Questionable Research Practices and the Reproducibility Crisis in Animal-Based Biomedicine: A New Perspective

Simon Lohse (simon.lohse@ru.nl)

Abstract: I discuss reproducibility issues in animal-based research in biomedicine and scrutinize the notion that the causes of non-reproducible results are the same as in other disciplines. I argue that there are aspects characteristic of animal experimentation that are important for analysing reproducibility problems but have not yet been discussed in this context. Using an approach that integrates epistemological and ethical questions, I explore these aspects and show that the prevalent focus on questionable research practices and methodological reforms falls short in understanding and managing key challenges to reproducibility in animal-based biomedicine.

Keywords: Replication crisis; animal experimentation; standardisation; 3R principle; philosophy of science in practice.

1. Introduction

In 2004, Pound et al. published the article "Where Is the Evidence That Animal Research Benefits Humans?" where they assessed existing systematic reviews of animal-to-human translation and painted a dire picture of the usefulness of animal-based research for translational purposes in biomedicine. Many of the reviewed animal experiments had methodological flaws and were of questionable reliability and predictive value. A decade later, Mak et al. (2014) came to the conclusion that more than 92% of animal-based cancer research gets "lost in translation", with 85% of new drugs that were successful in pre-clinical testing (including animal testing) failing in early clinical trials (also see Pound & Bracken, 2014). Translational issues of this kind have, together with more general concerns regarding progress in drug development, motivated much recent work on methodological issues in pre-clinical research and led to the finding that much if not most animal-based research is in fact irreproducible (Freedman et al., 2015). This fuelled talk about a replication (or reproducibility) crisis in biomedical research, which is widely seen as an important factor in failure to transfer results from pre-clinical research to the clinic (Baker, 2016; Engber, 2016a, 2016b; Fidler & Wilcox, 2018).[1]

The observed low degree of reproducibility in biomedical research has also been discussed by philosophers of science interested in the reproducibility crisis in science.

[1] I am aware that sometimes "replication" refers to repeating the experimental *procedure* as closely as possible, whereas "reproduction" addresses the reliability of experimental *results*. However, I use both terms synonymously in this paper.

Most discussions focus on issues in the psychological sciences and/or on general aspects of scientific reproducibility/replication. Several philosophers of science have provided conceptual analyses and taxonomies of different types of replication (e.g. Machery, 2020; Shavit et al., 2017). Others have debated the meaning of reproducibility or questioned the alleged key role of direct replication for scientific research (e.g. Leonelli, 2018; Feest, 2019). A third body of literature addresses underlying causes of the reproducibility crisis in science, whereby a majority position seems to have emerged stating that the causes of replication problems are to a large extent the same across fields (see below).

In this paper, I focus on reproducibility issues in animal-based research in biomedicine and scrutinise the notion that the causes of irreproducible results are indeed the same across disciplines. I argue that there are aspects characteristic of animal experimentation that are important for analysing reproducibility problems but have not yet been sufficiently addressed. Using an approach that integrates epistemological and ethical issues, I explore these aspects and resulting challenges to reproducibility by means of two mini case studies. This results in critical questions about the utility of focussing mainly on "questionable research practices" and methodological reforms to understand and overcome reproducibility problems in animal-based biomedicine.

I begin by sketching the replication/reproducibility discussion in animal-based biomedical research and review the most commonly mentioned underlying causes and

suggested fixes for irreproducibility in this context. Next, I introduce two challenges to reproducibility related to animal-based research that have not received much attention in philosophical debates. This also serves to cast a new light on deviations from "good research practice" in this context. *First*, I describe methodological challenges to standardisation practices in animal-based research. While the received view assumes the need for high standardisation in biomedical research to increase the validity and robustness of results, several animal researchers have argued that too much standardisation might be part of the problem for animal-based research. I discuss this tension and its implications using examples from animal experimental practice. *Second*, I discuss what I call "ethico-epistemic trade-offs" in research practice. These trade-offs are a consequence of the controversial nature of animal experimentation and manifest in situations where epistemic and non-epistemic values conflict. In the concluding section, I draw out normative consequences of my analysis for the discussion revolving around questionable research practices in animal-based biomedicine and argue for a more pluralistic and nuanced discussion of replication issues in the meta-scientific discourse.

2. The reproducibility crisis in animal-based biomedicine

As indicated in the introduction, it was translational problems to the clinic (combined with more general concerns regarding progress in drug development) that motivated work on methodological issues within the biomedical research community. In

particular two articles (Begley & Ellis, 2012; Prinz et al., 2011) have sparked debate and co-initiated talk about the replication crisis in biomedicine. These articles report the results of carefully conducted validation studies of cutting-edge biomedical research that could for the most part not be replicated. Many more studies have been conducted since, which corroborate the fact that much if not most biomedical science is in fact not replicable (for an overview see Begley & Ioannidis, 2015; Freedman et al., 2015). These studies were for the most part "direct replications" (Pashler & Harris, 2012), i.e. attempts to achieve (more or less) the same results by replicating the same general experimental design as closely as possible. Sometimes this involved contacting researchers of the original studies and even working in the same laboratories. However, there were also elements of "conceptual replication" where features of the experimental setup were systematically varied (e.g. using in vitro/vivo models) to test the robustness of results (see Feest, 2019 for conceptual challenges related to this distinction).

As many preclinical studies make extensive use of animal experiments, some authors have pointed out that precisely this could be a significant part of the problem (Macleod & Mohan, 2019; Spanagel, 2022). A provocative piece in *PLOS Biology* even suggests that the replication of certain pre-clinical results (involving animals) would be so low that it could theoretically be replaced by a coin flip (Piper et al., 2019). Hence, the assumption of a connection between irreproducible animal-based research and a lack of transferability to the clinic seems to have some plausibility. It should

not be overlooked, however, that other factors also play a major role in persisting translational problems (Leenaars et al., 2019), namely effectiveness issues in later phases of clinical trials (Magee, 2013), economic and organisational hurdles (Seyhan, 2019), sub-quality research materials (Guttinger & Love, 2019) and, of course, long-known foundational problems with transferability from animal models to humans (LaFollette, 2011; LaFollette & Shanks, 1993; Green, 2024).[2] I will not consider these factors further here, though, but focus on replication problems as they have been discussed in the context of biomedical research, i.e. problems in reproducing results in important respects using a comparable animal-based experimental design. Even if these are not the only problems of animal-based biomedicine, they can indeed be an indicator of systematic methodological problems in research and thus deserve attention.

2.1 Is the crisis real, though?

But let's not rush to conclusions. Are replication problems really such a big problem in biomedical research? Do these problem really amount to a "crisis"? Recent work in the philosophy of science invites scepticism. There are two arguments in particular that pose challenges to the diagnosis of a replication crisis. The first challenge results from the observation that there are many more successful replications in research

[2] Which is why it is misleading to claim that ≈ 90% of results from preclinical animal research do not translate to the clinic *because* animal-to-human translation does not work. Rather, it is a complex question of how exactly the 90% actually come together.

practice than one might think. Guttinger (2018, 2019) calls these “micro-replications”. These are partial replications of certain methods and results in the life sciences by other research groups on which further investigations are built. Micro-replications are not intended as full-blown replication studies but occur as an integral element of “everyday research”, for instance in form of setting-up and calibrating experimental systems. The implication of this is that there are considerably more successful replications in the life sciences (including biomedicine) than is generally assumed.

Although this analysis is convincing, in my view it still leaves room for the diagnosis of serious replication issues, even if these might be somewhat less serious than initially conceived. Recall that the diagnosis of a replication crisis was based on explicit attempts to reproduce key results in cutting-edge biomedical research. These attempts failed, leading to empirically supported irritations in the field: Even if replications of partial experimental designs often work, there may still be a real problem with top-drawer biomedical research.

A second challenge stems from Bird's claim (2021), based on considerations by Sterne/Davey Smith (2001), Ioannidis (2005) and others, that the replication crisis may in fact be a base rate fallacy. The idea is that due to the fact that biomedical hypotheses are rarely derived from well-established theories, these have a low prior probability of being true leading to “high proportion of positive test outcomes that are in fact false positives” (Bird, 2021, p. 971). Hence, failure to replicate should not

come as surprise indicating poor scientific methodology but as expected outcome in this field of study. No replication crisis![3] In response to this, however, Autzen (2021) points out that it is quite unclear whether Bird's assumption of low initial probability of biomedical hypotheses actually holds. He shows that biomedical hypotheses, although rarely derived from a well-corroborated theory, are often based on contextual and empirical evidence and can therefore have a higher initial probability than Bird assumes. It should be added that this consideration seems to be in particular relevant to research involving animal experiments where scientists need to make a strong case for the scientific rationale of a study in order to get the respective experiment approved by the authorities (at least in many OECD countries). Based on these considerations, it is unclear how conclusive Bird's analysis actually is. In any case, it is plausible to assume that not all observed replication issues can be explained (away) as a base rate fallacy and that the phenomenon should therefore be taken seriously – at least until there is more definitive evidence to the contrary. This is all the more true as any reproducibility problems in the context of animal experimental research directed at human health/disease touch on important normative concerns. In my view, these are the four most important ones:

[3] The high number of false positives could still constitute a problem, though.

(1) Science: The reliability and credibility of research results in biomedicine is called into question. This also applies to all studies that are based on the results of irreproducible results.

(2) Economic issues: Animal experiments such as research on new drug targets are very expensive. So there could be a considerable waste of resources and related opportunity costs regarding the pursuit of alternative methods for drug R&D (Akhtar, 2015).

(3) Animal ethics: Animal harm and suffering could be unnecessarily high.

(4) Human health: To the extent that lack of reproducibility indicates methodological problems and unreliable findings, this could lead to pointless, even harmful human trials and ultimately to (more) translational failure.

2.2 Causes and fixes

I thus assume that replication failures in animal-based biomedicine should be taken seriously. This corresponds, of course, to the received view that has emerged in meta-science, the biomedical research community and at least some areas of philosophy of science, a view that also agrees broadly on the idea that the underlying causes of the problem are quite similar to other fields where a replication crisis has been identified, for instance in social psychology.[4] Here is a neat summary from the *Stanford Encyclopedia of Philosophy* article on the matter:

[4] See Feest (2024) for an alternative diagnosis of replication issues in psychology.

> "*The causes of irreproducible results are largely the same across disciplines* we have mentioned. This is not surprising given that they stem from problems with statistical methods, publishing practices and the incentive structures created in a 'publish or perish' research culture, all of which are largely shared, at least in the life and behavioral sciences" (Fidler & Wilcox, 2018, my emphasis).

This assessment resonates with meta-science work on animal-based research in biomedicine, which diagnoses research problems that are mostly generic. As in other disciplines, "questionable research practices" are considered to be key factors responsible for replication problems. Questionable research practices are in the "grey area" between clear scientific fraud and good scientific practice, seen as detrimental to rigorous science and likely to lead to biased results (Banks et al., 2016). Although there is no agreement on what exactly falls under this concept (Ravn & Sørensen, 2021), there are a several practices that are regularly categorised as questionable research practice, including selective reporting of positive or "ground-breaking" results, cherry-picking of data, p-hacking and similar statistical manipulations, and violations of established methodological standards, such as using underpowered studies and a lack of blinding, randomisation or standardisation, especially in biomedical studies (Ioannidis et al., 2014; Sánchez Morgado & Brønstad, 2021).

It should come as no surprise, then, that suggested ways of dealing with the replication crisis target these very causes and are thus similar to the ones suggested in other fields (see, e.g., Begley & Ioannidis, 2015; Ioannidis et al., 2014): Researchers should improve research design (e.g. larger sample sizes), provide better reporting (e.g. describe experimental procedures in more detail), implement pre-registration protocols of experimental studies, improve standardisation and provide better training for statistics and good research practice. In addition to these remedies, more fundamental reforms (that are difficult to establish effectively) are also recommended. These aim at revising review processes, transforming the incentive structure and cultural aspects of science, for example by recommending to acknowledge the benefits of replication studies more strongly and abolishing the logic of "publish or perish".

The identified causes and fixes are persuasive – but they are also quite generic. Context-specific factors and specific challenges of animal-based science understood as *material research practice in concrete contexts*, on the other hand, seem to hardly play a role in the replication crisis. From the point of view of modern science studies this assumption should raise some eyebrows. Although generic factors are certainly relevant for replication problems in biomedicine, it is implausible to assume that they are the only relevant factors. This is implausible in view of the extensive research in philosophy of science, HPS and STS that demonstrates the influence of epistemic cultures, material aspects, specific historical conditions, political context etc. on scientific practice in general (representative for many more: Ankeny & Leonelli, 2016;

Hackett, 2008; Knorr Cetina, 1999; Latour & Woolgar, 1979) and animal research in particular (e.g. Davies, 2021; Green, 2024; Lohse, 2021; Lowe et al., 2019). One of the main aims of this paper is thus to shift the focus, so that some of these aspects of animal-based research can come into view. This will not only lead to questioning the received view regarding the replication crisis in animal-based biomedicine, but also shed new light on the usefulness of the concept of questionable research practices.

3 Zooming in: animal experimentation and reproducibility

To achieve these aims, I discuss two instructive challenges for animal-based research in biomedicine that bear on the replication issue. First, I explore standardisation issues and scrutinise the idea that more standardisation is the way forward. Next, I want to introduce the issue of ethico-epistemic trade-offs in animal-based biomedicine using examples from research practice that pose a challenge for prevalent suggestions for good research practice. In the latter case, I draw not only on literature analysis but also on my experience as an embedded researcher in two life science consortia.[5]

[5] The Transregional Collaborative Research Centre "Biology of Xenogeneic Cell and Organ Transplantation - from Bench to Bedside" (https://www.klinikum.uni-muenchen.de/SFB-TRR-127/de) and the research network "R2N – 'Replace' and 'Reduce' in Lower Saxony (https://r2n.eu).

3.1 Standardisation

Standardisation is a hot topic in animal experimentation and concerns many aspects of research practice. In an overview article, Sánchez-Morgado et al. (2021) describe the relevance of a variety of factors that can influence research results in rodent studies and whose inadequate control and standardisation can be a problem. Among other things, they highlight environmental enrichment, light regimes, animal handling, diet, noise level, temperature, humidity, and even cage position in the room, and how differences in these respects can influence the reproducibility of results. For instance, day/night rhythm affects sleep and behaviour patterns of mice, as does noise level. This can have drastic effects on the reliability of measured outcomes, especially in cases where effect levels of (say) a medical treatment are rather low.

I cannot deal with all these factors in this paper, but will concentrate on one aspect that has been much discussed recently, namely the question of microbiota standardisation in rodents. The microbiota refers to the entirety of microorganisms living in (e.g. gut) and on (e.g. skin) an animal. As is well known by now, even rodent animal models with identical genomes may exhibit differences in their microbiota composition that can affect their behaviour, inflammation levels and immune responses to an experimental procedure in biomedical research (Hansen, 2021). As might be expected, microbiota composition will be different in animals from different vendors, but there are also differences between different batches from the same vendor. In addition, diet, cage characteristics and other hard-to control factors can influence

microbiota composition. Finally, there is horizontal transfer of microorganisms between animals with different microbiota inducing variation in the laboratory. At the same time, there is only limited screening of microbiota variation and potential effects of such variation (Witjes et al., 2020). Rather, for economic and pragmatic reasons, screening will most of the time only be considered if microbiota composition is *expected* to be relevant, i.e. to have an effect on the outcome of interest. The issue with this is, of course, that scientists may not always be in a position to know when to expect a relevant effect. As a consequence, unknown variety in microbiota composition (in particular instable variety) may affect reproducibility in animal experimentation in more cases than researchers are aware of.

An obvious remedy for this problem is to increase standardisation. The idea is to create and maintain stable microbiota composition by stricter vendor and laboratory protocols, higher standardisation of housing, diet etc. (Bleich & Hansen, 2012; Macpherson & McCoy, 2015). This would limit the amount of (unknown) variation of microbiota in rodent models and might, therefore, improve reproducibility of experimental results. However, the standardisation approach seems to face serious challenges in the context of animal-based research. The first problem is that standardisation often results in very sterile environments. This leads to compromised immune reaction in animal models which in turn leads to (even) less predictability of human immune responses in the biomedical context (Rosshart et al., 2019). The second problem is what has been dubbed the "standardisation fallacy". In a series of

paper, Würbel, Richter and co-authors have argued (and empirically corroborated) that increasing lab-internal reproducibility through standardisation comes at the expense of robustness across laboratories (Richter et al., 2009, 2011; Voelkl et al., 2018, 2021; Würbel, 2000). Due to the phenotypic plasticity of animals, depending on environmental factors, rigorous standardisation leads to animal cohorts being more homogeneous *within* a laboratory than *between* laboratories. This is because several factors will be difficult to standardise between laboratories. Practical restrictions may play a role, but also limitations in describing all the *potentially* relevant details of a procedure or housing condition[6], or simply the fact that effects of an environmental condition are not even considered to be relevant – for instance, researchers have not considered the influence of the experimenter's sex (Georgiou et al., 2022) or them having cats as pets (Panksepp, 1998, p. XX) on stress levels of rodents for a long time.

As a result, test animals may become homogenised to specific laboratory conditions – for example they may react more sensitively to certain procedures than elsewhere – which means that test results are less reproducible *between* laboratories with (even slightly) different environmental conditions. Moreover, a similar effect could occur when conditions change within the same laboratory. This point can be linked to potential issues with microbiota variability:

[6] An additional challenge is that not all implicit knowledge about the details of experimental designs can be made explicit in a straightforward way.

> "Subtle changes in environment have more impact on microbiome composition in a standardized laboratory environment in comparison to a natural environment. Therefore*, standardization might have complicated reproducibility* by creating mouse models with unstable (less resilient) microbiomes" (Witjes et al. 2020, 10f, my emphasis).[7]

In light of these problems, some authors argue that we should indeed go for *less* standardisation to strengthen reproducibility. One way of doing this is to implement controlled heterogenisation of microbiota in animal models (and prospectively: other factors relevant for replication issues) instead of ever stricter standardisation regimes. The clue is to avoid over-standardisation at the expense of robustness across contexts by systematic environmental variation, making animals within a given experiment more heterogeneous (for details and challenges of this strategy, see Richter et al., 2009 and Voelkl et al., 2018). A related approach is to increase real-world heterogeneity of microbiota through the "wildling strategy". Here the idea is to make laboratory animals with highly standardised properties (for instance in infection research) more natural and their microbiota more diverse and resilient by surrogate mothering of bred mice by wild mice (Graham, 2021). Although this approach has several drawbacks, including increasing data noise and the risk of cross-contamination between different

[7] Note that this quote concerns the *microbiome* so the genome of the microbiota.

types of mouse models in the same laboratory (Hansen, 2021), the wildling strategy could lead to progress in tackling the over-standardisation issue described.

So is the "de-standardisation" strategy the right one for making animal-based biomedical research more reliable? There does not yet seem to be a conclusive answer to this question. Rather, we are dealing with an open methodological discussion, where different authors argue for different positions (see Vatsos, 2017; Witjes et al., 2020). In addition, it is likely that different recommendations regarding microbiota standardisation will be appropriate depending on research context and concrete epistemic aims of a study (e.g. studying immune response in cancer research vs. testing a new surgical procedure). However, if this is the case, it has implications for the replication discourse. In particular, universalistic recommendations for standardisation by (some) authors in meta-science research appear inappropriate then. This is simply because such recommendations are not sensitive enough to research context. This observation also raises questions concerning the classification of certain methodological procedures (standardised/non-standardised) as "questionable research practice". Such a generalised classification does not appear to be possible for the same reason: it does not take into account the respective research purpose, the details of experimental practices in different sub-fields etc. (I come back to this point in the concluding section.)

3.2 Ethico-epistemic trade-offs

Next, I discuss balancing issues between epistemic and ethical aspects of animal-based research and how these may relate to methodological questions. These issues are a consequence of the ethically controversial nature of animal experimentation and its strict regulation and they occur when epistemic and non-epistemic values are – or *appear* to be – in conflict.

To approach this issue, it is useful to consider the ARRIVE (Animal Research: Reporting of In Vivo Experiments) recommendations for research design and reporting (fig. 1). These recommendations push for better reporting on certain details of animal experiments, including information about control groups, randomisation, sample size, and blinding (Kilkenny et al., 2010; revised guidelines: Percie du Sert et al., 2020). These recommendations are not only intended to improve reporting, but also serve as an incentive to rethink and possibly improve research design, for example by introducing more control groups or blinding – to improve, amongst other things: reproducibility.

The ARRIVE Essential 10

These items are the basic minimum to include in a manuscript. Without this information, readers and reviewers cannot assess the reliability of the findings.

Study design	1	For each experiment, provide brief details of study design including: a. The groups being compared, including control groups. If no control group has been used, the rationale should be stated. b. The experimental unit (e.g. a single animal, litter, or cage of animals).
Sample size	2	a. Specify the exact number of experimental units allocated to each group, and the total number in each experiment. Also indicate the total number of animals used. b. Explain how the sample size was decided. Provide details of any *a priori* sample size calculation, if done.
Inclusion and exclusion criteria	3	a. Describe any criteria used for including and excluding animals (or experimental units) during the experiment, and data points during the analysis. Specify if these criteria were established *a priori*. If no criteria were set, state this explicitly. b. For each experimental group, report any animals, experimental units or data points not included in the analysis and explain why. If there were no exclusions, state so. c. For each analysis, report the exact value of *n* in each experimental group.
Randomisation	4	a. State whether randomisation was used to allocate experimental units to control and treatment groups. If done, provide the method used to generate the randomisation sequence. b. Describe the strategy used to minimise potential confounders such as the order of treatments and measurements, or animal/cage location. If confounders were not controlled, state this explicitly.
Blinding	5	Describe who was aware of the group allocation at the different stages of the experiment (during the allocation, the conduct of the experiment, the outcome assessment, and the data analysis).
Outcome measures	6	a. Clearly define all outcome measures assessed (e.g. cell death, molecular markers, or behavioural changes). b. For hypothesis-testing studies, specify the primary outcome measure, i.e. the outcome measure that was used to determine the sample size.
Statistical methods	7	a. Provide details of the statistical methods used for each analysis, including software used. b. Describe any methods used to assess whether the data met the assumptions of the statistical approach, and what was done if the assumptions were not met.
Experimental animals	8	a. Provide species-appropriate details of the animals used, including species, strain and substrain, sex, age or developmental stage, and, if relevant, weight. b. Provide further relevant information on the provenance of animals, health/immune status, genetic modification status, genotype, and any previous procedures.
Experimental procedures	9	For each experimental group, including controls, describe the procedures in enough detail to allow others to replicate them, including: a. What was done, how it was done and what was used. b. When and how often. c. Where (including detail of any acclimatisation periods). d. Why (provide rationale for procedures).
Results	10	For each experiment conducted, including independent replications, report: a. Summary/descriptive statistics for each experimental group, with a measure of variability where applicable (e.g. mean and SD, or median and range). b. If applicable, the effect size with a confidence interval.

Figure 1: ARRIVE essential recommendations (source: https://arriveguidelines.org/sites/arrive/files/documents/ARRIVE%20guidelines%202.0%20-%20English.pdf)

These and similar recommendations are being urged and discussed more than ever with the emergence of the replication crisis discourse. Yet, there does not seem to be extensive change regarding many of these practices. Although the ARRIVE guidelines are supported by several journals in the field, measures to reduce biases in particular are still too little reported, which presumably indicates that such measures are still not as widely used as one might hope (Frommlet & Heinze, 2021; Leung et al., 2018). While institutional factors and a degree of scientific conservatism are likely to play a

role here, I want to suggest that there are also certain ethico-epistemic trade-offs in animal-based research that contribute to lack of progress. I illustrate this claim with two examples.

The first example concern lack of blinding in animal-based research. Blinding is widely recommended (not least in response to the replication crisis) in experimental setups with treatment and control groups to avoid observer and performance bias (see, e.g., ARRIVE guidelines). So why is it not an established standard in most animal-based research in biomedicine? Is this just bad research practice? Starting points for an alternative view can be found in a *Nature* commentary by Nelson (2021), in which she reports on her exchange with researchers on questions about methodological conservatism in research. One reason given there for not blinding experiments is that problems with re-identification could occur leading to mix-ups with severe consequences for the validity of the study in question. But misidentification is not just a practical problem for research design. There also are ethical implications of possible misidentification – as of non-identifiability of treatment/control group through blinding in general, in particular when animal pain or distress is involved. There are concerns that blinding may lead to suboptimal severity monitoring and inadequate pain relief measures. Severity monitoring in laboratory animals consists in the observation and measurement of several behavioural and physiological indicators to detect pain, distress, suffering etc. resulting from an experimental procedure and assess their severity. It is used for scientific quality control, is ethically required for animal

welfare reasons and legally obligatory in many countries (see, e.g., Annex VIII of the EU DIRECTIVE 2010/63). However, to achieve these goals, it is frequently important to know whether animals showing unusual behaviour are in the treatment or control group of an experiment. Here is a concrete example to illustrate. In cancer research, certain behavioural changes such as hunched posture or social withdrawal will be tolerated for a certain time as a transient side effect of an oncological treatment, while in the control group it could indicate a serious problem to act on (Karp et al., 2022). In such a scenario, blinding poses a challenge for researchers who are concerned about animal welfare.

In addition to this challenge, rodents – the most widely used animal in biomedical experimentation – in particular are believed to be very good at hiding pain through masking behaviour.[8] As subtle behavioural indicators are often an important element of severity monitoring (Leenaars et al., 2019), blinding may thus (further) complicate the appropriate assessment of the painfulness of an experimental procedure. It will not only require more time and/or staff – which may not be available – for monitoring more animals (recall that we do not know which animals are in the control group). It may also require more frequent and longer observations interfering with the experimental setup to make up for the risk of missing behavioural cues when not

[8] I write "are believed to be" because the matter may be more complicated and depend, among other things, on the interaction between humans and animals. However, it is sufficient to assume that "pain masking behaviour in rodents" is a widely shared belief in animal researcher communities (see Carbone, 2020 for an interesting discussion).

knowing which animals have undergone a procedure. This may be especially relevant when new procedures or compounds are tested where adverse effects on behaviour are completely unpredictable. In such cases, it is important to include as many indicators and contextual factors as possible in pain monitoring (Hawkins et al., 2011) which may be made more difficult by blinding.

Note that (a) observing behaviour cannot always be substituted by measuring other pain indicators (e.g. corticosterone levels) as this may affect the scientific outcome (*ibid;* Carbone & Austin, 2016), and (b) providing pre-emptive analgesia for all animals in an experiment may not always be desirable as pain medication can affect experimental outcomes and increase data variability, e.g. via physiological and behaviour effects (Jirkof & Potschka, 2021). This is not to say that there cannot be experimental setups whose sophisticated design can mitigate or even solve the aforementioned severity assessment issues (e.g. by separating science and welfare management responsibilities). Rather, the point is that in cases such as the one just outlined, there can be real balancing problems between epistemic and ethical aspects from the actors' point of view; problems which can explain certain reservations regarding blinding procedures.

My second example for an ethico-epistemic trade-off concerns the demand to reduce animal numbers in experimental setups according to the 3R principle (in line with current regulation in many OECD countries). According to this principle, researchers

should attempt to **r**eplace, **r**educe and **r**efine animal experiments wherever possible (Russell & Burch, 1959). *Reducing* in this context means that we should use as few animals as possible for a given experimental purpose. So if it is possible to answer a research question by using 30 animals instead of 40, scientists are obliged to do so in order to minimise animal use and suffering. It is important, however, that the research design is not negatively affected by this. This means that researchers should *not*, for ethical or regulatory reasons, reduce the number of animals, control groups or animal-based robustness studies to such an extent that the scientific conclusiveness of a study is jeopardised.

Unfortunately, this occasionally happens and may lead to studies with uncertain reliability or validity. The reason for this is that scientists may, for example, try to reduce the sample size of a study in order to fulfil the 3R-goal of reducing animal experiments as much as possible and go too far, which may lead to unsound results and ultimately to a waste of animals (Eggel & Würbel, 2021). It is important to note, however, that such cases are not *just* regular cases of bad practice resulting from carelessness or sloppiness. And they are certainly not questionable research practices as sometimes characterised in the literature, i.e.

> "[…] a consequence of a system that is willing to overlook and ignore lack of scientific rigor and instead reward flashy results that generate scientific buzz or excitement" (Begley & Ioannidis, 2015, p. 118).

Rather, the pro-active but in this case sub-optimal balancing of ethical and epistemic aspects is responsible for the problem – which may be explained, at least in part, by the strong regulatory focus on the 3Rs and perhaps an unintended side-effect of a (desirable!) "culture of care" that emphasises responsible conduct in the context of animal research (Davies et al., 2018).

There is another, deeper problem with the methodological caveat regarding (over-)reduction: It may not always be clear-cut how much evidence is enough for a given purpose. There are cases where the decision on sample size or in-house replications etc. will need to be made by carefully balancing the inductive risks of false positives/negatives (Douglas, 2000; Elliott & Richards, 2017). Consider a scenario in which scientists want to investigate the possible side-effects of a new drug. The therapeutic in question has similar properties to already authorised drugs and is considered a promising candidate for the treatment of certain diseases in humans, but has never been tested on a whole organism. How much research on animals will it take before we have *enough* evidence to enter phase I trials in humans? How large does the sample need to be? How many in-house replication studies should scientists conduct to be certain enough that there will be no dangerous side-effects? Should other species be tested in addition to rodents? The answers to these questions naturally depends on established statistical and regulatory standards and a variety of details about the drug in question. But they also depend on risks projections, the experience of scientists and

regulatory authorities investigating similar drugs, and, last but not least, earlier setbacks in drug safety assessment etc. (cf. reforms of preclinical/clinical research practice following the infamous TGN1412 incident, see Lemoine, 2017). In short, these are questions whose answers *also* depend on balancing potentially unnecessary harm to animals and possible harm to humans. However, if this is true, it touches on questions of good scientific practice. These may not be answerable in a purely methodological way, for example by giving general advice on statistically optimal experimental design in light of replication issues, but must, at least occasionally, take ethical aspects into account too. Contrary to the received view in meta-science and biomedicine, the *right* experimental design may depend on striking the right balance between several epistemic *and ethical* aspects, where in this case, ethical aspects become relevant on both sides of the scale (harm to humans vs. animals).

4. Concluding thoughts

In a paper on reproducibility as a quality criterion for science, Leonelli (2018) points out that failure to reproduce does not necessarily imply bad science (although it certainly can be an indicator), but can also lead to productive scientific investigations, such as indicating limited generalisability. As we have seen, it can also lead to productive *philosophical* investigations with normative impetus. This is especially the case when there is an intertwinedness of epistemic, ethical and institutional (e.g. 3R guidelines) factors leading to deviance from what is generally believed to be good

research practice, as illustrated in the second mini case study in this paper. The identification of ethico-epistemic trade-offs opens up new research horizons in this context - both in cases of unavoidable value conflicts and in those that could ultimately be resolved[9], but until then may influence methodological decisions and actions of scientists in their everyday research practice. How could these be made more transparent? On what basis should the balance be struck? Who should do this? Such further-reaching questions could be fruitfully addressed by an approach that integrates philosophy of science in practice and research on ethical, legal and social issues in the life sciences (as suggested in Lohse et al., 2020).

As far as questionable research practices are concerned, my discussion supports the observation by Stefan Guttinger that this concept has limited utility, at least if it is to be understood as a universal recipe for avoiding bad science.[10] Although certain research practice are indeed almost always problematic (e.g. publication biases), in many cases, it will be highly context-dependent which methodological decisions are problematic – as their validity depends on research aim, local practices and challenges etc., and this may even change over time. As I have attempted to show, the assessment of a specific decision or procedure as problematic may, at times, also depend on ethical considerations. This conclusion casts doubt on the usefulness of developing guidelines

[9] See, e.g., Neumann et al.'s (2017) proposal to reduce the number of required animals according to the 3R-principle by a sophisticated statistical study design.

[10] This paragraph draws on a very insightful talk by Guttinger, "What are questionable research practices?", presented in the workshop series *The Statistics Wars and Their Casualties*. See https://phil-stat-wars.com & https://phil-stat-wars.com/wp-content/uploads/2022/12/guttinger-final.pdf [accessed 20 August 2024].

for good scientific practice with strongly universalistic tendencies, including those aimed at large scientific fields, such as medicine and psychology. Rather than relying pre-dominantly on such guidelines, we should pay more attention to local methodological norms engrained in specific epistemic practices and cultures; norms about which we do not know nearly enough, indicating the need for more qualitative studies informed by a philosophy of science/ STS perspective.

Such studies have the potential to re-orientate certain areas of the debate about replication, including the meta-science debate on animal-based biomedicine. They could enrich a new localism in reproducibility studies (Guttinger, 2020; Leonelli 2018) and help to paint a more pluralistic and nuanced picture of the state(s) of affairs. This paper is an attempt to contribute to this, not least because I believe that such a picture provides a more appropriate starting point for methodological critique of animal-based research. This is a form of critique that is empirically well-informed and context-sensitive, for instance regarding the *details* of how the 3R framework may affect replication and other methodological issues in animal-based research (Lowe et al., 2019). Such an approach will also have practical implications for policy proposals that aim at making science better, as it raises new, fundamental questions regarding ways to manage the described reproducibility challenges. For instance, given the discussed standardisation challenges and ethico-epistemic trade-offs in animal experimentation, how should we assess animal-based biomedicine and the possibility of reform? A proponent of this research practice might argue that we need different,

more appropriate scientific quality standards than in other disciplines with replication problems, a critic may interpret the described challenges as (additional) evidence for the deep flaws of animal-based research in biomedicine that should ultimately be discontinued. Questions like these are highly significant but become visible only against the backdrop of a context-sensitive analysis of reproducibility problems in science.

Acknowledgements

[omitted]

References

Akhtar, A. (2015). The Flaws and Human Harms of Animal Experimentation. *Cambridge Quarterly of Healthcare Ethics*, *24*(4), 407–419. https://doi.org/10.1017/S0963180115000079

Ankeny, R. A., & Leonelli, S. (2016). Repertoires: A Post-Kuhnian Perspective on Scientific Change and Collaborative Research. *Studies in History and Philosophy of Science Part A*, *60*, 18–28. https://doi.org/10.1016/j.shpsa.2016.08.003

Autzen, B. (2021). Is the replication crisis a base-rate fallacy? *Theoretical Medicine and Bioethics*, *42*(5), 233–243. https://doi.org/10.1007/s11017-022-09561-8

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: Evidence on Questionable Research Practices: The Good, the Bad, and the Ugly. *Journal of Business and Psychology*, *31*(3), 323–338. https://doi.org/10.1007/s10869-016-9456-7

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. https://doi.org/10.1038/483531a

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in Science: Improving the Standard for Basic and Preclinical Research. *Circulation Research*, *116*(1), 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819

Bird, A. (2021). Understanding the Replication Crisis as a Base Rate Fallacy. *The British Journal for the Philosophy of Science*, *72*(4), 965–993. https://doi.org/10.1093/bjps/axy051

Bleich, A., & Hansen, A. K. (2012). Time to include the gut microbiota in the hygienic standardisation of laboratory rodents. *Comparative Immunology, Microbiology and Infectious Diseases*, *35*(2), 81–92. https://doi.org/10.1016/j.cimid.2011.12.006

Carbone, L. (2020). Do "Prey Species" Hide Their Pain? Implications for Ethical Care and Use of Laboratory Animals. *Journal of Applied Animal Ethics Research*, *2*(2), 216–236. https://doi.org/10.1163/25889567-BJA10001

Carbone, L., & Austin, J. (2016). Pain and Laboratory Animals: Publication Practices for Better Data Reproducibility and Better Animal Welfare. *PLOS ONE*, *11*(5), e0155001. https://doi.org/10.1371/journal.pone.0155001

Davies, G. (2021). Locating the 'culture wars' in laboratory animal research: National

constitutions and global competition. *Studies in History and Philosophy of Science Part A*, *89*, 177–187. https://doi.org/10.1016/j.shpsa.2021.08.010

Davies, G., Greenhough, B., Hobson-West, P., & Kirk, R. G. W. (2018). Science, Culture, and Care in Laboratory Animal Research: Interdisciplinary Perspectives on the History and Future of the 3Rs. *Science, Technology, & Human Values*, *43*(4), 603–621. https://doi.org/10.1177/0162243918757034

Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science*, *67*, 559–579.

Eggel, M., & Würbel, H. (2021). Internal consistency and compatibility of the 3Rs and 3Vs principles for project evaluation of animal research. *Laboratory Animals*, *55*(3), 233–243. https://doi.org/10.1177/0023677220968583

Elliott, K. C., & Richards, T. (Eds.). (2017). *Exploring Inductive Risk: Case Studies of Values in Science*. Oxford University Press.

Engber, D. (2016a, April 19). Cancer Research Is Broken. *Slate*. https://slate.com/technology/2016/04/biomedicine-facing-a-worse-replication-crisis-than-the-one-plaguing-psychology.html

Engber, D. (2016b, April 19). *Think Psychology's Replication Crisis Is Bad? Welcome to the One in Medicine.* Slate Magazine. https://slate.com/technology/2016/04/biomedicine-facing-a-worse-replication-crisis-than-the-one-plaguing-psychology.html

Feest, U. (2019). Why Replication Is Overrated. *Philosophy of Science*, *86*(5), 895–905. https://doi.org/10.1086/705451

Feest, U. (2024). What is the Replication Crisis a Crisis Of? *Philosophy of Science*, 1–11. https://doi.org/10.1017/psa.2024.2

Fidler, F., & Wilcox, J. (2018). Reproducibility of Scientific Results. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Winter 2018 Edition)*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility/

Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015). The Economics of Reproducibility in Preclinical Research. *PLOS Biology*, *13*(6), e1002165. https://doi.org/10.1371/journal.pbio.1002165

Frommlet, F., & Heinze, G. (2021). Experimental replications in animal trials. *Laboratory Animals*, *55*(1), 65–75. https://doi.org/10.1177/0023677220907617

Georgiou, P., Zanos, P., Mou, T.-C. M., An, X., Gerhard, D. M., Dryanovski, D. I., Potter, L. E., Highland, J. N., Jenne, C. E., Stewart, B. W., Pultorak, K. J., Yuan, P., Powels, C. F., Lovett, J., Pereira, E. F. R., Clark, S. M., Tonelli, L. H., Moaddel, R., Zarate, C. A., … Gould, T. D. (2022). Experimenters' sex modulates mouse behaviors and neural responses to ketamine via corticotropin releasing factor. *Nature Neuroscience*, *25*(9), 1191–1200. https://doi.org/10.1038/s41593-022-01146-x

Graham, A. L. (2021). Naturalizing mouse models for immunology. *Nature Immunology*, *22*(2), Article 2. https://doi.org/10.1038/s41590-020-00857-2

Green, S. (2024). *Animal Models of Human Disease*. Cambridge University Press.

Guttinger, S. (2018). Replications Everywhere. *BioEssays*, *40*(7), 1800055. https://doi.org/10.1002/bies.201800055

Guttinger, S. (2019). A New Account of Replication in the Experimental Life Sciences. *Philosophy of Science*, *86*(3), 453–471. https://doi.org/10.1086/703555

Guttinger, S. (2020). The Limits of Replicability. *European Journal for Philosophy of Science*, *10*(2), 10. https://doi.org/10.1007/s13194-019-0269-1

Guttinger, S., & Love, A. C. (2019). Characterizing scientific failure. *EMBO Reports*, *20*(9), e48765. https://doi.org/10.15252/embr.201948765

Hackett, E. J. (Ed.). (2008). *The Handbook of Science and Technology Studies* (3rd ed.). MIT Press.

Hansen, A. K. (2021). Microbiology and Microbiome. In J. M. Sánchez Morgado & A. Brønstad (Eds.), *Experimental Design and Reproducibility in Preclinical Animal Studies* (Vol. 1, pp. 77–104). Springer International Publishing. https://doi.org/10.1007/978-3-030-66147-2_4

Hawkins, P., Morton, D. B., Burman, O., Dennison, N., Honess, P., Jennings, M., Lane, S., Middleton, V., Roughan, J. V., Wells, S., & Westwood, K. (2011). A guide to defining and implementing protocols for the welfare assessment of laboratory animals: Eleventh report of the BVAAWF/FRAME/RSPCA/UFAW Joint Working Group on Refinement. *Laboratory Animals*, *45*(1), 1–13. https://doi.org/10.1258/la.2010.010031

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., Schulz, K. F., & Tibshirani, R. (2014). Increasing Value and Reducing Waste in Research Design, Conduct, and Analysis. *The Lancet*, *383*(9912), 166–175. https://doi.org/10.1016/S0140-6736(13)62227-8

Jirkof, P., & Potschka, H. (2021). Effects of Untreated Pain, Anesthesia, and Analgesia in Animal Experimentation. In J. M. Sánchez Morgado & A. Brønstad (Eds.), *Experimental Design and Reproducibility in Preclinical Animal Studies* (Vol. 1, pp. 105–126). Springer International Publishing. https://doi.org/10.1007/978-3-030-66147-2_5

Karp, N. A., Pearl, E. J., Stringer, E. J., Barkus, C., Ulrichsen, J. C., & Sert, N. P. du. (2022). A qualitative study of the barriers to using blinding in in vivo experiments and suggestions for improvement. *PLOS Biology*, *20*(11), e3001873. https://doi.org/10.1371/journal.pbio.3001873

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., & Altman, D. G. (2010). Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. *PLoS Biology*, *8*(6), e1000412.

Knorr Cetina, K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press.

LaFollette, H. (2011). Animal Experimentation in Biomedical Research. In T. L. Beauchamp & R. G. Frey (Eds.), *The Oxford Handbook of Animal Ethics* (pp. 796–825). Oxford University Press.

LaFollette, H., & Shanks, N. (1993). Animal Models in Biomedical Research: Some Epistemological Worries. *Public Affairs Quarterly*, *7*(2), 113–130. JSTOR.

Latour, B., & Woolgar, S. (1979). *Laboratory Life: The Social Construction of Scientific Facts*. Sage.

Leenaars, C. H. C., Kouwenaar, C., Stafleu, F. R., Bleich, A., Ritskes-Hoitinga, M., De Vries, R. B. M., & Meijboom, F. L. B. (2019). Animal to Human Translation: A Systematic Scoping Review of Reported Concordance Rates. *Journal of Translational Medicine*, *17*(1), 223. https://doi.org/10.1186/s12967-019-1976-2

Lemoine, M. (2017). Animal Extrapolation in Preclinical Studies: An Analysis of the Tragic Case of TGN1412. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *61*, 35–45. https://doi.org/10.1016/j.shpsc.2016.12.004

Leonelli, S. (2018). Rethinking Reproducibility as a Criterion for Research Quality. In L. Fiorito, S. Scheall, & C. E. Suprinyak (Eds.), *Research in the history of economic thought and methodology. Including a symposium on Mary Morgan: Curiosity, Imagination, and Surprise* (pp. 129–146). Emerald Publishing. https://doi.org/10.1108/S0743-41542018000036B009

Leung, V., Rousseau-Blass, F., Beauchamp, G., & Pang, D. S. J. (2018). ARRIVE has not ARRIVEd: Support for the ARRIVE (Animal Research: Reporting of in vivo Experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia. *PLOS ONE*, *13*(5), e0197882. https://doi.org/10.1371/journal.pone.0197882

Lohse, S. (2021). Scientific Inertia in Animal-Based Research in Biomedicine. *Studies in History and Philosophy of Science Part A*, *89*, 41–51. https://doi.org/10.1016/j.shpsa.2021.06.016

Lohse, S., Wasmer, M., & Reydon, T. (2020). Integrating Philosophy of Science Into Research on Ethical, Legal and Social Issues in the Life Sciences. *Perspectives on Science*, *28*(6), 700–736.

Lowe, J. W. E., Leonelli, S., & Davies, G. (2019). Training to Translate: Understanding and Informing Translational Animal Research in Pre-Clinical Pharmacology. *Tecnoscienza*, *10*(2), 5–30.

Machery, E. (2020). What Is a Replication? *Philosophy of Science*, *87*(4), 545–567. https://doi.org/10.1086/709701

Macleod, M., & Mohan, S. (2019). Reproducibility and Rigor in Animal-Based Research. *ILAR Journal*, *60*(1), 17–23. https://doi.org/10.1093/ilar/ilz015

Macpherson, A. J., & McCoy, K. D. (2015). Standardised animal models of host microbial mutualism. *Mucosal Immunology*, *8*(3), 476–486. https://doi.org/10.1038/mi.2014.113

Magee, C. (2013, January 23). *Nine Out of Ten Statistics Are Taken Out of Context*. Understanding Animal Research. https://www.understandinganimalresearch.org.uk/news/nine-out-of-ten-statistics-are-taken-out-of-context

Mak, I. W., Evaniew, N., & Ghert, M. (2014). Lost in Translation: Animal Models and Clinical Trials in Cancer Treatment. *American Journal of Translational Research*, *6*(2), 114–118.

Nelson, N. C. (2021). Understand the real reasons reproducibility reform fails. *Nature*, *600*(7888), 191–191. https://doi.org/10.1038/d41586-021-03617-w

Neumann, K., Grittner, U., Piper, S. K., Rex, A., Florez-Vargas, O., Karystianis, G., Schneider, A., Wellwood, I., Siegerink, B., Ioannidis, J. P. A., Kimmelman, J., & Dirnagl, U. (2017). Increasing efficiency of preclinical research by group sequential designs. *PLOS Biology*, *15*(3), e2001307. https://doi.org/10.1371/journal.pbio.2001307

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press.

Pashler, H., & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science*, *7*(6), 531–536. https://doi.org/10.1177/1745691612463401

Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Karp, N. A., Lazic, S. E., Lidster, K., MacCallum, C. J., Macleod, M., … Würbel, H. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLOS Biology*, *18*(7), e3000410. https://doi.org/10.1371/journal.pbio.3000410

Piper, S. K., Grittner, U., Rex, A., Riedel, N., Fischer, F., Nadon, R., Siegerink, B., & Dirnagl, U. (2019). Exact replication: Foundation of science or game of chance? *PLOS Biology*, *17*(4), e3000188. https://doi.org/10.1371/journal.pbio.3000188

Pound, P., & Bracken, M. B. (2014). Is Animal Research Sufficiently Evidence Based to Be a Cornerstone of Biomedical Research? *BMJ*, *348*, g3387. https://doi.org/10.1136/bmj.g3387

Pound, P., Ebrahim, S., Sandercock, P., Bracken, M. B., & Roberts, I. (2004). Where Is the Evidence That Animal Research Benefits Humans? *BMJ: British Medical Journal*, *328*(7438), 514–517.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets? *Nature Reviews Drug Discovery*, *10*(9), 712–712. https://doi.org/10.1038/nrd3439-c1

Ravn, T., & Sørensen, M. P. (2021). Exploring the Gray Area: Similarities and Differences in Questionable Research Practices (QRPs) Across Main Areas of Research. *Science and Engineering Ethics*, *27*(4), 40. https://doi.org/10.1007/s11948-021-00310-z

Richter, S. H., Garner, J. P., & Würbel, H. (2009). Environmental Standardization: Cure or Cause of Poor Reproducibility in Animal Experiments? *Nature Methods*, *6*(4), 257–261. https://doi.org/10.1038/nmeth.1312

Richter, S. H., Garner, J. P., Zipser, B., Lewejohann, L., Sachser, N., Touma, C., Schindler, B., Chourbaji, S., Brandwein, C., Gass, P., Stipdonk, N. van, Harst, J. van der, Spruijt, B., Võikar, V., Wolfer, D. P., & Würbel, H. (2011). Effect of Population Heterogenization on the Reproducibility of Mouse Behavior: A Multi-Laboratory Study. *PLOS ONE*, *6*(1), e16461. https://doi.org/10.1371/journal.pone.0016461

Rosshart, S. P., Herz, J., Vassallo, B. G., Hunter, A., Wall, M. K., Badger, J. H., McCulloch, J. A., Anastasakis, D. G., Sarshad, A. A., Leonardi, I., Collins, N., Blatter, J. A., Han, S.-J., Tamoutounour, S., Potapova, S., Foster St. Claire, M. B., Yuan, W., Sen, S. K., Dreier, M. S., … Rehermann, B. (2019). Laboratory mice born to wild mice have natural microbiota and model human immune responses. *Science*, *365*(6452), eaaw4361. https://doi.org/10.1126/science.aaw4361

Russell, W. M. S., & Burch, R. L. (1959). *The Principles of Humane Experimental Technique*. Methuen.

Sánchez Morgado, J. M., & Brønstad, A. (Eds.). (2021). *Experimental Design and Reproducibility in Preclinical Animal Studies* (Vol. 1). Springer International Publishing. https://doi.org/10.1007/978-3-030-66147-2

Sánchez-Morgado, J. M., Brønstad, A., & Pritchett-Corning, K. (2021). Animal and Environmental Factors That Influence Reproducibility. In J. M. Sánchez Morgado & A. Brønstad (Eds.), *Experimental Design and Reproducibility in Preclinical Animal Studies* (Vol. 1, pp. 53–75). Springer International Publishing. https://doi.org/10.1007/978-3-030-66147-2_3

Seyhan, A. A. (2019). Lost in translation: The valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. *Translational Medicine Communications*, *4*(1), 18. https://doi.org/10.1186/s41231-019-0050-7

Shavit, A., Ellison, A. M., & Kress, J. W. (Eds.). (2017). *Stepping in the same river twice: Replication in biological research*. Yale University Press.

Spanagel, R. (2022). Ten Points to Improve Reproducibility and Translation of Animal Research. *Frontiers in Behavioral Neuroscience*, *16*. https://www.frontiersin.org/articles/10.3389/fnbeh.2022.869511

Sterne, J. A. C., & Smith, G. D. (2001). Sifting the evidence—What's wrong with significance tests? *Physical Therapy*, *81*(8), 1464–1469. https://doi.org/10.1093/ptj/81.8.1464

Vatsos, I. N. (2017). Standardizing the microbiota of fish used in research. *Laboratory Animals*, *51*(4), 353–364. https://doi.org/10.1177/0023677216678825

Voelkl, B., Vogt, L., Sena, E. S., & Würbel, H. (2018). Reproducibility of Preclinical Animal Research Improves with Heterogeneity of Study Samples. *PLOS Biology*, *16*(2), e2003693. https://doi.org/10.1371/journal.pbio.2003693

Voelkl, B., Würbel, H., Krzywinski, M., & Altman, N. (2021). The standardization fallacy. *Nature Methods*, *18*(1), Article 1. https://doi.org/10.1038/s41592-020-01036-9

Witjes, V. M., Boleij, A., & Halffman, W. (2020). Reducing versus Embracing Variation as Strategies for Reproducibility: The Microbiome of Laboratory Mice. *Animals*, *10*(12), Article 12. https://doi.org/10.3390/ani10122415

Würbel, H. (2000). Behaviour and the standardization fallacy. *Nature Genetics*, *26*(3), Article 3. https://doi.org/10.1038/81541