducibility practices in qualitative research may become more common.

Program Outcomes and Reproducibility: It is important for research programs to clearly state their expected outcomes, also in terms of reproducibility. Currently, reproducibility is not highlighted as a component in terms of research efficiency and quality.

Role of Funders: Funders have the power to change the incentive system in research and demand reproducibility practices from funders. There is optimism about funders' potential to influence positive changes, particularly in the context of Research Integrity (RI). More funders may follow the lead of organizations like the NIH in mandating data and materials availability, as a step toward reproducibility promotion.

Incentive Structures: While some actions, such as reporting standards and education, can be implemented on a discipline-specific scale, changing incentive structures is seen as more challenging. The emphasis is on recognizing that one size does not fit all in terms of data and analysis sharing.

Realistic future Scenarios: The expectation is that research with reproducibility components will receive more citations, leading to greater value for the authors and institutions. Over the next ten years, there may be a shift away from quantitative metrics toward qualitative peer review in assessing research impact.

Leverage as Publishers: Publishers may have leverage to require more reproducible reporting in published research.

> *Funders set the roles. It is in their power to change the incentive system. So, they need to act. (Funder, Scenario Planning)*
>
> *I think we actually have the power to change incentive structures because of the threat to RI. Things are actually changing and as publishers we are a well-placed stakeholder. (Funder, Scenario Planning)*
>
> *But if we worry about bad incentives in academia making people take shortcuts and risks and cheating, then I think I'm pessimistic about that changing. (Funders, Scenario Planning)*
>
> *Probably some reproducibility practices in qualitative research will start getting incentivized. (Qualitative Researcher, Scenario Planning)*

4.2.4. Funding reproducibility

Funding for Metaresearch: There is a call for more funding to support metaresearch, which involves research on research itself. This additional funding can be instrumental in addressing issues related to research quality, reproducibility, and integrity.

Data Stewards and Support Infrastructure: The Dutch Research Council is mentioned as an example of funding for data stewards. Funding for data stewards can contribute to the development of a more robust support infrastructure for reproducibility at the institutional level.

Program Funding and Expected Results: Funders often support research programs and look for specific outcomes from these programs. Consideration is given to what results a research program should yield, including factors related to reproducibility.

Other Job Titles: Besides data stewards, it was suggested that new job titles may emerge in the next ten years that would require funding. These positions would play a role in supporting research quality, integrity, and reproducibility.

Impact on Young Researchers: It was acknowledged that a lack of support for reproducibility and research quality can have a detrimental impact on young researchers. The challenges they face in attempting to replicate existing research can lead to frustration and may affect their research progress and well-being.

> *"More funding for metaresearch" (Qualitative Researcher, Scenario Planning)*
>
> *"I also think so I'm very like very much like this Note on funding for data stewards, something that's the Dutch Research Council is doing at the moment, and I think we will continue to do that. That will also lead to a much better support infrastructure for reproducibility on an institutional level." (Funder, Scenario Planning)*

4.2.5. Definitions and standardization

The issues around reproducibility will continuously evolve. New tools and practices (partially AI) driven will emerge and standards will improve. Possibly, free data sources will be no longer free (think of google/twitter etc).

4.3.Outlier Scenario

4.3.1. Culture

Collaboration: Researchers emphasize the importance of collaborative work and constant debate to ensure the robustness of research claims. This collaborative approach is considered essential for maintaining high research standards. Collaborative efforts are made to engage different groups of people in other parts of the worlds in reanalysing data to check if their conclusions differ. This approach aims to promote diverse perspectives on research. The academic landscape is compared to an industrial revolution with large collaborative teams, aiming to shift from a feudal system with hyper-competition.

Anonymous Reporting for Non-Reproducible Research: Suggestions are made for implementing anonymous tip lines or ombudsmen to report instances of non-reproducible research, creating a mechanism to address unethical practices.

Data Stewards and Skills: The role of data stewards, individuals with the skills and knowledge to manage research data effectively, is highlighted as crucial for ensuring data reproducibility.

Public Involvement: The conversation explores the idea that the public and society should play a more significant role in holding researchers accountable for reproducible research, rather than funders or publishers.

Global Inclusivity: Efforts are discussed to open discussions on open and reproducible research to researchers in Africa and the global South, promoting inclusivity in addressing research reproducibility.

> *And then on the more negative side, an anonymous tip line or ombudsman for, you know, people doing purposefully non-reproducible research to be to be called out this is hard. (Funder, Scenario Planning)*

> *So, academia looks more like a sort of it's gone through an industrial revolution with large factories of people rather than its current feudal system, which is driving the hyper competition. (Publisher, Scenario Planning)*

> *We have tried to get different groups of people to competitively reanalyse data to see if they come to different conclusions. (Qualitative Researcher, Scenario Planning)*

4.3.2. Definitions and standardization

The conversation emphasizes the optimistic view of AI's role in research, particularly in generating hypotheses and promoting collaboration between AI and researchers. It also underscores the importance of early education on research ethics and integrity. Finally, the potential of AI in qualitative data analysis and its impact on research outcomes is highlighted.

Embracing AI for Reproducibility: The idea of embracing AI as a utopian future is highlighted, indicating a positive view of AI's potential in research. AI can play a significant role in generating hypotheses from existing research, aiding researchers in their work. AI tools are envisioned as powerful resources for suggesting research hypotheses. AI can assist researchers in various aspects of their work, demonstrating the potential for a symbiotic relationship. AI can enable the analysis of qualitative data in various ways, leading to diverse insights and conclusions.

Advocacy for Reproducibility: Publishers are recognized as key players in advocating for reproducibility and research ethics. The suggestion is made that advocating for research integrity should start early in the education curriculum to instil these values in future researchers.

> *Like from the public demanding reproducible research because we make them literate enough to understand what this is about. (Funder, Scenario Planning)*

> *I put Robot Labs, so it's really the idea that, um, hypothesis can be generated from known research and hypotheses can be suggested using AI tools. (Publisher, Scenario Planning)*

4.3.3. Infrastructures

Mandatory Training for Principal Investigators (PIs): The idea of implementing mandatory training for PIs is mentioned as a potential solution to enhance research reproducibility. Challenges to this idea are discussed, including PIs' lack of interest or perceived need for such training.

Offices for Research Integrity: The importance and funding of offices for research integrity are brought up, highlighting their role in promoting ethical research practices.

Guidelines and Tools: The need for clear guidelines and tools to support reproducibility efforts is emphasized.

Initiatives in Publishing: The conversation mentions initiatives like eLife and PLoS that are willing to experiment with different approaches to research publication, potentially encouraging reproducibility. Publishers are considered as having a role in education by promoting research ethics and integrity at an earlier stage in the academic curriculum.

Data Sharing and Sustainability: A scenario is presented where everything is shared, addressing reproducibility issues, but it raises concerns about the sustainability of data storage, particularly related to energy consumption. The discussion highlights efforts to engage different groups of people in the reanalysis of data competitively, promoting diverse perspectives on the same dataset.

The Mars Reproducibility Network: The concept of a "Mars Reproducibility Network" is introduced, indicating a growing interest in reproducibility initiatives.

Technology and the Future of Research: Ideas about advanced technologies, including AI and brain-computer interfaces, are presented as potential tools to improve transparency and data sharing. The idea of tapping into researchers' brains to understand their decisions and workflows, essentially creating a "brain dump appendix" for research methods, is discussed.

> *That if somebody applies for funding at a funding institution, then they have to get trained (in reproducibility red. (Funder, Scenario Planning)*
>
> *Maybe we should work on the Mars Reproducibility network first. (Funder, Scenario Planning)*
>
> *So, you have labs that are that are connected to these hypothesis generation engines and that where the tools themselves in the Internet of Things type of way report back on what was done. Um, and because all the standards are so good, you really have emerging features of research that that can build on itself. Um, following from, from these types of labs. So that's, that's my utopia that I put in there in a bunch of stickers. (Publisher, Scenario Planning)*
>
> *And there have been many situations where I thought if only, I could sort of duplicate myself and do some of these things with another person, that'd be so great. (Qualitative Researcher, Scenario Planning)*

4.3.4. Incentives and recognition

The conversation highlights a shift toward more inclusive, transparent, and accountable research practices, with a strong emphasis on involving the public and addressing issues of reproducibility at various levels. Below you find some ideas how to shape this in the future:

Public Demand for Reproducible Research: The public becomes science-literate and actively demands reproducible research, recognizing the significance of robust scientific findings. The public is envisioned as the primary driver of reproducibility efforts, as they are the ultimate beneficiaries of scientific research.

Establishment of a Journal or Chair for Reproducible Research: Suggestions include creating a dedicated journal for reproduced research or establishing a funded chair in reproducibility. These initiatives aim to promote and reward reproducible research.

Nobel Prize for Reproducibility: The concept of a Nobel Prize for Reproducibility is introduced, recognizing collaborative teams for their commitment to producing reproducible research.

United Nations Recognition: The United Nations acknowledges the importance of reproducibility, creating a prize for teams that excel in reproducible research. This recognition includes permanent contracts, prestige, and funding for their work.

Hotline for Reporting Bad Practice (Ombudsman): A hotline or ombudsman is proposed to report unethical or non-reproducible research practices, providing a channel for addressing research misconduct.

Reform Funding practices: Proposals include grant funding specifically allocated for reproducibility efforts, ensuring that research projects have dedicated personnel addressing reproducibility issues. The idea of defunding research groups that engage in questionable practices is discussed, with the possibility of reclaiming funds allocated for such research.

Shifting Focus Away from Research as a Career Progression Tool: The conversation highlights the need to shift the focus of research away from being a primary tool for career progression, emphasizing quality and reproducibility over quantity.

Assignment of Reanalysers: Researchers can be awarded or assigned a "reanalyzer," an individual responsible for independently verifying and checking the reproducibility of their work.

> *Nobel Prize for Reproducibility. Okay, so. What does that look like? (Funder, Scenario Planning)*
>
> *Defunding, asking the money back from research groups who are following questionable research practices. (Funder, Scenario Planning)*
>
> *So, you could imagine a radically different setup for at least science. And therefore, papers would come to reflect that and they would look rather different. They would they would be a genuine report of something that had been done*

rather than something that was used primarily as a career currency. (Publisher, Scenario Planning)

5. The enablers and barriers towards reproducibility

Factors that have the potential to contribute to higher levels of reproducibility or that are considered barriers towards preferred future scenarios of reproducibility were intensively discussed during all workshops. In this section, we will answer the sub question 'What are the enablers and barriers on the way to the preferred future or reproducibility more generally?'.

Table 5 presents an overview of the results, outlining the most important and most frequently discussed barriers and enablers towards preferred future states of reproducibility. In brief, the findings we report in this subsection reveal that cultural and social aspects, alongside technological and infrastructural ones, are thought to be the most important factors facilitating or hindering both the preferred future and reproducibility more generally. Research culture is conceived of the space where reproducibility is understood and implemented, as well as the space that must be influenced by other facilitators for reproducibility to become mainstream, normative, and standardised (as much as possible). Meanwhile, technological and infrastructural aspects are understood to be the tools, systems and platforms with and within which researchers can implement reproducibility.

To achieve reproducibility in both realms, participants identified training and education (for researchers and all other stakeholders involved in support and assessment), systemic and policy-related factors (to drive and enforce best practices), and financial or economic factors (in terms of funding for reproducible research and for the work and infrastructures that support it) as important factors. In the subsections below, we discuss the barriers and enablers associated with these dimensions in more detail.

Table 5. Main dimensions considered to affect the journey towards preferred futures of reproducibility, both in terms of enabling or hindering such futures.

Dimension	Barriers	Enablers
Culture / Social	• Poor levels of awareness with strong variation across communities. • Lack of consensus on terminology. • Current incentives push researchers to value the wrong things. • Risk that if metrics for open/reproducible research are badly designed they could be another box ticking exercise. • Epistemic diversity and risk of marginalization.	• The context where reproducibility is understood and implemented. • Aspects that must be influenced by other facilitators for reproducibility to become mainstream, normative and standardised • Shared understanding, incentivising, and spread of reproducibility practices through shared project work or collaboration.
System / Policy	• Lack of research support services. • Need for a reform of the reward and recognition system. • Inequality in available resources.	• Systemic change is needed to drive best practices. • Debate about the extent to which practices should be mandatory: If not, then will actors care about it?

	• Lack of reward and recognition policies that foster reproducibility. • Lack of alignment of policies across types of institutions and disciplines. • The interaction between research actors and industry or policy actors. Commercial interest might form barriers for transparency.	But if one mandates a 'one-size-fits-all approach', one runs the risk of enforcing standards upon people in contexts where they are not relevant or undesirable.
Technology/ Infrastructure	• Lack of methods to assess levels of reproducibility. • Lack of rules and standards. • Lack of interoperability of tools/infrastructures. • Potential negative uses of AI.	• Providing the tools, systems and platforms, with and within which researchers can implement reproducibility practices. • Digital and technological aspects (like data repositories or links between outputs to enable interoperability) but also non-digital aspects (like data stewards and submission guidelines)
Finance/ Economy	• Costs of reproducibility include extra time and resources need to enable transparency. • Those with less resources fall behind. • Lack of funding for reproducible practices. • Reproducibility practices are not rewarded within funding streams, including pre-award and post-award process.	• Funding is needed both to support reproducibility research / replications and for the work and infrastructures that support it.
Training / Education	• Missing statistics and data science skills. • Lack of expertise in Open Science skills. • Bad scientific practices.	• For researchers (mainly ERCs but also others) on reproducibility practices and skills (mainly, but not exclusively, related to data skills) • For other stakeholders involved in support and assessment • Also need to teach all stakeholders that 'good behaviour pays off'.

5.1.1. Cultural and social factors

Research culture: norms, values and practices

In terms of how it affects reproducibility, our participants framed culture as both the 'place' where reproducibility takes place (or doesn't), and as an enabler of or barrier to it. It is both the space where reproducibility is understood and implemented (or not), as well as a set of values, norms and practices that must be influenced by other enablers for reproducibility to become mainstream, normative and standardised (as much as is possible).

Among our participants, there is a general understanding that research culture heavily shapes practices, and that anything that influences research culture is therefore either an important enabler of or barrier to reproducibility. They illuminated the enabling role of research culture by pointing out how reproducibility practices sometimes bleed over from one area of research to

another through publication trends, collaboration, or the development of “critical mass”. For example, a researcher in our Machine Learning workshop said:

> *But it's true that Machine Learning, the culture that we were talking about before in Machine Learning, it was very common from many years ago to just publish the code. So, you have the source code. And for us in recommender systems, until the Machine Learning community appear[ed] and started publishing a lot, we didn't really pay a lot of attention to that. (Machine Learning Researcher, Scenario Planning)*

In our workshop with qualitative researchers, one researcher said of collaboration:

> *And then I'm also thinking about researchers being my collaborative partners. Like, even if I'm doing qualitative work, there might be other people doing all sorts of different work and interdisciplinary settings where they might very clearly sort of incentivize me to do different things, or in order to be part of something, some sort of larger project. (Qualitative Researcher, Scenario Planning)*

Speaking to how reproducibility might become more common within research cultures, in the funders workshop, remarked on the existence of our project:

> *And then, of course, all of these projects like TIER2 and others will continue to be going forward. Right? There'll be more of them. And so, the awareness and critical mass around reproducibility will hopefully continue and increase until something bigger happens. (Funder, Scenario Planning)*

Conversely, both groups of researchers identified poor levels of awareness about reproducibility and its importance as a barrier to it. In our pre-workshop survey, one ML researcher suggested those immersed in such issues may exist in a “bubble”, while others remain relatively unaware: “The scientific community should take reproducibility more seriously – I think we are in a bubble of enthusiasts”.

Reflecting a shift in research culture, participants in our funders’ workshop imagined that, in their ideal future of reproducibility, it will be both mainstreamed and normalised. They imagine that reproducibility will be “the new normal” (Funder), where reproducibility is practiced both “by default” and “by design” (Funder). Describing this, another funder said:

> *So, people see that it's important that it becomes kind of normalised in terms of policies and documents and guidance and all that sort of stuff. So, people maybe aren't thinking, I need to focus on reproducibility, but they have it in the back of their head as, ‘Oh, this is not bad. This is something we should think about.' (Funder, Scenario Planning)*

A qualitative researcher imagined that researchers are doing reproducible research because “all their peers are choosing to use reproducible methods and approaches.” Elaborating on an imagined future wherein there is “intrinsic motivation” to engage in reproducible research, they continued:

> *So, we have this scenario where we look in 2033 at qualitative researchers across disciplines and we find widely used reproducible methods, but not necessarily because they have to, [not] because they receive funding or because publications require it, but just because they want to do it. Or it's what's done. (Qualitative Researcher, Backcasting)*

Articulating this perspective with both breadth and depth, a qualitative researcher stated in the same workshop:

> *So, we have a scenario where an early career researcher is ready to share their data in September 2033. Um, and they can do it because there are a multitude of examples: Pilots across qualitative research approaches, whether it's interviews, focus groups, ethnographies, and field notes and so on. And there are guidelines and infrastructures that make it easy to make your data available to the next researcher. So, this early career researcher is so prepared that because of this, um, that when they collected their data, they also collected necessary consent from their subjects for further data sharing. And they have also documented meta information on qualitative contexts. (Qualitative Researcher, Backcasting)*

Reflecting a ML perspective, a ML researcher spoke to a shift in research culture that they imagine would reflect the normalisation of reproducibility:

> *I wish in a culture perspective that every lab should have, okay, you do your research, but there will be external eyes to see what you did and whether what you did is right or wrong in terms of implementation... So, I wish in the future to see this happening in every lab. Um, having someone who is just working on reviewing the code we did and whether it makes what we say in the paper, for example, in the conference at least because we cannot think for all the accepted papers or all the papers that are submitted, we cannot go through their codes. But at least for the reproducibility track before accepting a paper, I think a lot of work should be done going through the implementation of that paper. It's not about the number of metrics or algorithms they propose in there, but it's about how they implemented it and whether it reflects what they mention. Again, I think [name] mentioned about advertising, it's not about just having the paper accepted with so many quantitative things inside your package, but it's about the quality of what is there. And this means a lot to me. (ML Researcher, Scenario Planning)*

This quote concludes with mention of another theme that we observe in how participants discussed changes in research culture that would foster reproducible research – changes in values. Primarily, participants who spoke about this described a shift in values from quantity of research outputs to quality of research processes and practices.

In our workshop with qualitative researchers, participants imagined a shift to “slow science”. A qualitative researcher said:

> *And then this is very crucial. There should also be something of a paradigm shift towards slow science as the sort of way of doing things well, which we believe is going to be the final facilitator to get to the scenario where this sector-wide review is going to find that this is all over the place just because people really, really enjoy it and it brings them good science. (Qualitative Researcher, Backcasting)*

Institutional culture: norms, values and practices

Participants also observed that for reproducibility to be normalised, it must be embedded in the cultures of institutions and of assessment. This differs from the culture discussed in the previous section by the fact that the former refers to research cultures made up of communities of peers within a same field or discipline, while the latter refers to cultures within a single organisation or institute. Note that individuals are always members of both types of cultures at the same time and both types of cultures tend to interact.

A funder said, “And because we deem that reproducibility is important, institutions and policymakers also need to, you know, embed that in their culture saying, ‘You know, this is important to us’.” While, in the same workshop, a funder said, “Our future hope is that reproducibility is fully embedded in how research is being evaluated by funding councils, not as an add-on, but really normalised within the standard procedures.” Later, the same funder clarified that, in their ideal future of reproducibility, it would be included in a “broadened” set of practices that are “recognised and rewarded... So that provides space for reproducibility in there. It's maybe not focused on reproducibility, but it allows that to be in.”

Participants observed that institutional cultures and practices that foster competition rather than collaboration, and that reward “high impact” publications rather than rigorous reporting of results, are detrimental to reproducibility. A publisher reported in the publishers’ workshop anecdotal evidence of groups intentionally withholding methods to prevent replication and linked this specifically to their perception of the current research landscape as one of “competition as opposed to collaboration, sort of as a foundation of why research is done and how research is done”. Reflecting the same concern, a funder noted in the funder workshop that the dystopian scenario they had formulated (of over-competition and sensationalism in science communication) was in fact quite reflective of the current reality:

> *It's weird that some of this dystopian, negative thing is the extension of how things currently are, […] forcing a really negative culture that forces over-competition, loss of reputation from reproducing stuff. Researchers are only trained to chase the new leading, to kind of, exaggerate[e] every little discovery in science communications where the public thinks, you know, everything is going to cure cancer and change the world and solve, you know, global warming and stuff. But none of it does because it's not reproducible. It's all just shiny, flashy, new things which causes a loss of trust and loss of faith in science. (Funder, Scenario Planning)*

Breaking with valuing quantity over quality, which fosters competition to publish more, and explaining how a shift in institutional culture can impact research practices and culture, a

researcher in the ML workshop said during Scenario Planning, “If we move on to a different [research and institutional] culture, probably that is not focused on number of papers and is not focused on impact factor or things like that, then authors will say, okay, if I publish something that is reproducible, it will get more citations and my funder or my institution will value this paper more.”

Participants in the qualitative researchers workshop also imagined a future in which values have shifted among funders, with a switch from focusing mostly on “sexy topics” that generate new and greater quantities of research outputs, to “funding replication” (Backcasting). Explaining this, a qualitative researcher said, “So the novelty value of topics is sort of not the only value for why we are funding things for research, but also to look at these more methodological issues” (Backcasting).

On the possibilities for reform, participants were mixed. While participants were all very clear on the need to realign incentives to achieve cultural change towards open and reproducible research, they noted a range of difficulties including the general difficulty of disincentivising “shortcuts” and “cheating” (Publisher) and resistance from elite researchers prospering within the current system. As a funder observed in the funders workshop, “Why do they push back? It devalues their work.” Superficial commitment to DORA (the San Francisco Declaration on Research Assessment) was noted by a publisher as an example of the difficulty of changing engrained habits: “Most people say that the institutions pay lip service to DORA. I would say that almost every scientist I ever encounter judges’ people on the journals in which they publish.”

Participants also expressed concerns regarding possible negative side-effects of overly rigorous reforms expressed through requirements, especially that they could be reduced to merely “ticking the box” (Qualitative Researcher, Scenario Planning). Further still, a ML researcher stated during Scenario Planning that requirements might be rejected by researchers if they are seen as overly onerous or limiting of creativity and diversity in research: “Many of the scientists I know, they are really creative and playful people. And if you put too much rigor on these researchers, they will not enjoy research and start doing some something else.”
We note that discussions about cultural norms and values as enablers of or barriers to reproducibility were mainly discussed in the workshops with researchers, rather than funders or publishers. In addition, and potentially consequently, most of the actions mentioned as being required or desired relate to researchers (see appendix C).

Standardisation of definitions, norms and practices
Participants viewed lack of consensus on terminology related to reproducibility as a key barrier. Firstly, the terms “reproducibility” and “replicability” themselves are sometimes used synonymously, or sometimes given broad or specific meanings. As stated in response to the pre-workshop survey by publishers, “Reproducibility” can have a broad or narrow meaning: “it is both a specific thing (computational reproducibility - same data + same code = same results), and an umbrella term that includes consistency of results.” In the ML workshop, one researcher expressed that there is a need to be clearer about which aspects of research are the target of reproducibility:

> *Are we talking about reproducibility of experiments, of empirical observations, or are we talking about reproducibility of the method? Is that method something computational and all these kinds of things? (Machine Learning, Survey Results)*

Confusion or lack of clarity about such issues was noted by participants as hampering common understanding. As a researcher put it in the ML workshop during the discussion of survey results, “Every paper about reproducibility, it starts with saying reproducibility is a cornerstone of science and then we don't agree on what it is.” One publisher advised such confusion can derail discussions of specifics like the effort and cost associated with ensuring reproducibility:

> *I almost feel it's so critical because it affects everything that is discussed subsequently, including the costs, et cetera, is this issue of what exactly we mean. I know there's some different definitions, but certainly when I speak with scientists, I don't even think there's like replicability and reproducibility. I think there's about five different things here. And, you know, the ways in which we frame the issue. And again, as you alluded to, this differs by discipline. You can get very different answers as to what the amount of effort that you're willing to spend, the amount of cost and whether you need to worry about this at all… You know, if you repeat your experiment, how often do you get the same results? Is that what we're talking about? Or if I take your data. Can I reproduce the results that you got in your analysis? If I do the same experiment as you? Do I get the same results? If I do a very different experiment that addresses the same question, and my results replicate yours? Is that it? Or if I do a similar experiment in a different system, and do I get similar results, the whole generalizability? […] Then we have to have absolute clarity of whether we're covering it all or some specific aspect. (Publishes, Survey Results)*

Within qualitative research the very language of reproducibility was at issue for participants. As we examine later, the relevance of reproducibility to such work is highly contentious. Hence, just using the terms ‘reproducibility’ and ‘replication’ can act as a barrier to engaging qualitative researchers in broader discussions of the benefits of greater transparency, which can create possibilities for secondary analysis of data. This was communicated in a discussion during the backcasting exercise with qualitative researchers:

> *A: The whole time I was sort of wondering whether we're talking about reproducibility as like, just taking a study and trying to reproduce that finding, which every time I pitch that project, I pitch it as that's exactly not what we're doing because people have such a gut reaction, that it's so negative, this idea of being reproduced. Right? Because I'm saying, oh, we're using your data, but we're asking a new question. So, we're not trying to find out whether you are wrong, but we're trying to—*
>
> *B: Yeah, we're not verifying. We are actually sort of further—*
>
> *A: Taking that, we're producing greater value from the data that you've produced right now, we can focus on that. I don't mind that at all. (Qualitative Researchers, Backcasting)*

Therefore, to make reproducibility mainstream within research cultures, many of our participants, across stakeholder groups, emphasised the social process of standardising definitions, norms and practices—the elements of research culture that they view as key in shifting to reproducible research. Broadly speaking, they discussed a need for standardisation (to some degree) of

reproducible research norms and (methods) practices, for infrastructure and services, as well as definitions of the term of how to apply it to evaluate research for funding or publication.

In our workshop with qualitative researchers, one spoke in depth about the need for standardisation.

> *Um, so in this future scenario, I imagine that there will be standardized methods, workflows that are shareable and standardized. For example, instructions on how to share step-by-step qualitative analysis, because that's also important for reproducibility and qualitative researchers just don't know how to share analysis. And I was talking to this one researcher at some point, and they were like, 'Oh yeah, I'm recording video walkthroughs of, for example, how I code in NVivo my data.' And I'm like, 'Oh, wow, that's amazing, right?' If something like that can be shared. So, this sort of, you know, more innovative maybe, and more standardized ways of sharing these types of outputs so that it's normal and natural and easy for qualitative researchers to share these types of outputs. (Qualitative Researcher, Scenario Planning)*

Speaking to the need for clear definitions (and standards), a ML researcher said in the ML workshop:

> *If there is a too vague definition, obviously that might actually lead to gaming the system. So, everyone can claim […] that they're doing reproducibility. […] So, a very weak definition might be an issue there and that requires some consensus-making. That requires some search for what could be the essence of what we actually mean with 'reproducibility'. (Machine Learning Researcher, Scenario Planning)*

Leaders in change

While there is consensus among our participants that standardisation is necessary to achieve a reproducible research culture, there is disagreement on who should lead in creating standards. More discussion was oriented around the idea of community-driven standards that reflect and respect epistemic diversity, but some others suggested that a more top-down approach, led by professional associations, groups of international experts and multi-stakeholder organisations should be pursued. Reflecting our discussions about research culture and institutional culture as enablers or barriers regarding reproducibility, the discussions in this section highlight different views among our participants about which of these two types of cultures is best positioned to lead the way towards a desirable future of reproducibility.

Describing a community-driven approach but also suggesting that a higher-level organisation could foster this, a ML researcher said:

> *So, I think the community, whatever the community […] for every community they should really meet. And start agreeing on things because, yeah. Even if you make a paper. Yeah, that has a limited impact, right? So, I think the community or maybe [professional organisation] or maybe. Yeah. Someone at a higher level. Should not directly set a standard but try to gather the community*

> *[…] A higher level institution like ACM [Association for Computing Machinery] should put together researchers. And start finding agreements and I would say if we make this, then start a little by little […] Then in CFP or reviewer criteria. (Machine Learning Researcher, Backcasting)*

Emphasising the need to respect epistemic diversity, this same ML researcher also said:

> *Also, a good decision from the community would be to agree that more transparency is better. And, that the community will be put together and decide. And some agreements related with these objectives. And once these agreements are met, then they will start putting them in the call for papers and trying to not force, because maybe that's, and we already mentioned that, not all research needs to be reproducible, but so that reviewers are aware of what needs to be valued. (Machine Learning Researcher, Backcasting)*

A funder also spoke to the importance of research communities in spreading reproducibility norms and practices (often Open Science practices) when describing the function of reproducibility networks. They said, “I think having those kinds of networks and ensuring that researchers can work on that together. It's very important science, very much into statements that have been brought forward about the importance of fostering communities when fostering Open Science” (Funder, Scenario Planning).

Reflecting the idea that professional societies should have a leadership role in defining standards, a publisher said during the stakeholder mapping exercise of the publishers’ workshop:

> *So, I was just going to say, I think, [name] you're spot on really about the change. I've moved my professional societies closer to the centre because I'm thinking particularly in health care research, for example, where you've got standards and guidelines laid down by professional societies that you've obviously got to abide by. (Publisher, Stakeholder Mapping)*

Granted, professional societies tend to be composed of researchers, yet they have authority and power that is markedly different from an independent research community, so we understand this view to be different from those stated previously.

In a similar vein, some participants suggested that international experts and multi-stakeholder organisations should lead in standard-setting. Of this, a publisher said:

> *...I think [domain specific international committees] can be crucial. I mean […] you have international committees of academics who will define standards that everybody should subsequently use, and then often they're enforced by publishers. But, you know, that's critical to saying, ‘What is the thing that one should do?’ You know, I mean, it could be as simple as, you know, in statistics saying ‘You need to have N equals three’ or whatever. But there's almost every level. There's some group of international experts who say this is the minimum. (Publisher, Stakeholder Mapping)*

Epistemic and geographic diversity

Yet, whichever approach to standardisation is pursued, our participants agreed that standards should be field-specific and context sensitive. Reflecting this, a publisher said:

> *If you work in this field, this is what you have to do. This is the accepted norm. And I think that, you know, we can do that on a field-specific basis, like people, you know, in certain areas, clinical trials, for example, you understand you would pre-register your hypothesis. You know, if you're doing a certain type of neuroscience experiment, you understand these are your, the fMRI parameters you should be using. So, you have all those standardisations. (Publisher, Scenario Planning)*

Similarly, a qualitative researcher discussed, of standards specific to qualitative research that would be developed in an ideal future of reproducibility:

> *Metadata standards have been developed. So, if you are providing some sort of contextual appendices or any other type of data that goes with your raw qualitative data that you are archiving and how to do that. And guidelines were also developed on how to reconcile consent and replicability across qualitative approaches. (Qualitative Researcher, Backcasting)*

Describing how epistemic diversity could be reflected in standards and requirements, a qualitative researcher described a flexibility to data-sharing expectations during Scenario Planning and said, “one project can share some sort of derived data and another project can share the sort of the analysis workflow or the step-by-step analysis description; that there are these different outputs depending on the type of the study that can be shared.”

Reflecting the same concern, one funder stated in the pre-workshop survey: “I tend to agree with Leonelli’s observation that commentators on the importance of reproducibility and its associated terms (replicability, repeatability) tend to overlook these differences and sometimes overemphasize its value as proxy for quality and reliability in science.”

Participants observed (as do we) that, historically, certain quantitative-oriented fields have led in addressing issues of reproducibility while others have thematised the issue less, due to a perceived lack of relevance or importance. As expressed by qualitative researchers, even the term ‘reproducibility’ can have a marginalizing effect. As one expressed in the pre-workshop survey: “I don't think reproducibility (even broadly construed) is a goal for the majority of qualitative research”, while another responded in the survey:

> *The concept of reproducibility as defined, even in the broadest sense, does not really hold as a meaningful and/or desirable concept for many working in the traditional of qualitative research - although note there are some more realist perspectives for whom the concept has some grounding. (Qualitative Researcher, Survey)*

Another qualitative researcher shared in the survey, reflecting the marginalizing effects of the current reproducibility discourse and movement:

> *For the majority of members of the qualitative psychology community the drive for reproducibility presents a threat/risk. Wholesale moves towards centring reproducibility at the heart of what counts as 'good' science risks eroding/ devaluing research where reproducibility is not a measure of methodological rigour or research quality. (Qualitative Researcher, Survey)*

Therefore, for reproducibility reforms to be relevant and valued across epistemologies, what the term means, how it is used, and recognizing that other terms might be more applicable and resonant for some epistemologies is important. As a qualitative researcher said, pointing out the importance of terminology:

> *There's this automatic gut reaction and defensive stance that I think a lot of qualitative researchers have at the moment. You say 'transparency' because they already feel threatened [by reproducibility]. (Qualitative Researcher, Pre-workshop Survey Discussion)*

Additionally, participants recognised that reproducibility reforms and expectations should reflect the realities of research around the world, and not just in the global north. As one qualitative researcher said:

> *I wanted to echo the point that [name] put on the comment that I think all of us are from basically European kind of backgrounds and maybe have worked in North America versus there might be a stark difference if we're thinking more towards low- and middle-income countries in considering these kinds of questions and also the sort of burdens that put upon them. (Qualitative Researcher, Stakeholder Mapping)*

As one publisher put it in the pre-workshop survey, there is a need to "urgently […] bring researchers in from the Global South and LMIC [low- and middle-income] countries into the conversation." When reflecting on the workshops and the next steps to be taken by the community, the participants once again noted that, despite some variation in terms of participants' background, the discussions in our workshops should probably be characterised as Global North discussions. In the wider discussions on and approaches to reproducibility, including voices from a more diverse background was considered to be essential. To not do so may mean that reforms are not applicable or possible across geographies, as well as epistemologies, and therefore could pose barriers to spreading reproducibility reforms.

5.1.2. Systemic and policy-related factors

A second factor discussed in all workshops concerns the implementation of adequate policy to foster reproducibility. In terms of policy initiatives, the participants of all workshops are very clear that the only sustainable and effective way to foster preferable reproducibility futures is to make proper behaviour mandatory. This can relate to a host of different practices, ranging from data sharing, transparent reporting, appointing data stewards, providing training, etc., and it involves various actors, though mainly the traditionally powerful ones (funders, publishers, institutions, or even national governments) as they are best positioned to actually enforce certain practices from actors that depend on them.

> *"We have only made progress with increasing things like Data Availability Statement by mandating via journal policy." (Publisher, Survey).*
>
> *"I think someone needs to enforce this, and that would be those with the most influence, sort of. And in my view, that was grant agencies and publishers, and they should require transparency." (Machine Learning Researcher, backcasting exercise)*
>
> *"… if the government says institutions, funders, we want you to focus on reproducibility and this is what we value, then people will do it." (Funder, Stakeholder Mapping)*
>
> *A: "So, I mean, first it has to be kind of mandated, right?"*
>
> *B: "Yeah."*
>
> *A: "So there has to be like, like a mandate from EU Global Research Council, uh, worldwide funders or something like this, right. Like the big, the big names." (Funders, Backcasting, Pair 1)*

It is recognised that one policy usually requires several stakeholders to cooperate to effectively implement it, because actors have different roles in terms of drafting, executing, implementing, monitoring, etc. the policy:

> *The scenario that we were thinking about is the kind of minimum publication standard requirement for publishers. So, our statement said in 2020, in 2033, every relevant research article has a minimum standard of reproducibility of publication, and the standards are shared across communities and implemented and supported by funders, institutions and checked and reported on by publishers. (Publisher, Backcasting).*

Despite a more or less general agreement on the effectiveness and necessity of mandates, this practice also led to discussion. Some participants realised that this can only be done if a certain standard applies to all stakeholders involved. If not, then certain stakeholders will be forced into doing things that are not relevant, feasible or desirable. However, when one starts allowing for exceptions, the suggested policy loses part of its effect. This delicate balance led to discussion in the ML researchers' workshop, during the backcasting exercise. It illustrates how policy initiatives can be both enablers and barriers towards desirable futures of reproducibility:

> *A: Reject if the paper does not have an appendix documentation.*
>
> *B: Yeah.*
>
> *A: Maybe that's. Too hard, right? Yeah. What the. I forgot his name, but. Yeah. What he was saying. I think, um. That not all research needs to be reproducible. Right? So, I mean if you, if you. Yeah. So, it's too high.*
> *B: But if you keep it for researchers to choose, most of the researchers will not choose to do it because it's extra work. (Machine Learning Researchers, Backcasting)*

Another argument in favour of mandating policies, is that implementing and mandating policy across the board, can make implementation easier because it removes 'first mover risks', which could be detrimental especially for researchers (being afraid of not being able to get their work funded or published) and publishers (being afraid of losing submissions). This also speaks for stakeholders to act in collectives, rather than as individual entities (e.g., associations of publishers, rather than individual publishers):

> *The American Geophysical Union in the kind of Earth environmental sciences area basically shifted a whole community to open data by taking all their own journals and all the publishers with them. So pretty much in that subject area, every journal has a very similar requirement. So, in a way it removes the option, removes the risk from the publisher, from saying, well, actually we're going to have these high standards, but we'll lose all our submissions because the researchers will just go somewhere else. So, I was wondering if there was the potential for like a minimum standard for publication based on reproducibility criteria. And if all the publishers and societies adopted the same minimum standard, then that would remove the first mover risk. (Publisher, Scenario Planning).*

Some participants therefore suggest a stepwise approach:

> *So, we might see more funders mandating data and materials availability. Like that's not reproducibility, but it's kind of a first step. Like if the stuff's available, then people can start to, if they're motivated, do checking and those sorts of things. It's not mandating reproducibility, but kind of materials availability is like a first step towards that. (Publisher, Scenario Planning)*

A particular focus area concerning policies and how they can enable or hinder reproducibility practices, relates to the implementation of adequate incentive structures. Participants across the four workshops saw the need to integrate reproducibility practices into the full system and lifecycle of assessment; including into values of assessment

> *If we want to improve the situation of reproducibility, it needs to be integrated in how research groups are awarded grants, followed up during a project, or measured in their institutions. (Publisher, Survey)*

> *P2: Um. Yes. And then maybe also regarding the evaluation.*

> *P4.: The evaluation? Yes. Then and then I guess, um, reproducibility becomes a is a, you know, is recognized as a, as an element of research quality, I guess. And therefore, it is treated accordingly in terms of program evaluation, project proposal evaluation." (Funders, Backcasting Exercise)*

> *P14: Okay, Give me a second. Okay. Um. I think the thing that needs fixing is the whole incentive structure. Um, and you know, in, in the future, I think research assessment should be based on, um, you know, the. The quality of the research rather than the impact. And, you know, it should be judged on, you know, is the methodology good? Has there been pre-registration? Things that*

> *actually help improve trust in science. And so that's where am. Yeah. (Funder, Backcasting)*

Related to this, institutions were particularly mentioned and perceived to be important facilitators of reproducibility practices. They do not have a primary role in conducting reproducible research, but they are considered to be important supporters of reproducibility practices, either through properly rewarding and incentivizing such practices, or by providing the institutional conditions to achieve reproducibility, e.g., through training, installing data stewards or providing adequate infrastructure. It is acknowledged that incentive structures are particularly important, which is partly a duty of institutional actors.

> *And then training and mentorship. Yeah, I think that's really one of the important responsibilities of the, of the research institutions themselves. And yeah, my idea would be that reproducibility as a, as a research practice should be embedded in, in graduate programs. (Funder, Scenario Planning).*

> *Because I fully agree with you that for many of us, that wasn't so much built into people's education before becoming a researcher. And I find that more and more universities have a statistics helpdesk or a whatever, helpdesk, right? Where also later on you can actually find that sort of support and get the constructive engagement and you don't just need to find it yourself." (Qualitative Researcher, Backcasting, Pair 1)*

In one workshop, a participant expressed a concern about the effectiveness of institutions as facilitating actors. While institutions are perceived to be important actors in the network, this participant doubts whether they are really capable of moving things forward:

> *I'm less optimistic about the role of institutions and institutional policy makers. Maybe that's my bias, but at least in [country]. Well, and when it comes to, uh, thinking about the bigger recognition and reward issue, I think in at least in [country], the institution does play an important role. (Funder, Stakeholder Mapping)*

Lastly, participants across the workshops discussed institutionalization in the sense of setting norms and standards on a community level and agreeing on these:

> *If you work in this field, this is what you have to do. This is the accepted norm. And I think that, you know, we can do that on a field specific basis. (Publisher, Scenario Planning)*

As discussed before in the subsection on policy, participants note that there is a need for some level of standardization across contexts, in order to remove barriers like 'first mover risk'. As we mentioned before, there is however also a risk of moving too much into a 'one-size-fits-all model', which comes with its limitations. Therefore, institutionalization of reproducibility norms on a discipline level was considered to provide a suitable balance.

5.1.3. Technological and infrastructural factors

The development of suitable infrastructures and technological advancements were often mentioned as potential enablers for reproducibility practices, while the lack thereof was reasoned as a barrier. Participants in all workshops discussed these kinds of enablers and barriers, though they differed slightly in the kind of tools or infrastructures they focussed on.

In several discussions, participants noted that infrastructures form the backbone on which any other development or reproducibility practice is based. Hence, getting the proper infrastructure in place is a prerequisite to fostering desirable reproducibility practices:

> *You have the infrastructure and then you assume. Yeah, I mean there's a bit of a chicken and egg with the data repositories, et cetera. But I think the infrastructure from the publishing perspective, the infrastructure has to be there. And then you go and find the right things to link to. (Publisher, Scenario Planning)*

In this case, the infrastructure was considered so important that the participants decided to devote their entire discussion during the backcasting exercise to it:

> *So, we decided that one thing that would really help was if there was a complete knowledge graph of data provenance, the article forward linking and alerts to anything pertaining to reproducibility around an article. (Publisher, Backcasting)*

One of the ways in which participants differed in their responses was the level of detail provided about the kind of infrastructures or tools that they envisioned to be needed for improved reproducibility standards or practices. For example, when prompted for solutions to the reproducibility crisis, one of the qualitative researchers responded in the pre-workshop survey: “The application of reproducibility tools”, without specifying what such tools should look like. Similarly, in the discussion during the scenario planning exercise of the workshop with ML researchers, respondents settled on the need for “[…], probably, yeah. Some public infrastructure based on European or whatever funding.” When providing more detail, respondents mentioned various kinds of tools or infrastructures, for example related to data sharing and curation:

> *It is encouraging that data sharing is increasingly encouraged and even mandated by some funders, but researchers' ability to re-use and replicate data is largely dependent on context and curation. Publishers can support this by linking out to the data set from the publication, but data repositories should also consider what additional resources they could provide to help interpret and re-use the data. (Publisher, Survey).*

The level of detail provided, seemed to be related to the extent that participants were actively engaged in the development or use of certain infrastructures or tools. For example, when discussing data infrastructure in the ML workshop, two participants that were both actively using such infrastructures discussed the details of what ‘data centres’ should really provide during their backcasting exercise:

> *P19: So, I'm trying to interpret here what you mean by data centres. Are data centres available for people to just run their own experiments or to share their data or to post their reproducible experiments?*

> *P16: All three of it. [followed by a lengthy explanation of what this means] (Machine Learning Researchers, Backcasting)*

Importantly, the infrastructures discussed by participants did not only comprise technical or digital infrastructures. Especially in the workshop with funders (though also in others), participants discussed the need for infrastructures of expertise and support, for example in terms of data stewards or university research support staff, as an important factor, which was supported by the participant's funding agency:

> *They were expected to hire data stewards, and these data stewards had to work together to collect data according to their principles (Funder, Scenario Planning)*

> *Note on funding for data stewards, something that's the Dutch Research Council is doing at the moment, and I think we will continue to do that. That will also lead to a much better support infrastructure for reproducibility on an institutional level. (Funder, Scenario Planning)*

Other non-technological infrastructures thought to facilitate reproducibility include the availability of 'templates for submitting articles' (Machine Learning Researcher, Survey) or evaluation protocols at funding agencies focusing on reproducibility practices during the application stage (Funder, Backcasting).

Infrastructures or tools to share and reuse data were most commonly discussed. Participants in all workshops mentioned that these kinds of facilities are required to promote reproducibility. However, participants in multiple workshops also mentioned that data sharing and reusing infrastructures on their own are not sufficient. For example, in the publisher workshop, one respondent commented on the need to connect data infrastructures to other tools during the stakeholder mapping exercise:

> *On the need to not only create additional tools or databases, but mainly to allow linking and connection of these: "So if there's going to be lots of different research objects out there online in different places, like you might have papers in one place, data in one repository code in another, that all link to other grants, like all of the linking of all the different objects, needs to be clearer and better to just stay discoverability and therefore reproducibility further down the line. (Publisher, Stakeholder Mapping)*

A particular challenge coming up in this context is the missing possibility of linking between scientific platforms, partly attributed to a lack of agreements on meta-data standards and guidelines. Discussed was also the need for instruments supporting the whole workflow towards reproducible science:

> *Infrastructure - better infrastructure is needed to make research workflows, both quantitative and qualitative, more reproducible (it should be easy to conduct reproducible research studies, right now there are too many barriers and too few smoothly working protocols, routines, workflows). (Qualitative Researcher, Survey)*

In a similar vein, one of the participants in the workshop with qualitative social science researchers mentioned that the existence of infrastructures might be an important – or even necessary – first step, but in itself it is useless if proper reproducibility practices are not actually implemented:

> *Yeah, I just wanted to say one other thing, which is in this ideal scenario, which we're talking about now, right, is I also think that what is really needed is, is not just guidelines and tools, because those are very important. Right? I agree that they're important, but to actually apply them. (Qualitative Researcher, Scenario Planning)*

Most prominent among the qualitative researchers but also a topic for the ML stakeholders is the lack of standardisation. This concerns all different levels of presentation, communication and evaluation.

Examples from qualitative research are the formatting of results or the software incompatibility:

> *Issue that transcripts shared are not uniform in format, style, presentation; no agreement on how to share – big issue if you want to reproduce or replicate. (Qualitative Researcher, Fieldnotes)*

> *Different software used for data management and analyses (software can affect data processing and analysis and consequently the results; proprietary software makes reproducibility difficult due to inaccessibility). (Qualitative Researcher, Survey)*

The ML researcher identify missing standardisation of evaluation and documentation criteria as an issue:

> *The main issue is that in my field (recommender systems) there are no public standard measurements anyone could compare against, so there are several "standards", and even though it might be proved that one of those standards is flawed, it takes time for the community to catch up, and in the meantime, more and more papers continue using those standards for compatibility or comparability. (Machine Learning Researcher, Backcasting)*

Related to this, participants also pointed out that merely having infrastructure in place to share code is insufficient. Checking the validity of code is a burdensome process, requiring dedicated infrastructure and appropriate reward mechanisms:

> *Having someone who is just working on reviewing the code we did and whether it makes what we say in the paper, for example, in the conference at least because we cannot think for all the accepted papers or all the papers that are submitted, we cannot go through their codes. (Machine Learning Researcher, Backcasting)*

Overall, participants mostly discussed rather mundane tools that either already exist or seem relatively easy to develop in the near future. However, in some cases, they also discussed very futuristic tools that are only in their very early stages of development:

> *A: And so now that Elon is developing brain implants, kind of facilitating that rather than having to write everything down while you're writing it, which makes it painful, kind of hooking into your brain and get the brain dump as a as an appendix automatically.*
>
> *B: And there have been many situations where I thought if only, I could sort of duplicate myself and do some of these things with another person, that'd be so great. (Qualitative Researcher, Scenario Planning).*

In terms of more novel or futuristic tools, particularly in qualitative research, participants name the uptake of new methods as essential to counteract the high workload of scientists. However, one concern that was discussed among publishers, funders and qualitative researchers is the (i) error-prone or (ii) improper (e.g., to fake results) application of AI technologies.

> *We will see a lot of sloppy throwing AI at qual data to code for you, will degrade quality of qual research – commercial products already have automated coding – this is a disaster and makes a mess that is hard to disentangle. (Qualitative Researcher, Fieldnotes)*
>
> *But it will also become a feasible to fake qualitative data, which I don't think currently is reasonably doable. (Qualitative Researcher, Backcasting)*

In multiple workshops (funders, publishers, ML researchers), the evaluation of reproducibility or of the reliability of results were raised as issues. This also connects to the “Improper Use of AI”, and how applying AI methods to fake and improve scientific results may alter the scientific landscape and its trustworthiness.

We conclude this section with two observations. First of all, we note that in most cases, participants refer to infrastructures that they either actively use themselves or have the possibility to develop. This was for instance the case in the example of the knowledge graph discussed in the backcasting exercise of the publisher workshop or the data stewards discussed in the funder workshop, both mentioned above. Hence, participants put a certain level of responsibility on their own shoulders by discussing a need for infrastructures that they themselves can or should develop. There were some exceptions to this, however, most notably in the publisher workshop, where pair 1 in the backcasting exercise decided to focus on a somewhat utopian future, which the publishers have only limited control over and hence responsibility for:

> *So, you could imagine a radically different setup for at least science. And therefore, papers would come to reflect that and they would look rather different. They would be a genuine report of something that had been done rather than something that was used primarily as a career currency. (Publisher, Backcasting)*

Lastly, we note that infrastructure was not only mentioned as a direct enabler of or barrier to reproducibility practices but also in connection with the changing temporalities of science. In this sense, the proper use of infrastructures and tools were thought to allow more effective use of time, which could ultimately benefit reproducibility practices and research quality more generally.

Reference was made to recent developments in AI, which could free up time for the researchers, e.g.:

> *Like technologies enable. Um. Technology enables kind of easy engagement? Um. Effective. Time use, you know. I suppose what you said is technologies enable working in reproducible ways. (Qualitative Researcher, Backcasting)*

5.1.4. Financial and economic factors

Our participants overwhelmingly framed financial and economic factors as the primary enablers of reproducible research and as key barriers to it. They stated that funders can drive reproducible research practices by rewarding them with funding and by requiring them in grant agreements (and hinder them by not doing so); that they can foster the spread of values, practices and awareness by funding certain tracks of research (like by funding our project and its sisters); that they can specifically fund reproduction and replication studies; and that they and other stakeholders can fund the work that supports reproducible research (like data stewards, trainers, etc.). When this work is not funded, it is difficult to achieve in the context of a rushed, publish-or-perish research culture.

To this latter point, one qualitative researcher referred to “the cost of reproducibility” in the pre-workshop survey, and explained that extra time and resources are needed to enable transparency of methods, data, code, etc. In the words of one funder, “it is still often costly, sloppy and takes time” (Funder, Pre-workshop Survey). One ML researcher noted during Scenario Planning that these “costs” to “go the extra mile” are at odds with current incentives to “publish a lot.” Another participant in the same workshop later noted that this means that not all researchers are able to enact reproducibility:

> *Reproducibility for now is not for everyone. Not everyone is doing reproducibility and not everyone can do it because of resources and costs. Not everyone has the money and funding to do it. (Machine Learning Researcher, Backcasting)*

Framing this as a systemic problem, one funder stated in the pre-workshop survey that reproducibility practices are “currently not really rewarded or recognized within funding streams and processes, including pre-award and post-award processes.”

However, participants believe that funders can solve this problem by specifically funding the labour costs associated with this work. One funder stated during Scenario Planning that in an ideal future of reproducibility, funded researchers would be “expected to hire data stewards, and these data stewards had to work together to collect data according to their principles. And if they wouldn't agree to work together in that scheme that we set up, they just wouldn't get the money for the research. So, I think if we would want to as funders, there's ways to do this.”

In a similar vein, a qualitative researcher commented during Scenario Planning, “One would wish there was like grant funding on top of, not just as part of what is already offered. You could actually have people as part of your project who are actually taking care of all the reproducibility issues, right?” And continued, “And that means that there is work hours that have to be put into this and that people are being hired to do this right with you or for you.”

Later, during Backcasting, another qualitative researcher pointed out that funding for research support positions would solve the problem of "people [not having the] time to do these extensive pilots for data sharing. They need to get it funded or like the time needs to be allocated for that."

Not only do participants believe that funders can support reproducibility by funding the work required to create it, but also that reproducible research should be valued and in assessment procedures by funders. Speaking within the funders workshop, one participant said:

> *My preferred scenario would be that reproducibility would really be recognized as one of a broader set of Open Science practices and really recognized and rewarded in funding in the grant awarding procedures of, of [institution]. We are very much looking at how the European Commission is doing that right now in the Horizon Europe program. We're having the same kind of discussions right now. Um, it's not the case, at the moment, that at least that open science is a part of the criteria or the procedures across the board. But we do feel that it's important to further Open Science at large and within that broader agenda also reproducibility. (Funder, Scenario Planning)*

In terms of the research content that is funded, participants spoke about the concept of funders creating traction or "critical mass" around reproducibility and related issues, like the funding stream that supports this project and others like it. One ML researcher commented during Backcasting, "So, the fact that we have this meeting and there's research projects funded by EU, there is some traction in this." And a funder commented similarly during Scenario Planning, "And then of course, all of these projects like TIER2 and others will continue to be moving, going forward. Right? There'll be more of them. And so, the awareness and critical mass around reproducibility will hopefully continue and increase until something bigger happens."

Finally, some participants suggested that funders should specifically fund reproduction and replication studies to effectively "kickstart" these practices. Of this, one funder said:

> *So, I put a call for replication studies. I totally agree that it should be maybe embedded in the funding instruments that are already there and it doesn't, we might not need it eventually. This calls for replication studies, but to make it maybe more the mainstream. It's like with, I don't know, gender quotas or stuff like that might be at the moment. It still needs these special calls and then eventually we can move to, to embed it into. Yeah. (Funder, Scenario Planning)*

A qualitative researcher agreed with this approach during Backcasting, and suggested it as an alternative to funders focusing on "sexy topics". They stated that if funders put out calls for such studies, then researchers would think, "Oh, we have to we have to now focus on, I don't know, levelling up in the UK or whatever. Right?"

Funders, meanwhile, recognized both pros and cons to this suggested approach:

> *I agree with most of what you're saying, but something that struck me was this funding for reproducibility research that you brought up? Yeah, I agree. I agree*

> *with you. There should be more. But if we're thinking about ideal scenarios, like in ten years, and I put a comment here, maybe, you know, it should just be embedded everywhere, right? So, it shouldn't have specific funding for it. (Funder, Scenario Planning)*

However, one funder noted during Scenario Planning that, despite an interest in supporting reproducibility, "it's harder for us to ask for [reproducibility] because of the way that we fund things, because it's less stable and more project based," suggesting that supporting reproducibility requires longer-term funding.

Another type of economic barrier identified by our participants is the proprietary nature of some data. Within our ML workshop, economic issues regarding data as intellectual property arose in two contexts by the same respondent. In general, collaboration with industry partners for whom "data is a digital gold and they don't want to have it everywhere but actually with them and not sharing" was seen to limit the data sharing that ensures reproducibility during Backcasting. The same respondent also cited a more specific case, namely the issue of research done on proprietary datasets where changes in licensing conditions can mean the data on which previous research rests is no longer available:

> *In our field, access to potential data that is not freely or economically viable, accessible [is an issue]. The classic example would be a lot of research has been done using almost public datasets such as Twitter, and this data is no longer freely available and has to be now bought in ways that would not allow most of the research projects that have been done in the past. (Machine Learning Researcher, Scenario Planning)*

ML researchers noted that industry-funded research also presents a barrier, in that sharing of data, software and other materials is often not allowed by industry partners for reasons of competition. Researchers working in such collaborations would not be able to create reproducible research.

5.1.5. Training and educational factors

In addition to the dimensions covered in the previous subsections, training and educational activities are considered another basic element required to effectively promote and scale reproducibility practices. Participants in all four workshops mentioned these elements as being critical and agreed that they comprise a core element on the road to a preferred future state of reproducibility. Formal training, as well as other educational practices like mentorship and role modelling, are considered a means to "get policy down to the individual researcher" and to "incorporate it in everyday research" (Machine Learning Researcher, Backcasting, Pair 1), making "those things normal" (Publisher, Scenario Planning).

This training can address multiple aspects of conducting and organizing research as well as addressing various actors. An often-recurring combination of training targets comprised the training of early career researchers or students on technical skills, research methods and statistics. Participants also frequently brought up the general need for Open Science skills and the skills required to use Open Science platforms (Publisher, Survey). Several participants gave examples of having been involved in this kind of training (either as trainer or trainee) or mentioned that such training should be more widely developed and implemented:

> *[…] the reproducibility practices are integrated also in the university curriculum, so that from early stages on, students who will become researchers at some points are already familiar with these practices. (Funder, Backcasting)*

> *[…] a positive driver would be that basically reproducibility being integrated in university curricula. Right. Somehow, I mean it I think it's important to make sure that already at the level of education…. at the level of university curricula, that there is some kind of training on how to ensure reproducibility of results. (Funder, Backcasting)*

Related to this, participants also discussed the extent to which this kind of training or mentorship should be mandated, mostly agreeing on the fact that mandatory training would be beneficial:

> *I think like a really important component here is that there is like training in data sharing and archiving at institutions, right? So, so there is like, um, data sharing and archiving training is obligatory. (Qualitative Researcher, Backcasting)*

Some participants even went a step further by mentioning the possibility of mandating the conduct of a replication study for PhD students to obtain their degree:

> *And I quite liked the provocative statement used in the example about replication studies being required of PhD candidates. (Publisher, Backcasting)*

Participants also noted that training should not only be provided to students or early career researchers. Qualitative researchers see a need for trained research support staff to facilitate tasks related to open science e.g., data management and analysis.

> *Research support staff is often not trained to help with reproducibility issues, or it is not a part of their tasks. (Qualitative Researcher, Survey)*

> *Division of labour - researchers need better-trained research support staff - data managers, data analysts, data curators, research software engineers - to get help to make their work reproducible. It is not one man's job. (Qualitative Researchers, Survey)*

In addition, senior researchers and principal investigators should also receive training, as well as actors in other roles, including evaluators at either journals or funding agencies:

> *Training for evaluators, I would say is an important step in the process. […] obviously, um, when implementing those kinds of measures, training is needed for evaluators. (Funder, Backcasting, Pair 1)*

Taking this a step further, one participant noted that teaching endeavours related to reproducibility could extend beyond the realms of academia. This participant mentioned that they could be used to create pressures from the general public to increase reproducibility levels, by making the general public understand what reproducibility is and why it is important. Note the traditional and much criticized science literacy model that underpins this comment:

> *Like from the public demanding reproducible research because we make them literate enough to understand what this is about. (Funder, Scenario Planning)*

Apart from extending beyond merely early career researchers and students, training should similarly not only address research practices or data-related skills. Participants of our workshops noted that training is required to accomplish a culture change, in which researchers know that the additional efforts required to make their studies reproducible, will be rewarded:

> *Like trying to teach people that the extra work will be rewarded by more citations (Machine Learning Researcher, Backcasting)*

Conversely, participants mentioned that, if implemented improperly, role modelling or mentorship could also be a barrier to reproducibility. In particular, some participants mentioned the risk of early career researchers being socialised into a culture where cutting corners is considered the norm. PIs might be tempted to train their PhD students to implement bad scientific practices or even commit fraud in order to achieve certain goals.

> *...what I wrote down in training and mentorship, um, is to yeah, so this is a bit of a different perspective so that if yeah, so what we don't want to see in the future is a push, push their students to conduct fraud or questionable research practices in order to, to get funding. (Funder, Backcasting)*

Finally, participants made some comments about how the resources required for these training and mentorship endeavours could or should be accumulated. Primarily, they point to funding agencies that could provide dedicated schemes or resources for these kinds of activities:

> *Then regarding training and mentorship, maybe special calls or funds that could again be embedded in normal funding schemes for training of reproducible practices. (Funder, Scenario Planning)*

6. Discussion

The stakeholder mapping exercise and the discussions that took place across all the workshops, illustrate that researchers are the stakeholders who are most influential for shaping the future of reproducibility and can take the steps towards creating the preferred future for reproducibility. The second most influential stakeholders were funders, and the third were publishers.

Participants highlighted throughout the workshops and the validation workshop that a change in culture is the most important aspect for the preferred futures. Different stakeholders discussed different aspects of the changes that need to occur culturally for their preferred futures to come fruition. However, there were a few overlapping ideas. First, the new culture must be quality over quantity to ensure for best practices. In corelation, reproducibility must become the new norm. As indicated by our participants, these cultural changes depend a lot on young and first stage researchers therefore these groups need to have reproducibility and Open Science training incorporated into their studies and PhD's to establish the aforementioned cultural changes in the research community.

The importance of standardization and a definition for reproducibility was a widely discussed topic. All stakeholders agreed on the importance on the need for a solid definition for reproducibility and reproducible research. The importance for journals, publishers, conferences, and funders to agree to and implement the definition. However, qualitative and machine learning researchers drew on the importance of variation across disciplines. In the preferred futures the definition for reproducibility needs to be solid but needs to have enough variety to account for different disciplines needs and to aid progression not hinder it.

Incentives played a big part in the preferred futures, especially in enabling a long-term stability for reproducibility. The most discussed incentives were increased recognition and visibility awarded to researchers who use reproducible and Open Science practices. Additionally, alternative outputs are rewarded and recognized. A key incentive discussed by all stakeholders, is for funders to embed extra funding and to require researchers to fulfil and comply to reproducibility standards and to make these actions visible and clear.

Discussions were dedicated as to how guidelines and infrastructures can be shaped to foster reproducibility across contexts, stakeholders, and disciplines. The main conclusions were for funders to have clearer guidelines and policies to provide researchers on what they need to comply with. The same is for publishers and journals on what they will assess articles on. Open-source tools will be available to aid reproducibility efforts in the long-term.

The study aimed to not only explore the preferred futures for reproducibility but also the barriers and enablers for them. For culture change a lack of consensus on terminology, poor levels of awareness, current incentives pushing researchers to value the wrong thing, epistemic diversity and the risk of marginalization are the greatest barriers for cultural change. Whilst a space where reproducibility is understood and implemented in the mainstream is spread through shared project work is seen as enablers for cultural change.

When reflecting on the policy incentives identified in the workshops, it is key to understand the critical role technology and infrastructures play, as lack of research support, lack of reward and recognition, lack of alignment of policies across types of institutions and disciplines and research actors are barriers towards the preferred future.

This study has several strengths. First our study is the first to have explored the future of reproducibility, concerning not only a preferred future but also alternative futures such as the undesired, realistic, and outlier futures. And thus, potentially exploring new insights that may impact how we see, perceive and view reproducibility from different stakeholder perspectives. The use of the future studies methods (see methods section) helped us to build a more complete picture of reproducibility, not only representing the ideal future but also the undesired futures and the enablers and barriers towards the preferred future. Furthermore, we included multiple important stakeholders, including funders and publishers, providing a broad range of perspectives on the future of reproducibility. Our aim is to help different stakeholder groups during the process towards their preferred futures of reproducibility. We also explored different domains (machine learning and qualitative research) to understand the epistemic diversity of reproducibility. Additionally, the Miro board provided the opportunity for us to collect different international perspectives by allowing us to include participants from a variety of countries.

The results should be considered in the light of some limitations. Firstly, only four workshops were performed. Whilst each workshop highlights important and different perspectives, we unfortunately cannot generalize our results as the perspectives are representative of a select set of individuals who are recruited because they are knowledgeable and have expertise on reproducibility. Therefore, a quantitative study with a larger sample size could be a next step to validate the perspectives from this qualitative work and see how generalizable the results are. Moreover, we failed to include other domains, such as biomedicine and humanities, in the study reducing the generalizability of our results. However, we still had a diverse setup of participants representing different disciplines (machine learning and qualitative research), different institutions and countries. This could be interesting in future research to explore differences between countries. Lastly, the scenarios that emerged from the backcasting exercise were focused only on one aspect of the of the preferred future limiting our understanding of the direct steps needed to be taken towards the preferred future, nonetheless, the future studies methods clarify the picture of the future.

7. Conclusion

In conclusion, we will use these results to further the discussions on reproducibility. Specifically, these findings will be used to develop tools in the next steps of the TIER2 project. The results at hand can be used to see how the developed tools can fit the preferred future. Furthermore, the results can help with the pilot testing and implementation of the tools. During testing, we can examine whether the tools are effective to reach the preferred futures for the stakeholders. The enablers and barriers identified during the workshops can have substantial influence in the implementation of the pilot tests in the next phase of the project, particularly for helping successfully implementing the tools per stakeholder group. This can help in steering towards the right direction and be prepared for certain barriers and to help use certain enablers. Lastly, the differences between stakeholders preferred futures will be used to gap bridges to help initiate the conversations between them to further reproducible efforts.

Acknowledgements

We would like to thank and acknowledge all our participants for taking time out of their schedules to participate and provide their expertise and insights during our workshops and validation workshop. We also would like to thank the internal reviewers for their feedback and suggestions.

References

Cobey, K.D., Fehlmann, C.A., Franco, M.C., Ayala, A.P., Sikora, L., Rice, D.B., Xu, C., Ioannidis, J.P.A., Lalu, M.M., Menard, A., Netizel, A., Nguyen, B., Tsertsvadze, N., & Moher, D. (2023). Epidemiological characteristics and prevalence rates of research reproducibility across disciplines: A scoping review of articles published in 2018-2019. *eLife*. https://doi.org/10.755/eLife.78518.

Deterding, N. M., Waters, M. (2018). Flexible coding of in-depth interviews: A twenty-first-century approach. *Sociology, 50,* 708-739. https://doi.org/10.1177/0049124118799377.

Haven, T.L., Gopalakrishna, G., Tijdink, J.K., van der Schot, D., & Bouter, L. (2022). Promoting trust in research and researchers: How open science and research integrity are intertwined. *BMC Research Notes, 15*(1), 302. https://doi.org/10.3122/osf.io/ngt3p.

Inayatullah, S. (2013). Futures studies: theories and methods. *There's a future: Visions for a better world,* 36-66.

Leonelli, S. (2018). Rethinking Reproducibility as a Criterion for Research Quality. Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise (Research in the History of Economic Thought and Methodology, Vol. 36B), *Emerald Publishing Limited*, Bingley,126-146. https://doi.org/10.1108/S0743-41542018000036B009.

Leonelli, S. (2019b). What distinguishes data from models? *European Journal for Philosophy of Science, 9*(2), 22. https://doi.org/10.1007/s13194-018-0246-0

Nichols, J.D.., Oli, M.K., Kendall, W.L., & Boomer, S.G. (2021). A better approach for dealing with reproducibility and replicability in science. *PNAS, 118* (7). https://doi.org/10.1073/pnas.2100769118. Nosek, B. (2019b). Strategy for culture change. *Center for Open Science* June 11. https://www.cos.io/blog/strategy-for-culture-change

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349* (6251). https://doi.org/10.1126/science.aac4716.

Peng, R.D. (2011). Reproducible research in computational science. *Science, 334* (6060), 1226-1227. https://doi.org/10.1126/science.1213847.

Popper, K. (2002). *The Logic of Scientific Discovery* (2nd ed). Routledge.

Walters, C., Harter, Z.J., Wayant, C., Vo, N., Warren, M., Chronister, J., Tritz, D., & Vassar, M.(2019). Do oncology researchers adhere to reproducible and transparent principles? A cross-sectional survey of published oncology literature. *BMJ Open*. https://doi.org/:10.1136/ bmjopen-2019-033962

Appendices

Appendix A: Workshop Materials

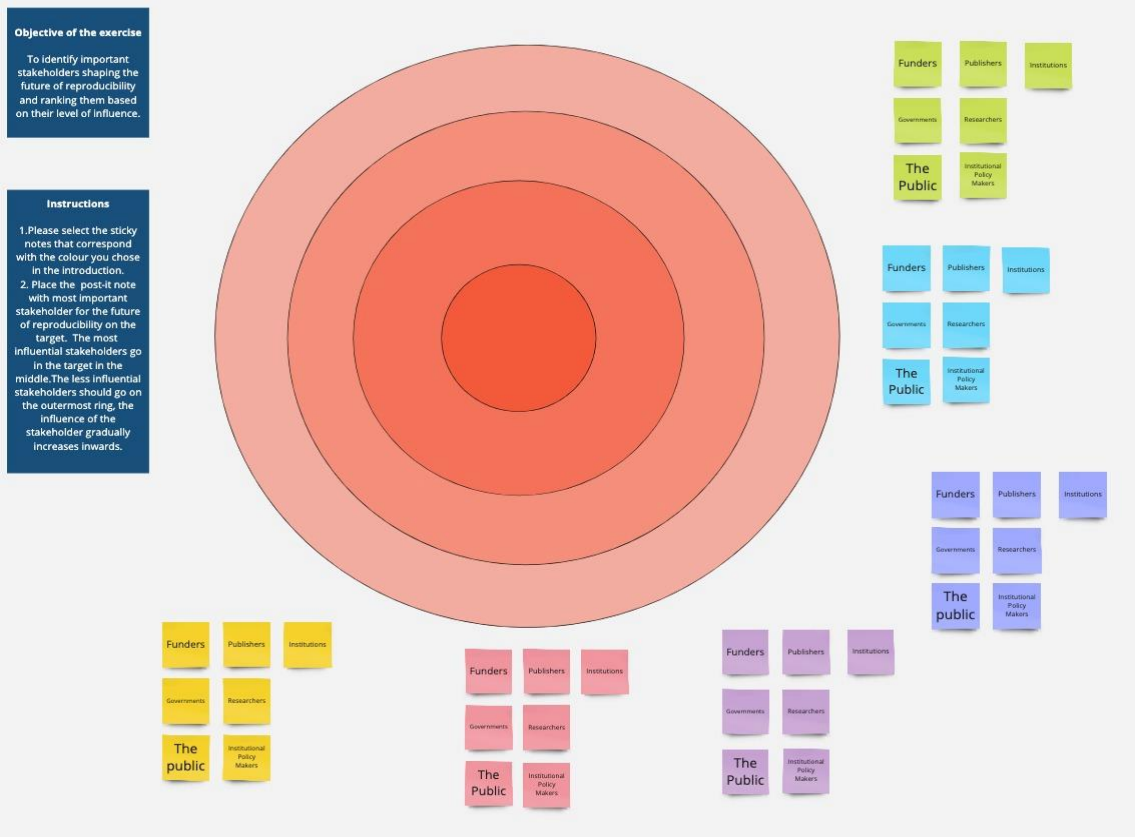

Figure A1: Stakeholder mapping exercise.

Figure A2: The four (preferred, dystopian, realistic, and outlier) scenario planning tables. The four orange sticky notes represented the four identified themes from the pre-workshop survey.

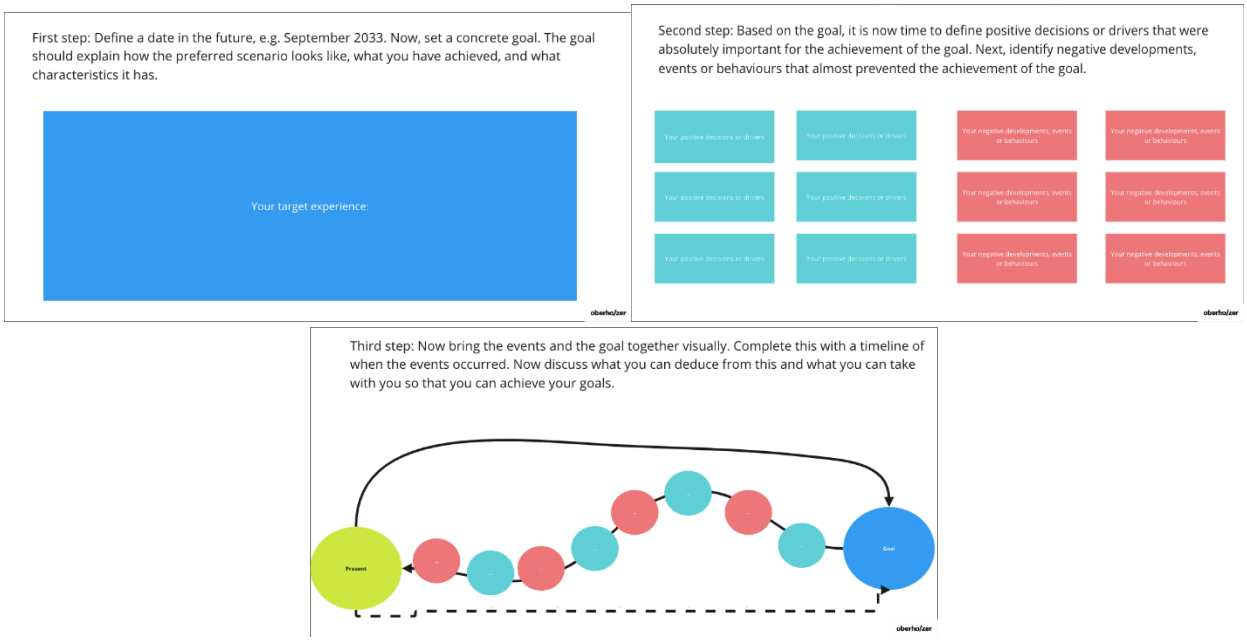

Figure A3. The three rectangles represent the three steps of the backcasting exercise. The first rectangle shows where participants wrote their target goal, the second rubric was where participants included their enablers and barriers. The last rubric was the timeline where participants backcast from the past to the future by adding the enablers and barriers.

Appendix B: Deductive Coding Structure

- Stakeholders (parent code—can use to capture longer discussions about stakeholders while also coding specific ones listed as subcodes)
    - Funders: code when the concept of funding or a funder or any specific funder is mentioned
    - Publishers: code when the concept of publishers or publishing or any specific publisher is mentioned (including journals and editors)
    - Institutions: code when the concept of an institution (scientific/research/higher ed) or any specific institution is mentioned
    - Scientific societies/associations: code when the concept of a scientific society/association (e.g., EUA, National Institutes of Health, etc.) or any specific one is mentioned
    - Infrastructure and service providers: code when the concept of one or a specific one is mentioned (e.g., open data platforms, EOSC, Embassy of Good Science, Web of Science, etc.)
    - Researchers: code when the concept of researchers or any specific one or group is mentioned
    - The public: code when the public (i.e., non-scientific stakeholders, non-governmental) is mentioned, either large or small
    - Governments: code when the concept of government (civic) or any specific arm or administration is mentioned
    - Policymakers: code when any kind of policymaker is mentioned (e.g., institutional policymakers, science policymakers, educational policymakers, etc.)

- Reproducibility practices (parent code – can use to capture longer discussions about various repro practices while also coding specific ones listed as subcodes)
    - Open data: code when any type of open data practice is mentioned (e.g., preparing metadata, sharing data, using repositories, reusing data, etc.)
    - Open methods: code when any kind of transparent or open methods practice is mentioned (e.g., Open notebooks, documentation of methods, etc.)
    - Open evaluation: code when any kind of open evaluation is mentioned (e.g., open peer review, open assessment, etc.)
    - open access: code when any reference to open access publishing or documents is made (excluding data)
    - Open code/software/tools: code when any reference to open code, software or tools is made (e.g., creating, sharing, using, etc.)
    - Open infrastructure: code when any reference to open, non-proprietary infrastructure is made (e.g., services, standards, databases and platforms (e.g., preprint servers) that are community owned and free to use and replicate)
    - Pre-registration: code when any reference to pre-registrations or sharing of study protocols is made.
- Enablers (parent code – can use to capture longer discussions about enablers while also coding specific ones listed as subcodes. Note that this includes discussions of both enablers that are already in place, enablers that could become available or those that should be implemented by some stakeholder.)
    - Technological tools and infrastructure: code when any reference is made to how technological tools or infrastructure can support reproducibility (e.g., AI, Open Science infrastructures and tools, preprint servers, repositories)
    - Cultural/social: code when any reference is made to culture broadly, or to norms, values, ideas, worldviews/ideology, or social relationships that support reproducibility (e.g., between researchers, mentoring relationships, etc.)
    - Institutional: code when any reference is made to something an institution can do to support reproducibility (e.g., changing hiring or promotion criteria, assessment procedures, etc.)
    - Systemic: code when any reference is made to systems (scientific or social) or elements of them that support reproducibility, e.g., assessment structures, incentives for replication and reproducibility
    - Financial/economic: code when any reference is made to financial or economic investments that support reproducibility
    - Policy: code when any reference is made to any kind of policy that supports reproducibility (e.g., any actor mandating or encouraging Open Science or reproducible practices)
    - Training and education: code when reference is made to any kind of training, education and supervision that supports reproducibility
- Barriers (parent code – can use to capture longer discussions about barriers while also coding specific ones listed as subcodes. Note that this includes discussions of both barriers that are already in place, barriers that could become available or those that might be implemented by some stakeholder.)

    - Technological tools and infrastructure: code when any reference is made to how technological tools or infrastructure are undermining reproducibility or reproducible practices
    - Cultural/social: code when any reference is made to culture broadly, or to norms, values, ideas, worldviews/ideology, or social relationships that get in the way of reproducibility (e.g., between researchers, mentoring relationships, etc.)
    - Institutional: code when any reference is made to something an institution is doing that undermines reproducibility (e.g., hiring or promotion criteria, assessment procedures, etc.)
    - Systemic: code when any reference is made to systems (scientific or social) or elements of them that undermine reproducibility, e.g., assessment structures, disincentives for replication and reproducibility
    - Financial/economic: code when any reference is made to financial or economic aspects that undermine reproducibility
    - Policy: code when any reference is made to any kind of policy that undermines reproducibility
    - Training and education: code when reference is made to inadequate or missing training, education or supervision
- Scenarios
    - Preferred: code when participants discuss the preferred scenario
    - Likely: code when participants discuss the likely scenario
    - Dystopian: code when participants discuss the dystopian scenario
    - Outlier: code when participants discuss the outlier scenario
- Methods: code any and all discussions of our methods (e.g., explanations of an exercise, discussion of how to do an exercise, questions about or feedback from participants about our methods)

Appendix C: Stakeholder Mapping Results

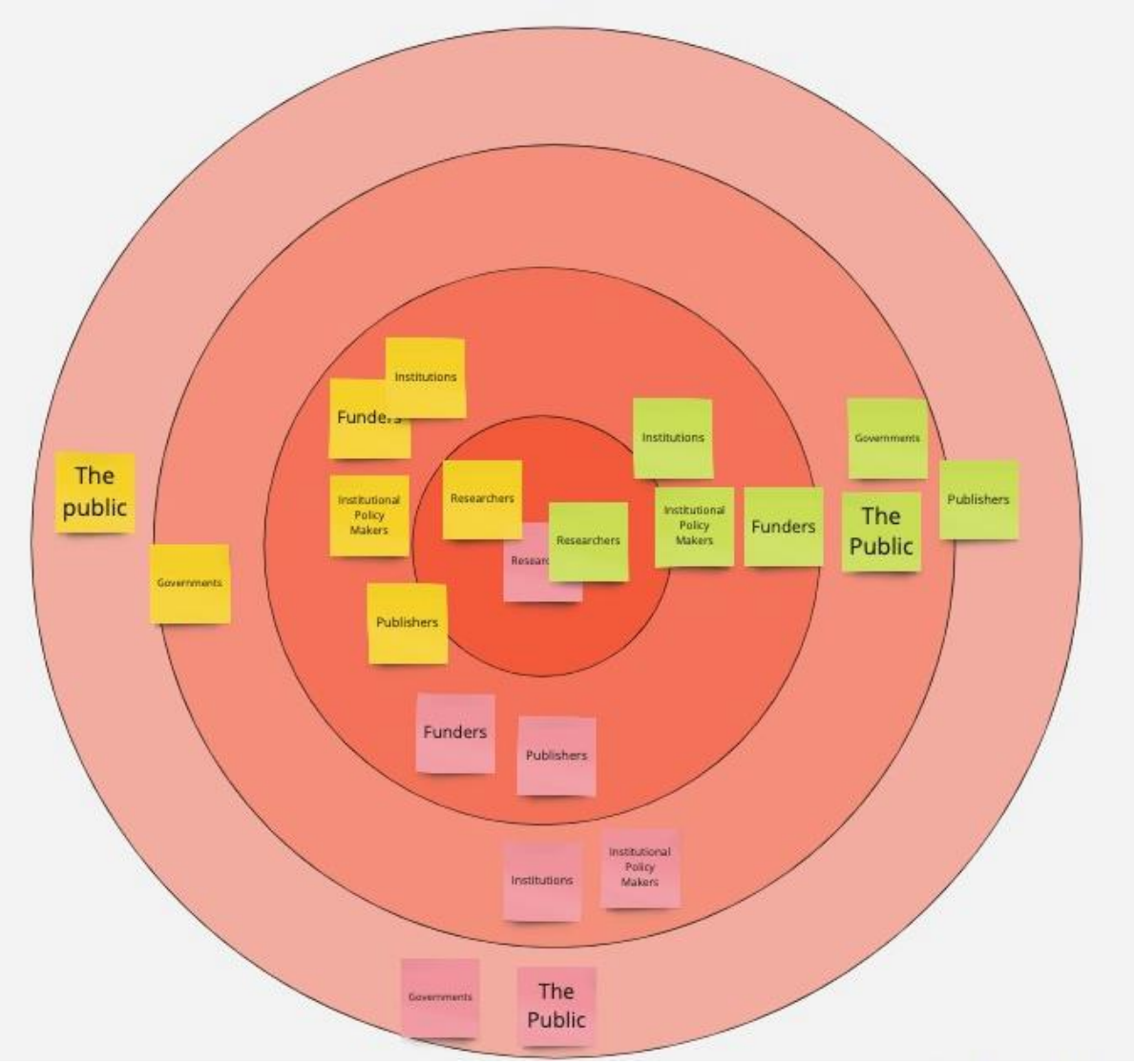

Figure C1: Results of the stakeholder mapping exercise for the funders.

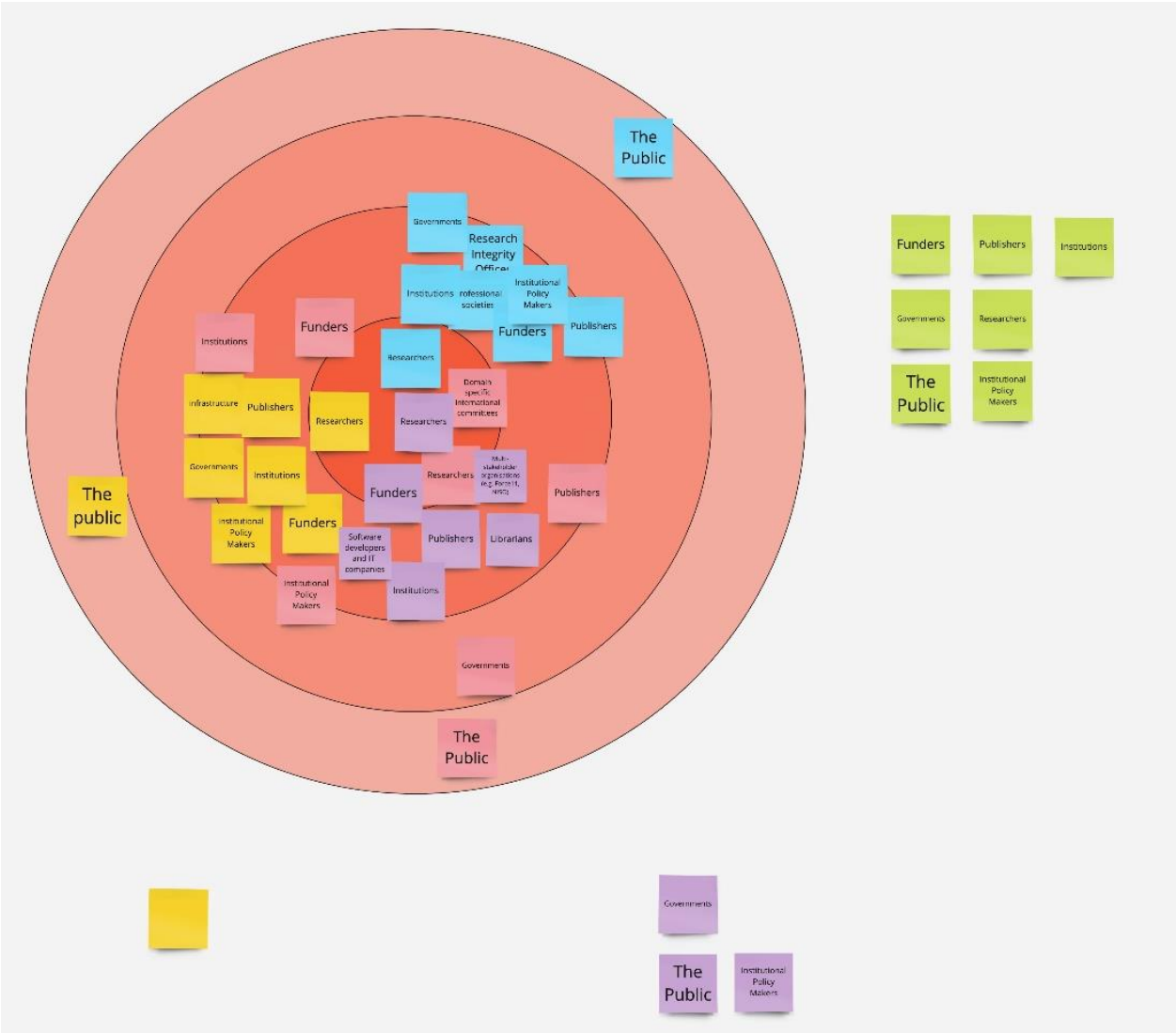

Figure C2: Results of the stakeholder mapping exercise for the publishers.

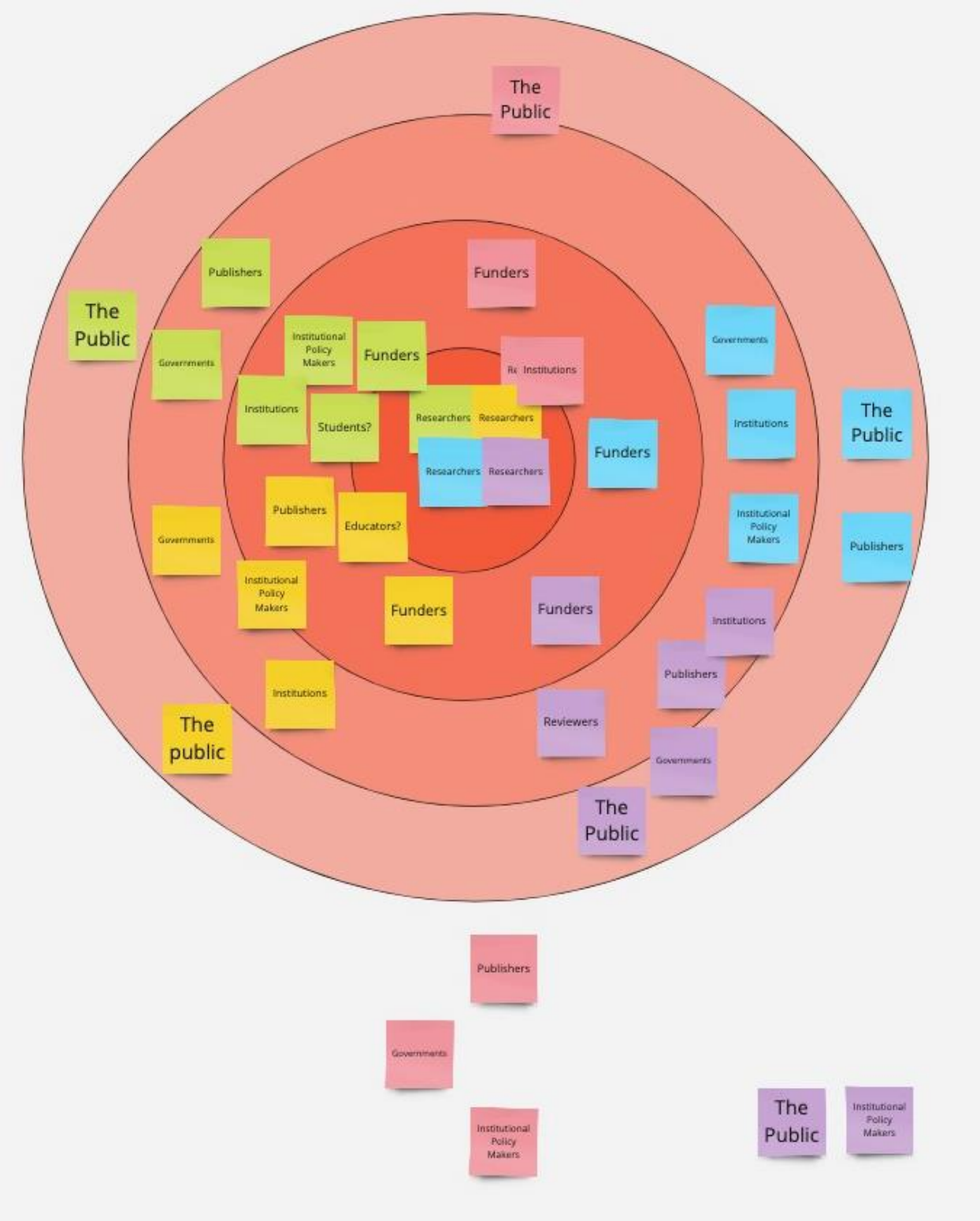

Figure C3: Results of the stakeholder mapping exercise for the qualitative researchers.

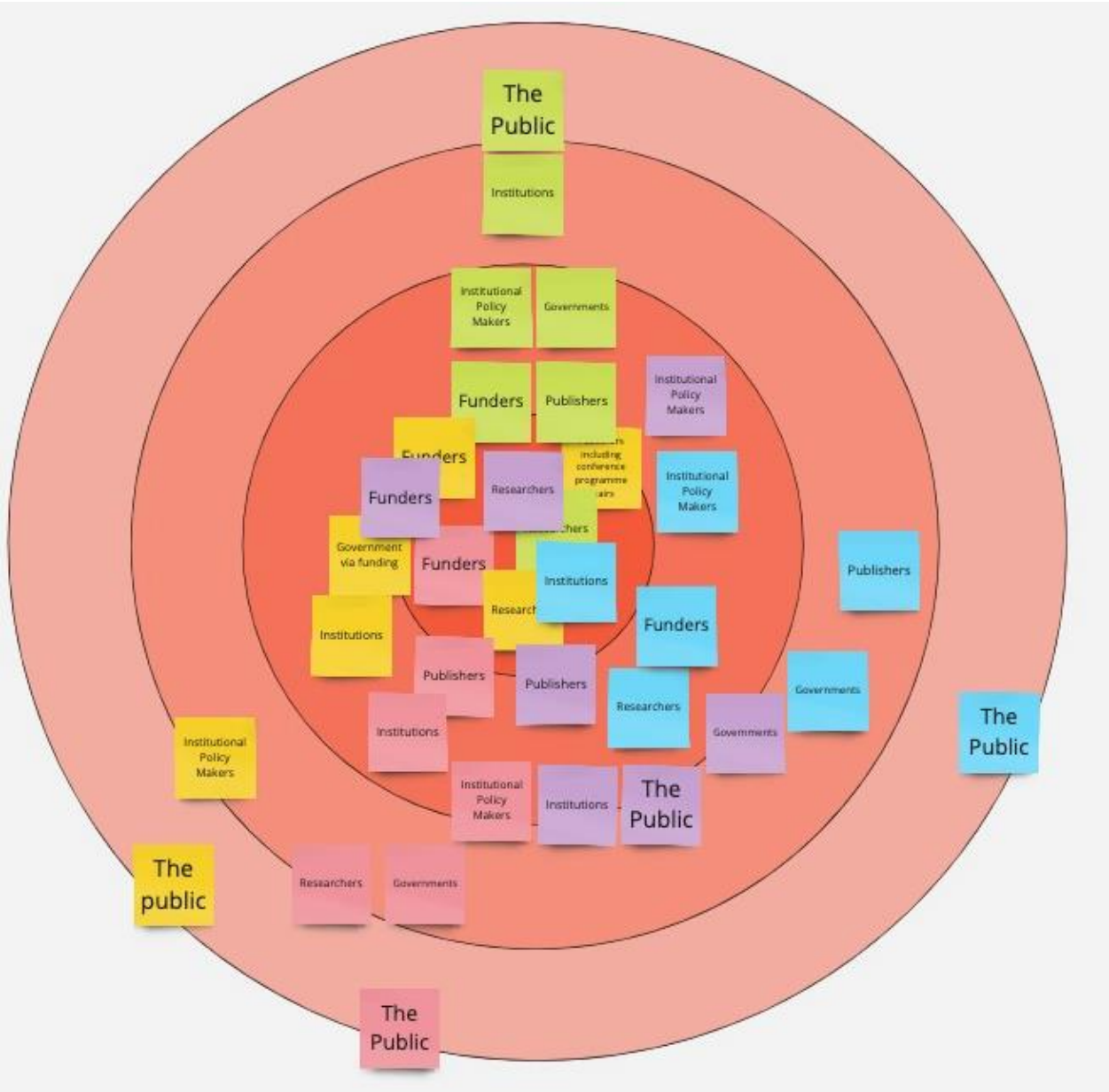

Figure C4: Results of the stakeholder mapping exercise for the machine learning researchers.

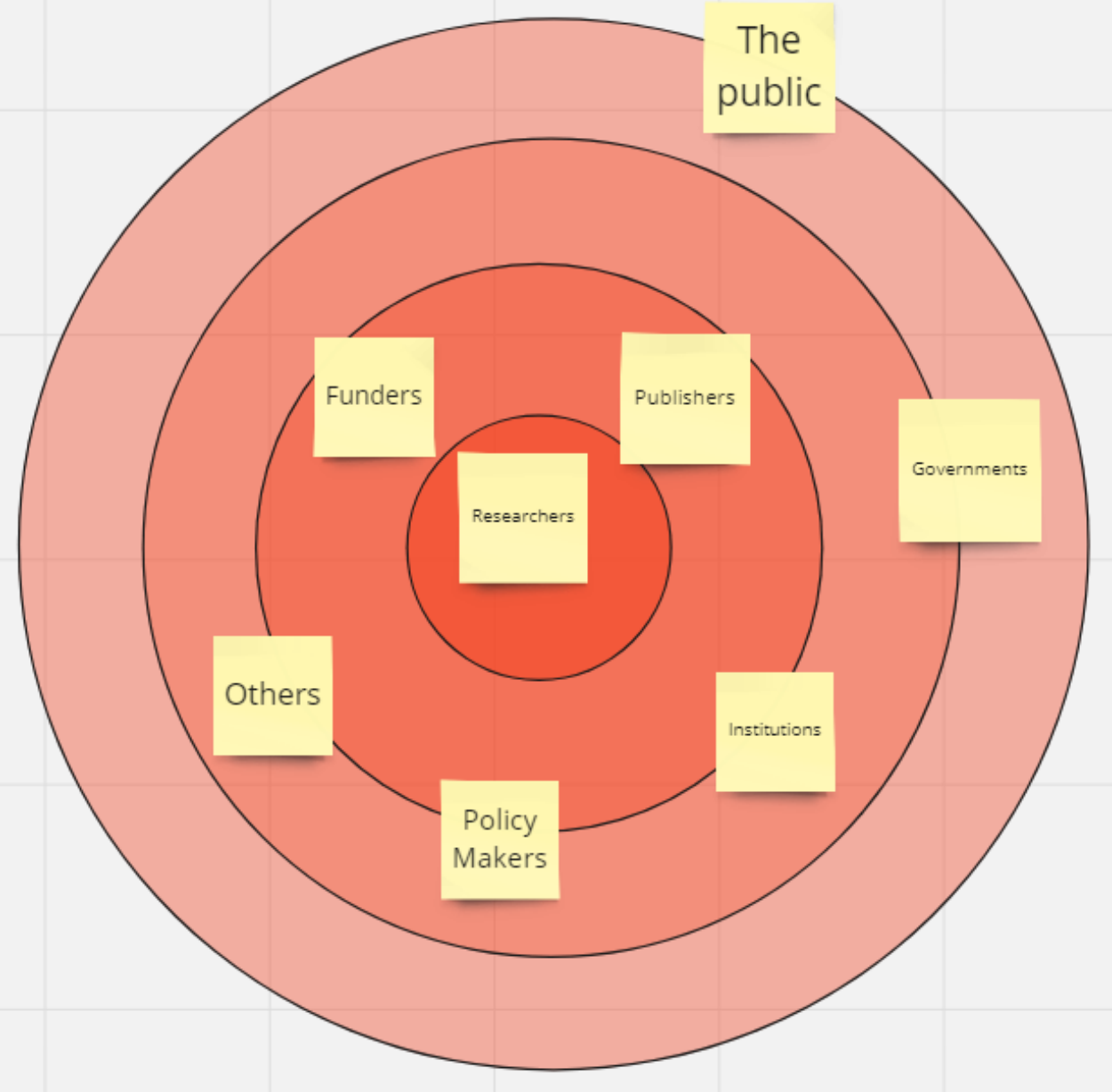

Figure C5: Stakeholder mapping results after analysis for all four workshops.

Appendix D: Specific Role of Stakeholders

The majority of enablers were mentioned in relation to researchers (91), funders (85), and publishers (67). Then institutions (37), infrastructure (22), and policymakers (18), scientific societies and associations (14), and governments (12).

Researchers were considered responsible for cultural/social enablers (35) and they were discussed in relation to policy (12), technological (13), training/education (9), institutional (8), and financial/economic (7) enablers.

Funders were considered responsible for policy enablers (20) and financial/economic (16) enablers, but also cultural/social (13) enablers, institutional (13) and technological (12).

Publishers were considered responsible for policy (22) enablers, technological (13) and cultural/social (12); also systemic (7).

Institutions were considered responsible for institutional (10), policy (8), and training/educational (7) enablers.

Policymakers were primarily associated with policy enablers (7).

Table C1: Enablers of reproducibility by stakeholder, NVivo Matrix Query export

	Cultural/social	Financial/economic	Institutional	Policy	Systemic	Technological tools/infrastructure	Training/education	Totals
Researchers	35	7	8	12	4	13	9	88
Funders	13	16	13	20	6	12	3	83
Publishers	12	4	2	22	7	13	3	63
Institutions	4	3	10	8	2	2	7	36
Infrastructure and service providers	2	0	2	1	0	17	0	22
Policymakers	2	1	2	7	2	0	3	17
Scientific societies and associations	6	1	0	5	0	1	1	14
Governments	2	1	1	5	2	0	1	12
The public	2	0	0	1	0	1	1	5
Totals	78	33	38	81	23	59	28	

The greatest number of mentions of barriers were linked to researchers (48), far more than for publishers (27), institutions (18), and funders (16). Few mentions of barriers were associated with policymakers (6), governments (5), infrastructure (5), support staff (5), and the public (5).

For researchers, the greatest numbers of mentions were for cultural/social (20), with some also for financial (5), institutional (5), policy (5), systemic (5), technological (4), and training (3).

For publishers, the greatest number of mentions pertained to policy (8) and cultural/social (7), with some mentions of financial, systemic, technological, and institutional.

For institutions, the greatest number of mentions were for cultural/social (5), policy (4), and systemic (4).

For funders, the greatest number of mentions were for cultural/social (4) and policy (4), then financial (3) and systemic (3).

Table C2: Barriers to reproducibility by stakeholder, NVivo Matrix Query export

	Cultural/social	Financial/economic	Institutional	Policy	Systemic	Technological tools/infrastructure	Training/education	Total
Researchers	20	5	5	5	5	4	3	47
Funders	4	3	1	4	3	1	0	16
Publishers	7	3	2	8	3	3	1	27
Institutions	5	1	2	4	4	0	2	18
Infrastructure and service providers	1	0	0	0	0	4	0	5
Policymakers	1	1	0	2	2	0	0	6
Scientific societies and associations	1	0	1	0	0	0	0	2
Governments	2	0	0	2	1	0	0	5
The public	2	0	0	2	1	0	0	5
Totals	43	13	13	28	19	13	8	