How to get there from here? Barriers and enablers on the road towards reproducibility in research

Authors

Serge P.J.M. Horbach[1], Nicki Lisa Cole[2*], Simone Kopeinik[3], Barbara Leitner[4], Tony Ross-Hellauer[2], Joeri Tijdink[4]

1 – Institute for Science in Society, Radboud University, Nijmegen, the Netherlands

2 – Open and Reproducible Research Group, Know Center Research, GmbH, Austria

3 - Fair AI, Know Center Research GmbH, Graz, Austria

4- Department of Ethics, Law and Humanities, Amsterdam University Medical Centers, Vrije Universiteit, Amsterdam, the Netherlands

* Corresponding author: ncole@know-center.at

Abstract

Reproducibility of research is a hotly debated topic, including aspects like causes and consequences of low levels of reproducibility, and the desirability of reproducibility across research fields. While some research fields have led the way and introduced various reproducibility practices and procedures, the call for efforts to 'improve' reproducibility in research has not come without criticisms. Generally, there is ample ambiguity about the desired and most productive way to make progress with reproducibility of research. To study the most desirable way towards this progress, the current study uses a future studies methodology to gather perceptions of developments in the research ecosystem related to reproducibility issues. It draws on input from representatives of four main stakeholder categories: scholarly publishers, funding agencies, qualitative social scientists and machine learning researchers. Particularly, it discusses the enablers and barriers that members of these stakeholder communities foresee on the road towards a research ecosystem that is more conducive to reproducibility.

Based on an initial survey and a series of workshops, the study finds that enablers and barriers can be categorised into five main clusters. The factors most prominently mentioned as potentially supporting or hindering a desired future are located within research culture, including norms, values and shared definitions; and in the infrastructure required to engage in reproducibility practices, including repositories, support staff, and digital infrastructure. Three other clusters of factors put forth by participants relate to policy efforts required to incentivise reproducibility practices; training and education to empower researchers and support staff to engage in reproducibility practices; and the financial resources required to facilitate the transition towards a desired future and to specifically fund replication studies. This manuscript also identifies several tensions between enablers and barriers perceived by diverse stakeholders and concludes with recommendations for addressing these.

1. Introduction

Reproducibility, or more specifically, concerns over the lack of it, have recently received ample attention from the research community and wider public alike (1). The concept of reproducibility, though interpreted in various ways, generally implies the scientific community's ability to achieve results that are either identical or similar to a study's original findings by repeating the research methods or analyses (2). It has been a subject of increasing concern across several fields, particularly in behavioural and medical sciences, due to perceived little success in replication studies (3). Researchers have suggested or identified various factors that are or might be contributing to low levels of reproducibility. These include insufficient clarity in reporting procedures, data, and analysis methods; a tendency to publish primarily positive outcomes; and problematic research practices (1,4,5).

Like the causes of low levels of reproducibility, its consequences are equally debated. Some view the lack of reproducibility as a major threat to the scientific process's self-corrective nature, research efficiency, and public trust in research findings (6,7). However, the relevance of reproducibility varies greatly depending on the research context and the type of research at hand (5,8,9). In some contexts, lower levels of reproducibility are not always deemed problematic, and reproducibility is not always considered a major concern (for example, this is often the view regarding qualitative research (10–13)). Factors influencing the applicability and desirability of reproducibility include the resources required for reproducibility, standardization levels within the research field, the philosophical underpinnings of the research, the subject matter being studied, as well as ethical and legal limitations on data sharing and transparency (8). In particular, there have been lively debates within the interpretative and qualitative social sciences and humanities about whether efforts to increase standards of reproducibility are relevant or even desirable and feasible for their fields of research (14–17).

1.1. Improving reproducibility

In the quest to improve levels of reproducibility, various interventions have been suggested and implemented across research communities. Most of these interventions have originated from the medical and behavioural sciences (18) and some have gained traction in various other disciplines. One significant intervention category is open methodology, where researchers are encouraged to make their research processes transparent. This includes detailing experimental procedures, analytical methods, and decision-making processes. The rationale is that such openness allows other researchers to understand, evaluate, and replicate the studies more effectively (19,20). For instance, various registries and platforms (e.g. the Open Science Framework) offer researchers tools to preregister study designs and methodologies, promoting transparency from the outset. This approach has been gaining traction, as evidenced by an increasing number of journals and funding bodies that now require or strongly recommend open methodology practices (21,22). Considering the debate about different reproducibility practices in diverse research traditions, efforts have been made to tailor preregistration formats to different ways of conducting research

(e.g., 23). Initial studies assessing the effectiveness of preregistration in terms of fostering reproducibility show small but positive effects (e.g., 24–27). Yet there remain concerns and questions about the validity of this intervention across diverse epistemic contexts (23,28–30).

Another critical intervention lies in the domain of open data and reporting standards. The movement towards open data involves making research data publicly available, thus enabling other researchers to verify results, conduct new analyses, or combine datasets for more robust findings. Initiatives like the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles have been instrumental in guiding researchers towards more accessible and reusable data practices, usually through data sharing policies by funders and journals. Additionally, the adoption of stringent reporting standards, like the CONSORT guidelines for clinical trials (31), the ARRIVE guidelines for animal research (32), or the PRISMA guidelines for reviews (33), aspires to set minimal standards for the level of methodological detail conveyed in published research, to allow for replication and evaluation. These guidelines are often endorsed or mandated by journals. Studies find mixed but mostly positive effects of these kinds of guidelines on reporting standards (e.g., 34–36). While most of these initiatives, again, originate in the (bio-)medical and behavioural sciences, some guidance also exists on reporting qualitative and humanities data. However, despite these interventions, studies indicate that the prevalence of open data practices remains low and debates about the desirability and usefulness of sharing qualitative data are ongoing (e.g., 37–40).

Furthermore, interventions in the form of improved infrastructures and open evaluation processes are reshaping research practices. Digital infrastructures like research data repositories and collaborative platforms aim to facilitate easier data sharing and collaboration. Many such infrastructures exist, some particularly tailored to the needs of specific research communities (e.g., 41). Open evaluation, including transparent peer review processes, has the potential to increase levels of scrutiny of research (42). This openness in the evaluation process aims to increase accountability and improve the quality of published research. Despite these advancements, the effectiveness of these interventions in terms of reproducibility remains largely unknown (43).

Lastly, several stakeholders have experimented with symbolic rewards, for example in the form of tokens or badges, that can be earned for acts considered to be good research practice (44). Such interventions, including badges for implementing practices like data sharing, code sharing, or pre-registration, aim at steering community norms and individual incentives towards higher engagement with reproducible research practices. Studies evaluating the effectiveness of these symbolic rewards to increase the likelihood of researchers engaging in the practices they are rewarded for, give mixed results. Indeed, some studies found positive effects on data sharing (45), but others found null effects on data sharing (46) and on code sharing (47). The effects of such incentive practices on reproducibility seem to be minimal, as evidenced by a study finding low levels of reproducibility, even among studies receiving badges (48,49).

As mentioned before, the call for efforts to 'improve' reproducibility in research has not come without criticisms. Both the initial identification of issues, quickly labelled as a 'reproducibility crisis', and the implementation of proposed solutions described above, have originated in certain disciplines, with their own characteristics, methodological procedures and normative frameworks (9). Scholars from other disciplines and scholarly traditions have repeatedly questioned the desirability of establishing universal standards and employing blanket strategies in relation to reproducibility. Criticism originated most notably from the qualitative social sciences (e.g., 15,50) and the humanities (e.g., 8,51,52). The latter critiques were voiced as a direct response to an explicit pledge for the desirability of reproducibility in the humanities by Peels and Bouter (6). Opponents of universal standards for reproducibility argue that the appropriateness and the meaning of reproducibility starkly depend on the context and unique conditions of a study (5). These conditions may vary even within disciplines and are closely related to ways of producing and certifying knowledge (9). Consequently, diverse research and research-supporting communities may have divergent understandings of the notion and appropriate standards of reproducibility, as well as ways of achieving it. This consequently calls for a community-driven approach towards understanding the specific enablers and barriers towards desired states of reproducibility in research.

1.2 Aims and objectives

The study described in this article uses a future studies methodology to gather perceptions regarding developments in the research ecosystem in terms of reproducibility issues. Particularly, it draws on input from representatives of four main stakeholder categories: scholarly publishers, funding agencies, qualitative social scientists and machine learning (ML) researchers. It aims to address the question: **what enablers and barriers do members of these stakeholder communities foresee on the road towards a research ecosystem that is conducive to reproducibility?**

This study was part of the TIER2 project (53). TIER2 focuses on enhancing reproducibility in research. Its objectives include advancing understanding of reproducibility, developing specialized tools, fostering community engagement, and executing policy interventions in diverse research contexts. It ultimately aims to improve the reusability and overall quality of research outcomes. A key aspect of the project concerns its investigation of the optimal adaptation of these interventions to the unique epistemic, cultural, and sociotechnical nuances present in various research contexts. The project centres on co-creative methods to achieve this goal. This study was preregistered, including a detailed description of its methodology (54).

This paper empirically contributes to the literature on reproducibility in two ways. It first provides an overview of what the participants of our future studies workshops consider desirable future states of academia in relation to the reproducibility debate. Second, we describe the enablers and barriers that our participants envision on the road towards these imagined futures. These contributions inform the debate on reproducibility and practically inform researchers,

policymakers and other stakeholders on how to transform academic practices and infrastructures to enable community-desired modes of working and foster reproducibility practices.

2. Methodology and data

2.1. Exploring the Future of Reproducibility through Futures Studies

Our study employs a futures studies methodology (55), a systematic approach to exploring possible, probable, and preferable futures. This method aims to envision the future by creating alternative scenarios that inform strategies for shaping desired outcomes at various levels, from individual to global (56,57). By envisioning both desired and undesired futures, futures studies challenge existing frameworks and assumptions, providing a foundation for rethinking the steps necessary to realize or avoid these potential futures. This study was approved by the Amsterdam UMC ethical review board on 9th August 2023.

We provide a condensed version of the methodological procedure here. Greater detail is provided in the preregistered protocol, cited above.

2.2. Methodological Approach: Workshops and pre-workshop survey

At the core of our method are scenario workshops, designed to explore the future of reproducibility over a ten-year horizon. We engaged four distinct stakeholder categories: machine learning researchers, qualitative social science researchers, research funders, and scholarly publishers. Funders and publishers were selected as core stakeholders of the TIER2 co-creation community, offering direct interventions to support reproducibility. Machine learning and qualitative social science researchers were selected as opposing ends of the research methodology spectrum, bringing epistemic diversity to our study. Participants were recruited based on the following inclusion criteria: people should be working in the specific stakeholder group targeted for the specific workshop 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.

This future study process began with a pre-workshop survey, allowing participants to familiarize themselves with the topic of reproducibility, the workshop's objectives, and the exercises planned (see Supplement 1 for survey questions). This survey collected demographic data and participants' perspectives on the current and future states of reproducibility in their fields or profession, feeding this data into the workshops.

Each workshop, conducted online in May and June 2023 over 3.5 hours using Miro and Zoom platforms, was divided into three substantive sessions and facilitated by Joeri Tijdink, PhD, MD (male), with various members of the authorship team present to support each workshop. The main researcher is trained as a psychiatrist and meta-researcher, and has ample experience in moderating workshops and focus groups. These sessions included a variety of exercises structured around a diverge-converge model, encouraging the individual or pair-wise generation of numerous ideas, followed by group reflection, and subsequent selection of the most promising scenarios for further development. Exercises ranged from mapping current reproducibility practices to scenario planning, where participants envisioned preferred, dystopian, likely, and outlier scenarios for reproducibility. See Supplements 2 and 3 for our workshop facilitation guide and template Miro board.

The workshops concluded with a backcasting exercise, where participants outlined necessary steps and identified potential enablers and barriers for achieving specific parts of their preferred future scenarios. This exercise was crucial for creating a roadmap to guide actions towards realising the envisioned future of reproducibility. Overall, the workshops employed participative and co-creative methods, aimed to gather diverse insights and collaboratively develop comprehensive scenarios, and contributed to a deeper understanding of the (possible) future landscape of reproducibility in research. The workshop series were concluded with a validation workshop, inviting members of each of the workshops to a final session in which initial findings were shared and adjusted based on participant feedback.

2.3. Recruitment and sample

A pool of potential participants was generated by the research team based on their existing professional networks and the stakeholders already involved in the TIER2 project, within which this study was carried out. Our aim was to generate a sample that was diverse in terms of gender, age/experience, and geography. We recruited participants using a template email initially sent by a member of the team affiliated or networked with the stakeholder group in question (see Supplement 4). Out of 67 people invited, we were able to recruit a sample of 19 participants (see Table 1). Despite our efforts in recruiting a gender-balanced sample, our sample included more men than women (12 vs. 9). It was diverse in terms of age and experience, and in terms of current geographic affiliation, with 9 countries represented overall (all within Europe, expect for the United States).

Table 1: Participant demographics

	Funders	Qualitative researchers	Publishers	Machine learning	Totals
Number of participants	4	5	5	5	**19**
Countries represented	3	5	3	3	**9**
Men	3	2	4	3	**12**

Women	1	3	1	2	**7**

Participants were presented with an information letter prior to granting informed consent to participate in this study (see Supplement 5).

2.4. Data and analysis

All workshop audio and video were recorded in Zoom and transcribed by an automated transcription tool (Amberscript) and checked for accuracy by members of the team. In addition to these transcriptions, our data comprises the responses collected through the pre-workshop survey and information created on the Miro boards during the workshops. All these materials were compiled in NVivo for coding and analysis. Data were first organized in NVivo using a flexible coding strategy (58) to facilitate coding and analysis within and across discrete aspects of the workshops by author NLC (as defined and illustrated in Supplements 2 and 3). We then deductively and collaboratively created a code system based on our organization of the workshops and existing knowledge of their contents (see Supplement 6 for our deductive code system). Then, each member of the team carried out coding using this system to an assigned section of the workshops. Members had the freedom to inductively create codes as themes emerged through this process. All data coded by individual team members were merged and four members (NLC, SH, SK and TRH) were then assigned specific themes to focus on for analysing, synthesising and reporting data for the purposes of this paper.

3. Results

3.1. The desired end point and responsibility for it

In the initial exercises of the workshop, almost all participants reported that researchers are the most influential stakeholders shaping the future of reproducibility. They agreed that what happens in terms of reproducibility is ultimately up to researchers, because it is they who manifest cultures of research. 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: "If researchers don't engage with reproducibility, then we will achieve nothing."

Most participants agreed that funders are the second most influential stakeholder group, placing them centrally, primarily 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 two groups 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. Indeed, we observed during our validation workshop, when presented with the results of this exercise, that some participants asserted that funders should be considered equally influential to researchers because of their power to influence how researchers behave.

Other stakeholder groups, which were largely considered to be secondarily influential, alongside funders, include research institutions and institutional policymakers, and (added by participants) 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, 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.

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.

In the scenario planning exercise, participants described their preferred futures of reproducibility as dynamic and identified various characteristics of them, which can be grouped under four main themes: culture, standardisation, incentives, and infrastructure. Though we distinguish them here for analytical purposes, participants stressed that they are intertwined and interdependent, while new issues will arise with time and progress. According to our participants, a preferred future has (some of) the following characteristics (see Table 2).

Table 2: Four themes of preferred reproducibility futures, with examples

Culture	• Research quality more important than quantity • Reproducible practices are centred • Reproducibility part of core training for researchers
Standardisation	• Minimum reproducibility requirements are present among funders, publishers and other stakeholders • Standardised and shareable methods and statistical software with detailed instructions are available to research communities • Guidelines account for epistemic and methodological diversity
Incentives	• Incentives are present to foster reproducible and open research practices • Recognition and visibility of reproducible actions • Alternative research outputs are rewarded • Collaboration is fostered

	• Reproducible and open research practices are supported by funding
Infrastructure	• Guidelines and policies for using infrastructures are clear • Open-source tools that support reproducible and open research are widely available

The stakeholders' preferred futures showed significant similarities but also some differences, particularly in the definition of reproducibility. Qualitative researchers emphasized the need for enough variability in qualitative research methods, highlighting that "one size fits all" does not apply to their preferred futures. Machine learning researchers shared similar concerns for computational reproducibility. Publishers and funders, in contrast, focused on the role of standards, guidelines and incentives in fostering reproducible research.

3.2. Enablers and Barriers

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 preferred futures and reproducibility more generally. Research culture is conceived of as 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.

3.2.1. Cultural and social enablers and barriers

3.2.1.1. *Research culture: norms, values and practices*

In terms of how it affects reproducibility, our participants framed culture as both the ‘place’ where reproducibility happens (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 reproducible research and workflows to become mainstream, normative and standardised (as much as is possible).

Respondents illuminated the enabling role of research culture by pointing out how reproducibility practices sometimes spill over from one area of research to another through publication trends or collaboration. For example, a researcher in our machine learning workshop said:

> *But it's true that 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, Scenario Planning)*

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

> *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 incentivize me to do different things, or in order to be part of something, some sort of larger project [one has to adopt reproducibility practices]. (Qualitative researcher, 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 machine learning researcher suggested those immersed in such issues may exist in a "bubble", while others remain relatively unaware, pointing to the existence of differing and not overlapping research cultures.

Reflecting a desired shift in research culture, participants in our funders' workshop imagined that, in their ideal future of reproducibility, it would be both mainstreamed and normalised. They imagined that reproducibility will be "the new normal", where reproducibility is practiced both "by default" and "by design". A qualitative researcher imagined that researchers are doing reproducible research because "all their peers are choosing to use reproducible methods and approaches", highlighting the potential for research culture to act as an enabler for reproducibility. 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. (Backcasting)*

In contrast, a machine learning researcher suggested that aspects of research culture that are extrinsic could play an important role:

> *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 reproducibility track [of a conference] 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. (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 framed this as a paradigm shift to "slow science". One said:

> *And then this is very crucial. There should also be something of a paradigm shift towards slow science as the 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 [reproducibility] is all over the place just because people really, really enjoy it and it brings them good science. (Backcasting)*

3.2.1.2. *Institutional culture: norms, values and practices*

Participants observed that for reproducibility to be normalised, it must be embedded in the cultures of institutions and assessment. This differs from how culture was discussed in the previous section, in that the former refers to research cultures made up of communities of peers within a field, discipline or research area, while the latter refers to the culture 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 interact. Maybe owing to the absence of participants directly responsible for institutional cultures of reproducibility, we observed participants mainly voicing barriers rather than enablers in relation to this theme.

Referring to the importance of institutional culture, a funder said, "Because we deem that reproducibility is important, institutions and policymakers also need to embed that in their culture saying, 'You know, this is important to us'." While, in the same workshop, another 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. In a workshop, a publisher reported 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 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. (Scenario Planning)*

Participants in the qualitative researchers workshop also discussed this point but turned responsibility for it back onto funders. Imagining 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), 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, an 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 rigour on these researchers, they will not enjoy research and start doing 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.

3.2.1.3. *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 survey an ML 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?*

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... (Survey Results)*

Within qualitative research the very language of reproducibility was an 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.*

This discussion suggests that standardising one particular definition of reproducibility would be neither appropriate nor effective as an enabler, however, standardising other things, like epistemically diverse practices that enable reproducibility, and what is offered and expected by infrastructure and services, would be helpful.

In our workshop with qualitative researchers, one spoke in depth about the need for this kind of 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. (Scenario Planning)*

3.2.1.4. Epistemic diversity

Building on the results reported in the previous subsection, our participants agreed that standards should be field-specific and context-sensitive for them to be appropriate and effective. 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. (Scenario Planning)*

Similarly, a qualitative researcher discussed 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. (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 [Sabina] 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 traditions of qualitative research - although note there are some more realist perspectives for whom the concept has some grounding.*

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.*

Therefore, for reproducibility reforms to be relevant and valued across diverse research epistemologies, it is important what the term means, how it is used, and to recognize that other terms might be more applicable and resonant for some.

3.2.1.5. Leaders in change

Our participants envision preferred futures of reproducibility will be enabled by community-driven standards that reflect and respect epistemic diversity. Some conceived of this in a more grass-roots fashion, while others suggested that a top-down approach, led by professional associations, groups of international experts and multi-stakeholder organisations should be pursued. Reflecting the 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, even if you make a paper, that has a limited impact, right? So, I think the community or maybe [professional organisation] or maybe 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 calls for papers or reviewer criteria. (Backcasting)*

A funder also spoke to the importance of research communities as leaders in spreading reproducibility norms and practices (often Open Science practices) when describing the function of reproducibility networks. They said, "We recently funded provided funding for a national reproducibility network in [country]. We think that's a very promising instrument to raise awareness and bring together people working on this topic. We've been very much inspired by network[s] that already exist in, for instance, Germany and the UK, I think. So yeah, I think having those kinds of networks and ensuring that researchers can work on that together [is important]" (Scenario Planning).

Reflecting the idea that professional societies should have a leading role in defining standards, a publisher said during the stakeholder mapping exercise:

> *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. (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. One publisher suggested, during stakeholder mapping, that, when international committees of academics/experts define and set standards, these become the norm within the relevant communities, and these will often be taken up and enforced by publishers, hence flagging them as a potential enabler for reproducibility.

3.2.2. Systemic and policy issues

A second factor discussed in all workshops concerns the implementation of adequate policy to foster reproducibility. 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 actually to enforce certain practices from actors that depend on them. Participants gave various examples in which they "only made progress" when certain policies were mandated, or projected expectations about what would happen if certain powerful actors would start enforcing or demanding reproducibility practices.

> *"... 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)*

Although generally agreed upon for their efficacy and necessity, mandates spurred debate due to the need for universal standards among all stakeholders to prevent irrelevant or infeasible tasks. Yet, permitting exceptions undermines the policy's effectiveness. This tension became evident in the ML researchers' workshop backcasting exercise, highlighting policy's dual role as both enabler and barrier in achieving reproducible futures

> *A: Reject if the paper does not have an appendix documentation.*
> *B: Yeah.*
> *A: Maybe that's too hard, right? I think not all research needs to be reproducible.*
> *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.*

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). Participants stressed that this speaks for stakeholders to act in collectives, rather than as individual entities. They suggested, for example, that associations of publishers, rather than individual publishers, are better placed to act.

As it is recognised that one policy usually requires several stakeholders to cooperate to effectively implement it, participants also suggested cooperation across stakeholder categories:

> *The scenario that we were thinking about is the kind of minimum publication standard requirement for publishers. [...] the standards are shared across communities and implemented and supported by funders, institutions and checked and reported on by publishers. (Publisher, Backcasting)*

A recurring theme in relation to effective implementation of policies related to the establishment of adequate incentive structures. Participants across the four workshops saw the need to integrate reproducibility practices into the full system and lifecycle 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)

I think the thing that needs fixing is the whole incentive structure. [...] in the future, I think research assessment should be based on the quality of the research rather than the impact. And it should be judged on whether the methodology is good. Has there been pre-registration? Things that actually help improve trust in science. (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.

In one workshop, a participant expressed a concern about the effectiveness of institutions as facilitating actors, though. This participant doubted whether they are capable of moving things forward, thereby questioning institutions' role as potential enablers of reproducibility:

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 doesn't 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)

Participants note that there is a need for some level of standardization across contexts to remove barriers like 'first mover risk'. However, as we mentioned before, they also recognised a risk of moving too much into a 'one-size-fits-all model', which comes with its own limitations and is considered an important barrier for widespread uptake of diverse reproducibility practices. Institutionalization of reproducibility norms on a discipline level was by some considered to provide a suitable balance.

3.2.3. Technological and infrastructural enablers and barriers

Participants discussed the significance of suitable infrastructures and technological advancements as enablers of reproducibility practices and the absence thereof as barriers. They agreed upon and emphasized the essential nature of having proper infrastructure for reproducibility in place. Some mentioned the need for reproducibility tools and public infrastructure, while not specifying concrete features. Others mentioned the importance of data sharing and curation, highlighting the need for additional resources to help store, share and reuse data. The level of detail provided in discussions seemed to relate to the degree of participants' engagement in developing or using specific infrastructures or tools, with those more actively involved in such processes, more likely to provide detailed accounts of how technological and infrastructural elements could act as either enablers or barriers of reproducibility practices.

Infrastructures or tools to share and reuse data were discussed most. Participants in all workshops mentioned that these facilities are required to promote reproducibility. However, participants in multiple workshops also mentioned that infrastructures for data sharing and reusing on their own are insufficient. For example, in the publisher workshop, one respondent commented on connecting research items (e.g., data, papers, code) by linking platforms, tools, and repositories. A particular challenge named in this context was the missing possibility of linking between scientific platforms, partly attributed to a lack of agreements on meta-data standards and guidelines.

The need for instruments (tools, protocols, routines) supporting the whole workflow towards reproducible science was also discussed. Participants brought attention to challenges such as the lack of agreements on metadata standards and guidelines and the need for instruments supporting the entire workflow towards reproducible science. For example, a publisher said during a workshop:

> *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. (Stakeholder Mapping)*

Additionally, the discussion emphasized the need for publicly funded infrastructures to support reproducibility on an institutional level. Connecting to another type of barrier discussed previously, a lack of standardization was identified as a prominent issue affecting dissemination, communication, and evaluation methods and tools. In the pre-workshop survey, one qualitative researcher identified the problem this way: “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)”

Participants also pointed out that meaningful code sharing requires more than just having an infrastructure in place, emphasizing the need for appropriate reward mechanisms for checking the validity of code.

> *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, Backcasting)*

Finally, the discussion touched upon the improper use of AI technologies, its potential impact on the scientific landscape and its trustworthiness. Particularly in qualitative research, participants name the uptake of new methods as essential to counteract the high workload of scientists. However, one concern discussed among publishers, funders, and qualitative researchers is the (i) error-prone or (ii) improper (e.g., to fake results) application of AI technologies. Imagining a dystopian future for reproducibility, one qualitative researcher identified a barrier to reproducibility by expressing concern about "sloppy" use of AI technology to auto code qualitative data. For example, "you read through four [transcripts] and then you tell ChatGPT to do the rest of the coding for you." Noting that proprietary QDAS (Qualitative Data Analysis Software) already have auto-coding features, they continued, "It's terrible for qualitative research, but it's going to happen and it's going to be a mess to review and to disentangle."

We conclude this section with two observations. First, 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 of a so-called knowledge graph discussion by the publishers or a data stewards discussion in the funder workshop. Hence, participants put a certain level of responsibility on their shoulders by discussing the need for infrastructures that they can or should develop. Second, we note that infrastructure was mentioned not only 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 was 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 qualitative researchers.

3.2.4. Financial and economic factors

Our participants overwhelmingly framed financial and economic factors as important 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 the TIER2 project in which this research is situated, and its sister projects, iRise and OSIRIS); that they can specifically fund reproduction and replication

studies; and that they and other stakeholders can fund the work and initiatives that support reproducible research (like data stewards, trainers, etc.). When this is not funded, reproducibility 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" (Survey). An ML researcher 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. (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." While having financial implications, this point relates to wider incentive and assessment structures that participants recognized as potential enablers of or barriers to reproducibility practices.

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 ha[ve] 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."

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. (Scenario Planning)*

A qualitative researcher supported this approach 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 respond.

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. (Scenario Planning)*

However, one funder noted 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 and too short programmes could act as barriers to reproducibility.

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 participant. 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. 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.

3.2.5. Training and education

In addition to the previously discussed dimensions, training and educational activities are considered crucial for promoting and scaling reproducibility practices. Participants in all four workshops stated that formal training, mentorship, and role modelling are essential to "get policy

down to the individual researcher" and to "incorporate it in everyday research" (ML researcher, Backcasting), making "those things normal" (Publisher, Scenario Planning).

Participants suggest that training can address multiple aspects of research and involve various actors. The primary focus could be on early career researchers or students, teaching them technical skills, research methods and statistics, as well as Open Science skills (Publisher, Survey). Several participants experienced such training (either as trainer or trainee) and described how these are essential elements of a desired future:

> *[...] 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)*

Some participants took it a step further by suggesting that PhD students should be required to conduct a replication study to obtain their degree. Additionally, senior researchers, principal investigators, and others in various roles, such as evaluators at journals or funding agencies, should also receive training.

> *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)*

Especially among qualitative researchers, this was even extended to non-academic staff at universities, including support staff facilitating tasks related to open science e.g., data management and analysis. One participant suggested that efforts to teach about reproducibility could go beyond academia to raise public awareness about the importance of reproducibility. This aligns with traditional science literacy models.

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

Participants emphasized the need for training to go beyond early career researchers and students and to not only focus on research practices or data-related skills. They believe that training should lead to a cultural shift, where researchers understand that the additional effort required to make their studies reproducible will be rewarded.

On the other hand, some participants expressed concerns that improper implementation of role modelling or mentorship could hinder reproducibility. They highlighted the risk of early career researchers being influenced into a culture that normalizes cutting corners.

> *...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)*

In addition, participants discussed the need for resources to support these training and mentorship efforts. They suggested that funding agencies could provide dedicated schemes or resources for

these activities, possibly embedded within existing funding schemes for training reproducible practices.

> *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)*

4. Discussion and conclusion

This study set out to identify the enablers and barriers that members of four key stakeholder communities (scholarly publishers, funders, qualitative researchers, and machine learning researchers) foresee on the way toward a desired future state of the research ecosystem that fosters reproducibility. We found that enablers and barriers can be categorised into five main clusters. The factors most prominently mentioned as potentially supporting or hindering a desired future are located within research culture, including norms, values and shared definitions; and in the infrastructure required to engage in reproducibility practices, including repositories, support staff, and digital infrastructure required for sharing research materials. Three other clusters of factors put forth by participants relate to policy efforts required to incentivise reproducibility practices; training and education to empower researchers and support staff to engage in reproducibility practices; and the financial resources required to facilitate the transition towards a desired future and to specifically fund replication studies.

The future of reproducibility that participants imagine:

- Has a particular research culture that prioritizes quality over quantity and centres reproducibility in research practice and in training;
- Has standardised reproducibility requirements that account for methodological and epistemic diversity and standardised and shareable methods, tools and workflows;
- Incentivises reproducible, open and collaborative practices by providing recognition for them, funding them, and making them visible (these include alternative research outputs); and
- Has infrastructure that is designed for ease of use with clear guidance, policies and training, hosts FAIR and open tools and workflows, and sufficient resources are available to develop and maintain such infrastructure.

Our results generally align with previous studies that have assessed the implementation of open science practices (e.g., 59) and reproducibility (e.g., 60), indicating a need for a culture change and training for all actors involved to achieve the desired goal.

As mentioned before, several of the themes or topics discussed by the participants transgress the boundaries of the analytical categories used for our discussions. For example, research assessment and incentive structures relate to multiple of the categories described above. Moreover, some of the topics were highlighted both as enablers and potential barriers. We note a few tensions between and across the enablers and barriers identified by participants. First, there remains ambiguity about the

level of standardisation or flexibility that should be maintained in the pursuit of reproducibility. Both in terms of infrastructures and cultural aspects, discussions amongst participants exemplified the pros and cons of a common set of standards shared between researchers and other stakeholders in diverse disciplines and contexts, versus a more flexible approach, catered to the specific needs of diverse communities, potentially involving distinct approaches in different settings. Thus, echoing the work of Leonelli (8,61), our findings suggest great value in context-sensitive solutions and expectations that respect the diversity of research practices and epistemologies.

Linked to this are questions of ownership and collaboration: to what extent and on what scale should stakeholders join forces to address reproducibility standards? Participants recognised the need for cross-cutting approaches involving different stakeholders from different communities. However, they also recognised the risk of ignoring the specificities of individual research contexts if efforts are coordinated too centrally and debated the merits of bottom-up versus top-down innovations and standards. Therefore, the feasibility and desirability of developing standards, tools and guidelines in collaboration with different actors remained ambiguous. All in all, we echo our participants' view that researchers, with the many hats that they wear, should be in leading positions to develop these. That being said, guidance from institutional actors would be useful in setting basic standards and expectations, linking and syncing with infrastructures and services, centring epistemic diversity, and fostering community-driven initiatives. Researchers in their roles as members of reproducibility networks, referees in grant or manuscript review, or as members of institutional review boards, are well-positioned to contribute to this agenda. This could create a dialectic of providing structure and empowering creativity and specificity at the local level.

This relates to the discussions about the desirability of mandating reproducibility efforts, and if so, against which standards. Some participants forcefully argued that the only way to achieve higher levels of reproducibility was by mandating efforts to foster it, for example, as an explicit requirement in grant funding, journal publication or tenure processes. Others, however, were much more sceptical, arguing for the need to remain flexible. This sceptical view included maintaining the option to opt-out of reproducibility efforts and standards in cases where these are not deemed relevant or feasible. This finding supports recent trends in the literature, which frame reproducibility (and replicability) as neither universally applicable nor feasible across diverse epistemic contexts (8,9,17,62). Consequently, we recommend the development of guidelines for reproducibility (and/or transparency) practices tailored to specific domain, methodological and epistemic contexts.

The results of this study speak to broader, ongoing conversations in the literature about the importance of research culture, in either fostering positive outcomes regarding reproducibility, or inhibiting them (63). Our participants flagged, as have many others, that the fast-paced, time-poor nature of research, tied to funding cycles and the "publish-or-perish" mentality, often stands in the way of rigorous and reproducible research. For example, a study focused on the prevalence of questionable research practices (QRPs) found that publication pressure appears to increase the frequency at which researchers engage in QRPs (64). Reflecting research culture, evidence suggests that mentoring plays a role in shaping how junior researchers operate. The same study from Gopalakrishna et al. (64) found that "survival mentoring" (i.e. the practice of learning mentees how

to survive in academia) is associated with increased rates of QRPs, while "responsible mentoring" exerts a (weaker) influence in reducing the rate of QRPs.

Recognizing the connection between a hurried research culture and research rigour, integrity, quality and reproducibility, some of our participants called for a shift to "slow science" – a solution that has been posited by others in recent years (65–67). According to Frith (67), to do slow science, one would need to produce less but better research, orient to longer timescales and bigger horizons, shift to valuing quality over quantity, value collaboration, and teach and mentor in ways that reflect these values. This is echoed by some of our participants. Importantly, this view raises the issue of research assessment and its role in shaping research culture. As our participants pointed out, what is valued in assessment procedures – be they within research institutions, funders, or publishers – wields a heavy influence on the values, norms and practices of research cultures. If reproducible research practices are not valued in assessment, and the time that is required to implement them is not accounted for in metrics used for assessment purposes, then reproducible research will not be realised. Related to this are concerns around the time and financial cost of creating and sharing reproducible research, which, though extensive for all, are not born evenly (more so by early career researchers (68) and by women, in the case of data management (69)) with implications not just for inequity in the undervaluing of this work in assessment procedures, but also in terms of which researchers are better positioned to implement them, depending on the level of resources available at their home institutions (70).

This, therefore, calls for approaches to reproducibility that focus on the full research ecosystem and lifecycle, including assessment procedures as a prime lever for initiating transformation. Transforming assessment procedures arguably creates both the opportunity and motivation for scholars to engage in reproducibility. The Behaviour Change Wheel framework (71) suggests these are two of the three main influences on behaviour – together with capability, which can be fostered through relevant training and infrastructure.

In terms of the latter, we agree with our participants' view that initiatives are needed that focus on equipping researchers and support staff with the necessary skills to engage in reproducibility practices, starting from early-career stages. These must be community defined, driven and delivered to be effective. Echoing participants, we recommend that training begin within the research education system.

In sharing these conclusions and making these recommendations, we acknowledge the limitations of this study. With 19 participants across four stakeholder categories, the sample size per stakeholder group is relatively small. And, while diverse in some respects, our pool of participants was limited in terms of geographic, cultural, and gender representation. Participants were predominantly from Europe and the United States, with a higher representation of men than women. In addition, participants were selected for their expertise and experience in relation to reproducibility, thereby likely engaging those that feel reproducibility is important. This limited diversity likely affected the range of perspectives included in the study, particularly regarding how reproducibility challenges are perceived in different global research contexts. We further acknowledge that the

stakeholder workshops were conducted online and within a limited time frame, which may have influenced the depth and quality of the discussions.

Despite the limitations, our findings confirm other findings already established in the literature and add to it by synthesising perspectives from a diverse stakeholder community, leading to actionable recommendations to transition research cultures and infrastructures towards a future state of research that fosters reproducibility, where and when appropriate.

Acknowledgements

We are grateful to our study participants for dedicating the time and effort to engage in our study and providing invaluable insights. We also want to express our gratitude to Eva Kormann for their help in preparing this manuscript.

Funding

This work was supported by the project TIER2, funded by the European Union's Horizon Europe research and innovation program under grant agreement No. 101094817.

Supplementary material

Supplements referred to in this manuscript are available at: https://osf.io/n28sg/

References

1. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016 May 26;533(7604):452–4.

2. Plesser HE. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. Front Neuroinformatics [Internet]. 2018 Jan 18 [cited 2025 Feb 6];11. Available from: https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2017.00076/full

3. Klein RA, Ratliff KA, Vianello M, Adams Jr. RB, Bahník Š, Bernstein MJ, et al. Investigating variation in replicability: A "many labs" replication project. Soc Psychol. 2014;45(3):142–52.

4. Fanelli D, Costas R, Ioannidis JPA. Meta-assessment of bias in science. Proc Natl Acad Sci. 2017 Apr 4;114(14):3714–9.

5. Guttinger S. The limits of replicability. Eur J Philos Sci. 2020 Jan 15;10(2):10.

6. Peels R, Bouter L. The possibility and desirability of replication in the humanities. Palgrave Commun. 2018 Aug 7;4(1):1–4.

7. Sarewitz D. Saving Science [Internet]. The New Atlantis. 2016 [cited 2025 Feb 6]. Available from: https://www.thenewatlantis.com/publications/saving-science

8. Leonelli S. Rethinking Reproducibility as a Criterion for Research Quality. In: Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise [Internet]. Emerald Publishing Limited; 2018 [cited 2025 Feb 6]. p. 129–46. Available from: https://www.emerald.com/insight/content/doi/10.1108/s0743-41542018000036b009/full/html

9. Ulpts S, Schneider JW. Knowledge Production Modes: The Relevance and Feasibility of Reproducibility [Internet]. OSF; 2023 [cited 2025 Feb 6]. Available from: https://osf.io/ujnd9_v1

10. Reischer HN, Cowan HR. Quantity Over Quality? Reproducible Psychological Science from a Mixed Methods Perspective. Vazire S, Vazire S, editors. Collabra Psychol. 2020 Jan 1;6(1):26.

11. Tamminen KA, Poucher ZA. Open science in sport and exercise psychology: Review of current approaches and considerations for qualitative inquiry. Psychol Sport Exerc. 2018;36:17–28.

12. Davies D, Dodd J. Qualitative Research and the Question of Rigor. Qual Health Res. 2002 Mar 1;12:279–89.

13. Huma B, Joyce JB. 'One size doesn't fit all': Lessons from interaction analysis on tailoring Open Science practices to qualitative research. Br J Soc Psychol [Internet]. 2022 [cited 2023 Feb 16]; Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/bjso.12568

14. Bazzoli A. Open science and epistemic pluralism: A tale of many perils and some opportunities. Ind Organ Psychol. 2022 Dec;15(4):525–8.

15. Bennett EA. Open Science From a Qualitative, Feminist Perspective: Epistemological Dogmas and a Call for Critical Examination. Psychol Women Q. 2021 Dec 1;45(4):448–56.

16. Drummond C. Is the drive for reproducible science having a detrimental effect on what is published? Learn Publ. 2019;32(1):63–9.

17. Penders B, Holbrook JB, de Rijcke S. Rinse and Repeat: Understanding the Value of Replication across Different Ways of Knowing. Publications. 2019 Sep;7(3):52.

18. Fidler F, Wilcox J. Reproducibility of Scientific Results. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy [Internet]. Summer 2021. Metaphysics Research Lab, Stanford University; 2021 [cited 2023 Jul 13]. Available from: https://plato.stanford.edu/archives/sum2021/entries/scientific-reproducibility/

19. Fuller T, Peters J, Pearson M, Anderson R. Impact of the Transparent Reporting of Evaluations With Nonrandomized Designs Reporting Guideline: Ten Years On. Am J Public Health. 2014 Nov;104(11):e110–7.

20. Pownall M, Azevedo F, Konig L M, Slack H R, Evans T R, Flack Z, et al. Teaching open and reproducible scholarship: a critical review of the evidence base for current pedagogical methods and their outcomes. R Soc Open Sci. 2023;10(5):221255.

21. Giofre D, Boedker I, Cumming G, Rivella C, Tressoldi P. The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices: Five years later. Behav Res Methods. 2022;17:17.

22. Veroniki A A, Tsokani S, Zevgiti S, Pagkalidou I, Kontouli K M, Ambarcioglu P, et al. Do reporting guidelines have an impact? Empirical assessment of changes in reporting before and after the PRISMA extension statement for network meta-analysis. Syst Rev. 2021;10(1):246.

23. Haven TL, Errington TM, Gleditsch KS, Van Grootel L, Jacobs AM, Kern FG, et al. Preregistering Qualitative Research: A Delphi Study. Int J Qual Methods. 2020 Jan 1;19:160940692097641.

24. Schäfer T, Schwarz MA. The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. Front Psychol [Internet]. 2019 Apr 11 [cited 2025 Feb 6];10. Available from: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00813/full

25. Scheel AM, Schijen MRMJ, Lakens D. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. Adv Methods Pract Psychol Sci. 2021 Apr 1;4(2):25152459211007467.

26. Soderberg CK, Errington TM, Schiavone SR, Bottesini J, Thorn FS, Vazire S, et al. Initial evidence of research quality of registered reports compared with the standard publishing model. Nat Hum Behav. 2021 Aug;5(8):990–7.

27. Obels P, Lakens D, Coles NA, Gottfried J, Green SA. Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology. Adv Methods Pract Psychol Sci. 2020 Jun 1;3(2):229–37.

28. Branney PE, Brooks J, Kilby L, Newman K, Norris E, Pownall M, et al. Three steps to open science for qualitative research in psychology. Soc Personal Psychol Compass. 2023 Apr;17(4):e12728.

29. Evans TR, Branney P, Clements A, Hatton E. Improving evidence-based practice through preregistration of applied research: Barriers and recommendations. Account Res. 2023 Feb 17;30(2):88–108.

30. Karhulahti V, Branney P, Siutila M, Syed M. A Primer for Choosing, Designing and Evaluating Registered Reports for Qualitative Methods [Internet]. MetaArXiv; 2022 [cited 2023 Aug 16]. Available from: https://osf.io/2azkf

31. Betz CL. Adoption of CONSORT Statements for Randomized Control Trials Published in the Journal of Pediatric Nursing. J Pediatr Nurs. 2011 Jun;26(3):177–8.

32. Percie Du Sert N, Ahluwalia A, Alam S, Avey MT, Baker M, Browne WJ, et al. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. Boutron I, editor. PLOS Biol. 2020 Jul 14;18(7):e3000411.

33. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021 Mar 29;n71.

34. Cobo E, Cortes J, Ribera JM, Cardellach F, Selva-O'Callaghan A, Kostov B, et al. Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: masked randomised trial. BMJ. 2011 Nov 22;343(nov22 2):d6783–d6783.

35. Hopewell S, Boutron I, Altman DG, Barbour G, Moher D, Montori V, et al. Impact of a web-based tool (WebCONSORT) to improve the reporting of randomised trials: results of a randomised controlled trial. BMC Med. 2016 Dec;14(1):199.

36. Van Der Braak K, Ghannad M, Orelio C, Heus P, Damen JAA, Spijker R, et al. The score after 10 years of registration of systematic review protocols. Syst Rev. 2022 Sep 5;11(1):191.

37. Open Data Barometer. Global Report Fourth Edition [Internet]. 2017 [cited 2025 Feb 6]. Available from: https://opendatabarometer.org/doc/4thEdition/ODB-4thEdition-GlobalReport.pdf

38. Hardwicke TE, Thibault RT, Kosie JE, Wallach JD, Kidwell MC, Ioannidis JPA. Estimating the Prevalence of Transparency and Reproducibility-Related Research Practices in Psychology (2014–2017). Perspect Psychol Sci. 2022 Jan 1;17(1):239–51.

39. Bergeat D, Lombard N, Gasmi A, Le Floch B, Naudet F. Data Sharing and Reanalyses Among Randomized Clinical Trials Published in Surgical Journals Before and After Adoption of a Data Availability and Reproducibility Policy. JAMA Netw Open. 2022 Jun 2;5(6):e2215209.

40. Haven TL, Abunijela S, Hildebrand N. Biomedical supervisors' role modeling of open science practices. eLife. 2023 May 22;12:e83484.

41. Zuiderwijk A, Janssen M, Parnia A. The complementarity of open data infrastructures: an analysis of functionalities. In: Proceedings of the 14th Annual International

Conference on Digital Government Research [Internet]. New York, NY, USA: Association for Computing Machinery; 2013 [cited 2025 Feb 6]. p. 166–71. (dg.o '13). Available from: https://dl.acm.org/doi/10.1145/2479724.2479749

42. Ross-Hellauer T, Horbach SPJM. Additional experiments required: A scoping review of recent evidence on key aspects of Open Peer Review. Res Eval. 2024 Jul 22;33:rvae004.

43. Ross-Hellauer T, Bouter LM, Horbach SPJM. Open peer review urgently requires evidence: A call to action. PLOS Biol. 2023 Oct 4;21(10):e3002255.

44. Eich E. Business Not as Usual. Psychol Sci. 2014 Jan;25(1):3–6.

45. Rowhani-Farid A, Aldcroft A, Barnett AG. Did awarding badges increase data sharing in *BMJ Open* ? A randomized controlled trial. R Soc Open Sci. 2020 Mar;7(3):191818.

46. Kidwell MC, Lazarević LB, Baranski E, Hardwicke TE, Piechowski S, Falkenberg LS, et al. Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. Macleod MR, editor. PLOS Biol. 2016 May 12;14(5):e1002456.

47. Rowhani-Farid A, Barnett AG. Badges for sharing data and code at Biostatistics: an observational study. F1000Research. 2018 Mar 7;7:90.

48. Hardwicke TE, Bohn M, MacDonald K, Hembacher E, Nuijten MB, Peloquin BN, et al. Analytic reproducibility in articles receiving open data badges at the journal *Psychological Science* : an observational study. R Soc Open Sci. 2021 Jan;8(1):201494.

49. Crüwell S, Apthorp D, Baker BJ, Colling L, Elson M, Geiger SJ, et al. What's in a Badge? A Computational Reproducibility Investigation of the Open Data Badge Policy in One Issue of *Psychological Science*. Psychol Sci. 2023 Apr;34(4):512–22.

50. Pratt MG, Kaplan S, Whittington R. Editorial Essay: The Tumult over Transparency: Decoupling Transparency from Replication in Establishing Trustworthy Qualitative Research. Adm Sci Q. 2020 Mar;65(1):1–19.

51. Holbrook JB, Penders B, De Rijcke S. CWTS. 2019 [cited 2025 Feb 6]. The humanities do not need a replication drive. Available from: https://www.cwts.nl:443/blog?article=n-r2v2a4&title=the-humanities-do-not-need-a-replication-drive

52. De Rijcke S, Penders B. Resist calls for replicability in the humanities. Nature. 2018 Aug 2;560(7716):29–29.

53. Ross-Hellauer T, Klebel T, Bannach-Brown A, Horbach SPJM, Jabeen H, Manola N, et al. TIER2: enhancing Trust, Integrity and Efficiency in Research through next-level Reproducibility. Res Ideas Outcomes. 2022 Dec 8;8:e98457.

54. Tijdink JK, Leitner B, Horbach S, Cole NL, Kopeinik S, Ross-Hellauer T. Future Studies - The Future(s) of Reproducibility in Research. 2023 [cited 2025 Feb 11]; Available from: https://osf.io/wgtku/

55. Bell W. Foundations of Futures Studies, Volume 1: Human Science for a New Era. Transaction Publishers; 2009. 406 p.

56. Boulding E, Boulding KE. The Future: Images and Processes. SAGE Publications; 1995. 248 p.

57. Inayatullah S. Futures Studies. Theories and Methods. Theres Future Vis Better World. 2013 Jan 1;36–66.

58. Deterding NM, Waters MC. Flexible Coding of In-depth Interviews: A Twenty-first-century Approach. Sociol Methods Res. 2021 May 1;50(2):708–39.

59. Drude N, Martinez-Gamboa L, Haven T, Holman C, Holst M, Kniffert S, et al. Finding the best fit for improving reproducibility: reflections from the QUEST Center for Responsible Research. BMC Res Notes. 2022 Aug 3;15(1):270.

60. Lalu MM, Presseau J, Foster MK, Hunniford VT, Cobey KD, Brehaut JC, et al. Identifying barriers and enablers to rigorous conduct and reporting of preclinical laboratory studies. PLOS Biol. 2023 Jan 5;21(1):e3001932.

61. Leonelli S. Open Science and Epistemic Diversity: Friends or Foes? Philos Sci. 2022 Dec;89(5):991–1001.

62. Cole NL, Ulpts S, Bochynska A, Kormann E, Good M, Leitner B, et al. Reproducibility and replicability of qualitative research: an integrative review of concepts, barriers and enablers [Internet]. OSF; 2024 [cited 2025 Feb 11]. Available from: https://osf.io/n5zkw_v1

63. Uttley L, Falzon L, Byrne JA, Tricco AC, Munafò MR, Moher D, et al. Research culture influences in health and biomedical research: rapid scoping review and content analysis. J Clin Epidemiol. 2025 Feb 1;178:111616.

64. Gopalakrishna G, Ter Riet G, Vink G, Stoop I, Wicherts JM, Bouter LM. Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. PloS One. 2022;17(2):e0263023.

65. Stengers I. Another science is possible: a manifesto for slow science. English edition. Cambridge, UK: Polity; 2018. 1 p.

66. Leite L, Diele-Viegas LM. Juggling slow and fast science. Nat Hum Behav. 2021 Apr;5(4):409–409.

67. Frith U. Fast Lane to Slow Science. Trends Cogn Sci. 2020 Jan 1;24(1):1–2.

68. Schimanski LA, Alperin JP. The evaluation of scholarship in academic promotion and tenure processes: Past, present, and future [Internet]. F1000Research; 2018 [cited 2025 Feb 11]. Available from: https://f1000research.com/articles/7-1605

69. Larivière V, Pontille D, Sugimoto CR. Investigating the division of scientific labor using the Contributor Roles Taxonomy (CRediT). Quant Sci Stud. 2021 Apr 8;2(1):111–28.

70. Ross-Hellauer T, Reichmann S, Cole NL, Fessl A, Klebel T, Pontika N. Dynamics of cumulative advantage and threats to equity in open science: a scoping review. R Soc Open Sci. 2022 Jan 19;9(1):211032.

71. Michie S, van Stralen MM, West R. The behaviour change wheel: A new method for characterising and designing behaviour change interventions. Implement Sci. 2011 Apr 23;6(1):42.