European Journal for Philosophy of Science (2020) 10:10
https://doi.org/10.1007/s13194-019-0269-1

PAPER IN GENERAL PHILOSOPHY OF SCIENCE

The limits of replicability

Stephan Guttinger[1]

Received: 14 June 2019 / Accepted: 11 November 2019/Published online: 15 January 2020

Abstract
Discussions about a replicability crisis in science have been driven by the normative claim that all of science should be replicable and the empirical claim that most of it isn't. Recently, such crisis talk has been challenged by a new localism, which argues a) that serious problems with replicability are not a general occurrence in science and b) that replicability itself should not be treated as a universal standard. The goal of this article is to introduce this emerging strand of the debate and to discuss some of its implications and limitations. I will in particular highlight the issue of demarcation that localist accounts have to address, i.e. the question of how we can distinguish replicable science from disciplines where replicability does not apply.

Keywords Replication crisis · Replicability · New localism · Demarcation problem · Trust-establishing practices

1 Introduction

The debate about replicability issues in the experimental sciences goes back many decades (in particular in fields such as social psychology), but it has erupted with particular force after 2011 with the publication of several key studies in both biomedicine and the psychological sciences (Prinz et al. 2011; Simmons et al. 2011; Begley and Ellis 2012; Pashler and Wagenmakers 2012). Since then talk of a 'crisis' has rapidly accelerated (Fanelli 2018) and expanded to other fields (e.g. health informatics (Coiera et al. 2018) or artificial intelligence research (Hutson 2018)). This expansion of the debate has been accompanied by the frequent use of generalising terminology, with key commentators talking of 'the' replication crisis in science and proposing general measures to tackle it. Underlying this talk of a crisis is not just empirical data on replication failures but also the fundamental assumption that replicability represents a universal epistemic standard for science.

✉ Stephan Guttinger
s.m.guettinger@lse.ac.uk

[1] London School of Economics, Centre for Philosophy of Natural and Social Science, Houghton Street, Lakatos Building, London WC2A 2AE,, UK

Recently, however, this broad way of approaching the issue has come under increased scrutiny. Based on empirical and conceptual analyses, several authors have argued a) that issues with replicability are not a general problem in science and b) that the ideal of replicability does not universally apply to all disciplines. This new localism not only implies that significant problems are limited to specific sub-fields, but also that high failure rates are not automatically a sign of crisis. Both claims together raise important challenges for current debates about the status quo of science and about the development of adequate policy measures.

The goal of this paper is to analyse the growing literature on the new localism in the replication crisis debate; to discuss its implications for the debate about science policy; and to address open questions the new accounts still face.

In the first part of the paper, I will analyse the two claims that I see as central to the new localism: the empirical claim a) (section 2), and the normative claim b) (section 3). I will argue that these claims are not only plausible but also important, as they directly affect the debate about science policy. In the second part of the paper, I will focus on a key question that the normative claim has to address, namely, the question of demarcation: how can we identify fields of science where replicability does not apply as an epistemic standard and for which a high failure rate for replications might therefore be the norm? I will show that most existing accounts draw this line using the living/non-living distinction and argue that new findings from the postgenomic life sciences suggest that this line is too narrow (section 4).

2 The reproducibility crisis in context

The replication crisis is often discussed in broad terms, with many commentators presenting the problem as one of science in general and possible solutions as widely applicable measures. Daniel Sarewitz, for instance, claims that:

"Science, […] our one source of objective knowledge, is in deep trouble [as] much of this supposed knowledge is turning out to be contestable, unreliable, unusable, or flat-out wrong" (Sarewitz 2016, p. 5).

Similarly, Roger Peng talks about 'the' crisis in science when he discusses how better statistical training could improve science's status quo (Peng 2015). Francis Collins and Lawrence Tabak claim that "[r]eproducibility is potentially a problem in all scientific disciplines" (Collins and Tabak 2014).

These worries about a crisis in science are based on the idea that replicability itself is a universal standard for reliable science. Replicability is often called a "cornerstone of science" (e.g., Simons 2014; eLife editorial 2017) and unbiased replication mechanisms are seen as essential for "maintaining high levels of scientific credibility" (Ioannidis 2012, p. 645). Without replications and replicability, so the idea, science as a whole lacks trustworthiness and credibility.

But despite the pervasiveness of generalisations, the debate about replication in science has also been characterised by an awareness of the importance of local context. This localism was not always explicit, but it can, for instance, be seen in the fact that whole fields, such as physics or chemistry, are usually excluded from

the debate.[1] As I will show in this and the next section, the emphasis on the importance of the local research context has gained more urgency and support in recent years. Here, I will first look at the empirical claim that there is no general crisis in science and that replicability issues are likely a local problem. Before I do so, however, I will also make a few remarks on the conceptual landscape of the debate.

2.1 Understanding replicability

The debate about the replication crisis is complicated by the fact that there is still no firm consensus on how exactly to define the terms used in the debate (Schmidt 2009; Goodman et al. 2016; Nosek and Errington 2017). What is clear is that 'replicability' and 'replication' are related but different concepts: a replication is an actual attempt to re-produce an earlier finding/experiment (or the outcome of such an attempt), whereas 'replicability' is a quality of an experiment/observation or a scientific finding.[2] Importantly, replicability is often also turned into a norm, i.e. the assumption that a finding or experiment should be replicable in order to be reliable.

This last point matters as it highlights two connected but different strands of the debate: first, there is a general debate about the different forms that replication can take on in research. This can be seen as a descriptive exercise, trying to capture the practices researchers associate with the term 'replication'. Second, there is a normative debate about the forms of replication that scientists should adhere to. It is within this second strand of the debate that the idea of a crisis has emerged.

Scientists and other commentators have identified a variety of replication types over the years.[3] Two types that have come to dominate the discussion are 'direct' and 'conceptual' replications. Both terms have received somewhat differing definitions and there is in particular no consensus on what 'direct' means in the context of replication. However, there are some key features that many authors seem to agree on, namely 1) that in a direct replication the same experimental protocol should be applied to the same kind of materials (for instance, individuals taken from the population originally studied, or the antibodies or cell lines used in the original experiment) and 2) that such an experiment should give an outcome that is the same or at least similar to that originally obtained. 'Similar' here means that the measured effect size could be smaller or bigger than the original, but that the direction of the effect should be the same (for instance, 'Overexpression of gene X accelerates cell growth', or 'Individuals presented with intervention X are more likely to do Y'). Findings and experiments are seen as reliable and trustworthy if they can be replicated in this particular manner.

A 'conceptual' replication, on the other hand, is often defined as an attempt to see an effect in the same direction as that originally reported using a different experimental protocol and/or materials. This type of replication is often linked by commentators to

[1] But note that there have been discussions in chemistry regarding the extent to which the field could be affected by the issue (Bergman and Danheiser 2016).

[2] Note that some authors distinguish between 'replications' and 'reproductions' (or 'replicability' and 'reproducibility'), whereas others – including the author of this paper – use the terms interchangeably.

[3] See (Barba 2018; Fidler and Wilcox 2018; Plesser 2018) for analyses of this diverse and complex conceptual landscape.

the goal of generalizing a finding or of testing its robustness, rather than assessing its reliability (Schmidt 2009). Some have suggested that such experiments should therefore be labelled as 'extensions' rather than 'replications' (Zwaan et al. 2018).

Apart from these two forms of replications there is a wide variety of further meanings associated with the term 'replication'. Leonelli (2018), for instance, describes 'scoping reproducibility', which consists of re-running experiments in order to identify potential sources of variation in experimental outcomes. Whilst this and other practices will form part of the experimental repertoire, this does not mean that they are also elevated to the level of a general epistemic norm. If researchers fail to identify the factors that cause variation in an outcome, this is unlikely to trigger talk of a fundamental crisis in science. The case is different, however, for the idea of direct replication. What triggered and further fuelled the crisis narrative were reported failures to directly reproduce existing results. And it is also this strict ideal of replicability that the new localists have set their eyes on. When I therefore speak of the replicability norm, I refer to this narrower understanding of replicability.

2.2 Understanding the actual extent of the problem

Following the emergence of the replication crisis narrative in 2011, several large-scale replication studies have been set up to assess the actual extent of the problem in different disciplines, ranging from the psychological sciences, to cancer research or even experimental economics. These studies, however, have so far failed to produce a clear picture of science's status quo.[4]

The largest of these studies, the 'Reproducibility Project: Psychology' (RPP), attempted to replicate 100 studies published in three psychology journals in the year 2008, an effort that resulted in a worryingly low success rate of only 39% (Open Science Collaboration 2015). Other studies that also focused on the psychological sciences painted a slightly more optimistic picture. The Many Labs project, which analysed how differences in settings/samples affect the variation of experimental outcomes, found success rates for replications between 50% (if a strict significance criterion was applied) (Klein et al. 2018), and 77% (Klein et al. 2014). These studies, however, used relatively small and non-random samples of studies (28 and 13, resp.) and deliberately included findings that were known to be robust (Klein et al. 2018). This could explain, in part at least, the higher success rates that were observed, compared to the RPP.[5]

Preliminary results from an ongoing replication study in pre-clinical cancer research ('Reproducibility Project: Cancer Biology; (Errington et al. 2014)) also indicate a higher success rate for replications than originally reported in this field (early studies by (Begley and Ellis 2012) and (Prinz et al. 2011) reported surprisingly low success rates for replications of 11% and 20-25%, resp.). Of the 14 replications completed as of September 2019 two could not be interpreted. Of the remaining 12 studies, three failed to reproduce the original experiments they intended to reproduce, resulting in a failure

[4] Fanelli is more optimistic as he claims the findings are "either reassuring or inconclusive" (Fanelli 2018, p. 2629).

[5] On the other hand, several authors argued that the findings of the RPP might be underestimating the actual amount of replicable data in psychology. See (Fanelli 2018) for a discussion.

rate of 25% (based on a total of 12 useable studies). Of the successful nine replications five were qualified as a full success whereas four managed to reproduce some results, but not others. Depending on how these mixed outcomes are interpreted, the success rate for the 'Reproducibility Project: Cancer Biology' is currently between 42% and 75%. Again, the relatively low sample size somewhat limits the conclusions one can draw from this study at this point in time.

Other fields appear to be doing marginally better than psychology (or cancer research, if a strict success measure is applied to the above studies). A study on experimental economics indicated a relatively high success rate for this particular field of 61%–89%, depending on the success measure used (Camerer et al. 2016). A similar study on the literature in social science found a success rate for replications of 62% (Camerer et al. 2018).

Overall, these results give an inconclusive picture of the state science is in. For both the social and behavioural sciences (Camerer et al. 2018), as well as pre-clinical cancer research, the best estimate of the amount of replicable studies is somewhere between 35% and 75%. Given this relatively weak and vague empirical basis, it is difficult to draw precise and definitive conclusions about the state of current research. At most, what we can derive from the data we have is that some disciplines *could* be in trouble, whereas other fields might not necessarily be in a significant crisis.

2.3 A local crisis?

To circumvent some of the above-mentioned problems researchers have relied on alternative methods for assessing the extent and the distribution of replication issues in science. Daniele Fanelli, for instance, uses metascience analyses, and in particular data on the prevalence of the drivers of the crisis (or what are presumed to be drivers of the crisis), as a proxy measure for replication issues.

One potential driver of the crisis that Fanelli and his co-workers looked at in more detail is the prevalence of publishing and reporting biases. These biases are often seen as potentially distorting the reliability and integrity of the scientific literature, because they influence what is being published and what is not. One example is the concept of a 'grey literature bias', which postulates that studies with small effect sizes are difficult to publish in traditional peer-reviewed journals and are therefore more likely to appear in less accessible outlets, such as personal communications, PhD theses, or other so-called 'grey literature'. This could mean that the scientific literature published in traditional outlets is biased towards studies that report large effect sizes. If that is the case, then the landscape of widely accessible findings would be significantly distorted.

Interestingly, Fanelli's analysis of the literature in different disciplines showed that whilst this and other publishing biases are present in science, they are not as prevalent as it has been feared. Moreover, the analysis showed that they are not equally pronounced in each field analysed (Fanelli et al. 2017). The same is the case for other potential drivers of low replicability, such as the low statistical power of studies (Fanelli 2018).

Whilst Fanelli previously supported a general crisis narrative, he now argues, based on his analysis of metascience data, that this narrative is at least partially misguided (Fanelli 2018). Some sub-disciplines might be facing significant issues, but we are not warranted to claim that whole disciplines or all of science are in crisis. As he puts it:

"Recent evidence from metaresearch studies suggests that issues with research integrity and reproducibility, while certainly important phenomena that need to be addressed, are […] heterogeneously distributed across subfields in any given area, which suggests that generalizations are unjustified" (ibid, 2628).

Fanelli stresses that the problem has to be thought of in a local rather than a generalising manner. Based on this analysis, he argues that the negative narrative of 'science-in-crisis' should be replaced with a more positive narrative of challenges and transformation. I will return to what his take on the replication crisis means for current policy debates in section 3.2.

3 Replicability as a local standard

A similar claim about the importance of the local research context has recently been made based on conceptual considerations rather than empirical analyses. This second component of what I call the new localism in the replication crisis debate targets the idea of replicability as a universal standard (de Rijcke and Penders 2018; Leonelli 2018; Nadin 2018; Penders et al. 2019).

The question of when replicability is a suitable epistemic norm is complex, but the critics of the norm have identified three aspects of research practice that can serve as guides for this debate: the type of questions researchers ask, the experimental setups they use, and the nature of the objects they analyse.

3.1 Working with different questions and setups

The first reason why not all research might have use for the ideal of replicability is the type of question it addresses. As de Rijcke and Penders (2018) point out, in the humanities the goal is often not just to develop single factual statements, such as 'Person X has created artwork Y', but to generate insight into questions of meaning or style. To answer such questions requires the researchers to take historical and social context into account. This also means that there can be multiple valid answers to the same question. Such plurality is often seen as an inherent feature of this type of research, and not as a sign of immaturity or flawed research practice. When addressing complex issues like, for instance, migration and what it means for society on a global scale, diverse positions are needed to develop sensible solutions that can work (Penders et al. 2019). This type of research cannot lead to 'the' single answer that scientists can arrive at with a highly standardized protocol. To impose a strict replicability criterion would therefore be counterproductive, as it would endanger the production of diverse insights and hypotheses in these fields (de Rijcke and Penders 2018; Penders et al. 2019).

Besides the type of research questions asked it is also the research setup itself that has to be factored in when assessing the suitability of the replicability criterion. Leonelli (2018) distinguishes between six different research settings that are relevant for this debate: 1) the validation and development of software in computer science, 2) standardised experimentation, 3) semi-standardised experimentation, 4) non-standard experimentation, 5) observational case studies, and 6) participant observation. The key dimension of this typology is standardization and its related issue of control: depending

on how much control each research setting allows the researcher to exert over key variables, the replicability criterion is more or less suitable.

In Leonelli's account, software development and participant observation represent the extreme ends of a spectrum: in the case of computer science, researchers can exactly control what is being used (existing data and code) and how a test is run. This means they have a high level of control. Because of this control, the ideal of direct replicability can be applied (and achieved in actual practice).

In the case of participant observation such tight control is not possible. There are too many variables that the researcher cannot fully control (including the effects of her own presence). As a result, each observer will encounter a highly idiosyncratic situation and come away with their own experience, observations, and interpretation of the case. Leonelli argues that the direct replicability of a finding is therefore not even expected by researchers in such research settings.[6]

This, however, does not mean that participant observation is less trustworthy and/or accountable. Because of their awareness of the importance of context, researchers engaged in participant observation invest more resources into making sure that the data they collect are carefully documented and preserved. The idea is to ensure accountability. As Leonelli explains:

"In non-standard types of inquiry, researchers typically recognize that direct reproducibility cannot function as an epistemic criterion for research quality, and instead devote care and critical thinking on documenting data production processes, examining the variation among their materials and environmental conditions, and strategizing about data preservation and dissemination" (Leonelli 2018).

This goes as far as documenting the researcher's personal circumstances during the observation process. The idea is to enable readers of the report to judge how factors such as the emotional state of the researcher might have affected the results and interpretations that were reported. In line with this, Penders and colleagues have highlighted that there are other 'technologies of accountability' – apart from replication – that researchers can use to ensure the production of reliable and trustworthy findings (Penders et al. 2019). Similarly, Jim Bogen has pointed out, based on an in-depth study of nineteenth century research in physiology, that replications might not always be necessary to assess the reliability of existing evidence, as researchers have other ways of doing so (Bogen 2001).[7]

Whilst the above arguments against replicability as a universal epistemic norm seem convincing when applied to the observational sciences and the humanities, it is not clear if and how this criticism extends to fields such as biology or chemistry, where semi- and non-standardised setups abound. It is here that the third aspect highlighted above, i.e. the nature of the object of interest, becomes a central factor in the debate. Before I turn to this issue in section 4, I will first draw an initial summary of what the new localism means for the debate about a replication crisis and science policy.

[6] There are two different claims mixed into arguments for localism: on the one hand the authors make an empirical claim, i.e. that researchers in particular fields don't actually use replicability as a standard. On the other hand, there is a normative claim that the standard should not be applied to certain fields.

[7] This raises an important issue that any localist approach has to address, namely the question of the nature and the importance of alternative 'trust-establishing practices' (TEPs), which might be used in areas where replicability does not apply as a norm. I will return to this point in section 5.

3.2 The implications of the new localism

A key claim of the new localism in the replication debate is that it does not make sense to assume (or impose) replicability as a general epistemic criterion for the quality of scientific findings; in some areas of research, such as observational research, the ideal simply is not applicable (and not applied by researchers in actual practice).

This claim has important consequences for the debate about a crisis in science as it shifts the goal posts: if we take the conceptual component of the new localism seriously, then concrete numbers for failure rates might not mean much anymore. A high failure rate for direct replications of, e.g., 80% could simply mean that the field is working as it should. It might not be a sign of a crisis because the standard that guides such a reading of these numbers does not apply in the first place. Once we factor in localism, the focus is no longer just on the actual failure rates in each discipline or sub-discipline, but on what these numbers might mean.

Some authors have gone further than this and claimed that failure in general is not a sign of crisis but just part of normal science (see, e.g., Firestein 2015; Redish et al. 2018). Whilst these views share a rejection of general crisis talk with the new localism, the latter represents a middle position in this often-heated debate, as its main point is not to question crisis talk per se. Rather, what is questioned is the indiscriminate expansion of the crisis narrative and the application of 'replicability' as a universal epistemic norm.

Understanding the limits of replicability matters for current policy debates. A key point of the new localism is that some fields should be kept out of the ongoing replication drive (see also (Bissell 2013) for an early critique of the replication drive). If, for instance, direct replicability is expected of all research then this could skew funding decisions against specific parts of science that are by nature less standardized (but not less important or accountable). A replication drive that is too broad could also lead researchers to pursue research protocols that ignore the inherent variation of the local context studied (Leonelli 2018). Furthermore, it might lead scientists to abandon the disciplinary standards of their field and lead to a reduction in the diversity of the findings and interpretations they produce (Penders et al. 2019).[8]

The conclusions we can draw from Fanelli's empirical analysis are similar: if the extent of the replication problem is not the same across science or even sub-disciplines, then broad changes to how research is funded and regulated might not be justified, especially as they could create harm in fields that don't need correcting in the first place. Furthermore, if intervention is required, the measures taken would have to be sensitive to the local context and to the concrete problems it suffers from, as the drivers of the problem are not always homogeneously distributed across science.

In sum, the focus on the importance of the local research context matters, because it changes the debate about how we measure and talk about reliable and trustworthy science; because it draws our attention to potentially negative side effects of the replication drive; and because it highlights the importance of local measures to tackle the problem.

[8] Stuart Firestein makes a similar case when he discusses the importance of failure for science and the need for a funding system that allows error-prone research to go ahead (Firestein 2015, ch. 12).

4 Drawing the line

Whilst localism is a credible and promising approach to the replication crisis debate, it is also an approach that raises new questions. One issue that stands out is that of demarcation: if replicability is not a universal ideal, then where do we draw the line? How can we know to which fields the norm applies and to which it doesn't?

The discussion in section 3 has shown that there might be well-defined poles where things are relatively straightforward – research in computer science at one end and participant observation in fields like anthropology at the other. But once we enter the realm of semi-standardised or non-standardised experiments, we encounter a large grey zone where it becomes much harder to decide if replicability should apply or not. This uncertainty matters because it directly affects debates about policy measures and about how they should be designed.

In this section, I will have a closer look at how different authors address this issue. I will show that the nature of the object of interest becomes crucial in this context. I will also highlight some key issues this particular demarcation criterion faces.

4.1 The question of standardisation

When it comes to the suitability of the replicability ideal, we have seen that high standardization and control over variables are crucial. Leonelli describes one class of experiments (apart from work in computer science) that can achieve such levels, namely what she calls 'standardised experiments'. Examples of such experiments can be found in the clinical sciences or in physics. In randomized controlled trials (RCTs), for instance, researchers apply rigorous controls to ensure that results are as reliable as possible. Leonelli argues that in such standardised experiments the idea of direct replication can be applied and is expected by scientists in actual practice.

However, such high standardization and control can only be achieved in a select few experimental contexts. In the majority of cases, especially in fields such as biology, researchers are dealing with what Leonelli calls 'semi-standardised' or 'non-standardised' experiments. The latter include, for instance, exploratory experiments where researchers investigate new entities or phenomena on which they have little or no background information. In these cases, standardization is not possible because the researchers might not know what to expect or what to control for. Leonelli argues that in such non-standardised settings the idea of direct replication (or conceptual replication for that matter) is 'not helpful' (Leonelli 2018, p. 10).

In the case of semi-standardised experiments there is also limited control that can be exerted (even though standardisation is possible and actually implemented). The problem here is not the availability of information or materials but the nature of the objects of interest. In particular, when working with living entities, such as model organisms, researchers are unable to control each and every aspect of the intervention because the objects of interest are highly context-sensitive (Leonelli 2018, p. 9). Animals, for instance, are responsive to changes in lighting, nutrition, or even the gender of the people handling them (Chesler et al. 2002; Lewejohann et al. 2006; Sorge et al. 2014). Such context-sensitivity imposes constraints on the level of control that can be achieved. Similar issues also affect experiments in psychology, where the human research subjects can be influenced by the research setting and the interaction with other people.

Leonelli does not make a normative statement when considering these grey zone cases. She simply makes the empirical claim that many researchers who work with semi- or non-standardized setups don't aim for direct replicability. As I will show in the following section, other authors who have commented on this issue argue that the living/non-living distinction can and should be used as a demarcation line in this debate.

4.2 Living entities and the problem of replicability

The idea that the nature of the entity of interest is crucial for any debate about replication and replicability is not new. Schmidt (2009) begins his discussion of the topic by highlighting that the idea of direct replicability builds on the fundamental assumption of the uniformity of nature. This assumption, he claims, is problematic as many fields deal with what he calls 'irreversible units' (ibid, p. 92). These are entities that are complex and not time-invariant (Schmidt does not specify what 'complex' means in this context). A key feature of these entities is that they have a memory of some sort; they 'accumulate history', as Schmidt puts it. In the case of human test subjects this historicity means that they might remember – consciously or sub-consciously – previous experiences and that this memory can affect their behaviour. Testing the same entity at a later time point might therefore not produce the same results, simply because the entity has transformed in key aspects. This historicity creates a problem for the idea of direct replicability as it undermines uniformity. Importantly, it goes beyond the context-sensitivity that Leonelli emphasises in her discussion of animal model research (which focuses on present influences on the test subject). Schmidt focuses his debate on the social sciences and only uses human test subjects as an example. It is therefore not clear what other entities he would include in his category of 'irreversible units'.

Crandall and Sherman (2016) approach the issue in a similar way. They claim that the idea of direct replication is a 'sensible proposition' in fields such as physics or biology, where the processes that matter for the outcome of an experiment are trans-historical and transcultural; changes in politics or language don't change the weight of an electron or the fold of a protein (Crandall and Sherman 2016).[9] But in a field like social psychology, a shift in language or socio-economic circumstances can profoundly affect the behaviour of the entities studied (a blog post by Michael Ramscar provides an in-depth analysis of how this might work (Ramscar 2015)). The entities studied in this field change over time, and their internal makeup depends not only on the context they are in but also the contexts they have experienced in the past. Crandall and Sherman not only highlight the importance of memory and experience, but also the fact that cultural factors, which shape a person's behaviour, can change over time. Their account is also more specific than Schmidt's, as they seem to propose a sharp line between natural sciences, such as physics or biology, and research in the social sciences.

Looking specifically at the humanities, de Rijcke and Penders (2018) follow a similar approach to that of Crandall and Sherman when they talk of 'interactive' and 'indifferent' kinds. Humans are examples of the first kind, DNA molecules of the

[9] Scholars in Science and Technology Studies, who have looked closely at the interconnections between science and politics, might disagree with such a strong realism.

second. Replicability, they argue, can only be used as a standard for the quality of data when doing research on indifferent kinds. In the humanities, where interactive beings are studied, this standard should not apply.

Whilst the above authors mainly focus on the role of historicity and plasticity in psychology and the humanities, others have, like Leonelli, focused more specifically on the situation in biology. Mihai Nadin, for instance, singles out biology because he draws a sharp line between the realm of living entities and that of 'dead matter' (Nadin 2018). He argues that there are fundamental differences in how change and causality work in these different realms, linking the idea of historicity and plasticity exclusively to living entities. Of particular importance to his account is the idea that the space of possibilities of living systems is continuously changing, an idea he takes from Giuseppe Longo's work (see, e.g., Longo 2017). He argues that:

"[T]he expectation of experiment reproducibility – legitimate in the decidable domains of the non-living (physics, chemistry) – is a goal set in contradiction to the nature of the change processes examined [in biology]" (Nadin 2018, p. 467).

Contrary to Crandall and Sherman (2016), he thus includes biology in the set of disciplines that pose significant problems for reproducibility.

The special status of the entities biology studies is also emphasised by Maël Montévil, who analyses the concept of 'measurement' in biology in the context of the replication crisis (Montévil 2019). He points out that the behaviour of systems analysed in physics is guided by an invariant underlying structure that can be captured in mathematical terms. This invariance (and invariance-preserving transformations) allows physicists to assume that generic conditions can be applied to generic objects when they deal with their objects of interest. This also means that replicability can be expected when particular features of physical systems are measured.

In biology, the situation is different. Here the organization of an entity depends on its past and current contexts, meaning that history matters for the (living) object of analysis (Montévil calls them 'diachronic objects' (ibid, p. 3)). Change also happens in physics of course, but there it is based on an unchanging mathematical structure, which is generic and therefore not context-sensitive (ibid, p. 5).[10]

The history-dependent nature of organisms also explains certain research practices in biology, for instance why researchers often exchange cell lines or other living model systems with each other (see also (Bissell 2013) on this point). The researchers have to make sure that they work with materials which have had the same experiences, and which are therefore more likely to display similar behaviours. The recent genealogy of the specimen is a feature of biological systems that has to be controlled as tightly as possible by the researcher to increase reliability (Montévil 2019, p. 10).

In summary, we see that a range of authors emphasise, in different ways, the importance of historicity and plasticity for debates about replicability. They highlight the fact that some types of entities are fundamentally time-dependent and that this interferes with the idea of uniformity that underlies the replicability ideal. Some authors, such as Nadin (2018) and Montévil (2019), link these features explicitly to living systems, whereas others talk more generally of 'irreversible' or 'interactive' entities.

[10] The importance of history for biological objects is also captured by Steven Rose's notion of 'lifelines' (Rose 1997).

Much of this debate implies that these distinctions should define a relatively clear boundary between science that can be treated as replicable and disciplines to which the replicability norm does not apply. The nature of the objects of interest affects the level of standardization and control that is possible in a field. This, in turn, affects the level of replicability that can be expected. However, as I will show in the next section, the practice of animal model researchers suggests that this line is not as clear as it might seem at first.

4.3 Rescuing replicability by abandoning standardization

The way in which researchers in animal model research deal with the problem of plasticity and historicity shows that they don't abandon the ideal of replicability when standardization and control become problematic. Similarly to what the above authors have emphasized, these scientists stress the importance of the history of the organism and its plasticity. As Voelkl and Würbel put it:

"[T]he response of an organism to an experimental treatment (e.g., a drug or a stressor) often depends not only on the properties of the treatment but also on the state of the organism, which is as much the product of past and present environmental influences as of the genetic architecture." They go on to conclude that "we should expect results to differ whenever an in vivo experiment is replicated" (Voelkl and Würbel 2016).

Even though this sounds like these researchers are ready to abandon the ideal of replicability, the opposite is the case: rather than turning their backs on replicability, they abandon the idea of standardization. Instead of seeing control in the form of uniform parameters as the solution (see, e.g., Festing 2004), some animal model researchers now see standardization as part of the problem (Würbel 2000; Richter 2017). This led them to coin the term 'standardization fallacy', which is defined as "the erroneous belief that reproducibility can be improved through ever more rigorous standardization" (Voelkl and Würbel 2016).

This shift in thinking leads to intriguing methodological consequences: to increase the reliability of their findings these researchers now introduce systematic heterogenization in their experimental setups, for instance through the use of animals with different genotypes, gender, or housing conditions. The idea is that the results of the experiment should become less sensitive to variations in these parameters, as the variation is already factored in.

Several studies using this approach showed that it can lead to a significant increase in the reproducibility of specific results. Using different mice strains in the same cage, for instance, led to a reduction of the variation in experimental outcomes (Walker et al. 2016). Varying environmental factors the animals are exposed to also had a positive effect on replicability (Richter et al. 2009, 2010, 2011). Furthermore, simulations suggested that multi-laboratory experiments, which automatically sample different housing and handling conditions, could increase reproducibility from 50% to 80% (Würbel 2017).

Whether this approach is applicable to other cases of semi-standardised research remains to be seen. What it shows, though, is that we are unlikely to find a general approach to the issue of control and replicability, as researchers will abandon the ideal in some cases, and re-invent their own experimental approaches in others in order to

stick to the ideal. When it comes to the grey zone of semi- and non-standardised experiments, a case-by-case analysis that pays close attention to actual research practice will therefore be more important than a single, general demarcation criterion. There is no general 'ought' that can be imposed here, and the living/non-living distinction can only serve as a rough guide for demarcation.

In the last section, I will turn my attention to a second issue this demarcation criterion faces, namely that of scope. I will argue that plasticity and historicity, which have been exclusively attributed to living systems, apply to other systems as well, in particular to macromolecular complexes such as DNA or proteins.[11] This extension matters as it suggests that the problem of localism is relevant to more fields than initially thought, extending beyond the realm of living things. The rise of post-genomic approaches to biological research, and in particular the field of environmental epigenetics, has had a huge role to play in this context.

4.4 Extending the lines

The shifts that were brought about by the postgenomic revolution over the last 10–15 years had several effects on biological theory and practice. The one that matters most for our discussion here is the shift in our understanding of historicity: some of the dynamics that were exclusively ascribed to living systems are now seen as features of other elements of biological systems as well.

This shift has been mainly due to methodological developments. New technologies, such as microarrays or high-throughput sequencing, have allowed researchers to gain new insights into the dynamics of the organism and into the importance of phenomena such as symbiosis (Guttinger and Dupré 2016). This has led researchers to a new understanding of the importance of context and history for the makeup of what were previously seen as 'mere' molecular systems.

Especially macromolecular complexes such as genomes or proteins are no longer seen as passive junks of matter. The genome, for instance, is now seen by some as a 'reactive' entity that is co-produced and maintained by a range of different processes (Gilbert 2003; Stotz 2006; Keller 2014). The genome has a sort of memory of past exposures (through epigenetic modifications of nucleotides or histone proteins) and its structure and behaviour are therefore not only defined by its sequence and its present context but also by past events; as some authors have argued, the genome has a lifespan of its own (Lappé and Landecker 2015). Because of these empirical and theoretical developments, fields such as molecular biology, genomics, or even biochemistry now have to be considered as areas of science where the ideal of replicability might hit its limits.

Interestingly, most of the accounts that I discussed in section 4.2 don't leave room for such an extension, as they insist on using the living/non-living distinction as a demarcation criterion. Montévil (2019), for instance, explicitly excludes biochemistry from the problems biological measurement faces. Nadin (2018) also seems to exclude the physico-chemical realm of peptides and other molecules from the problems living systems pose. Similarly, de Rijcke and Penders (2018) count DNA molecules as part of

[11] The claim is not that these systems are living. The point is rather that they share those features of living systems that affect experimental standardization and replicability.

the class of 'indifferent' entities. In essence, anything molecular is excluded in these accounts from the realm of plasticity and historicity.

What the recent developments in the postgenomic life sciences highlight is that these lines might be too narrow and that important questions about historicity, plasticity, standardization, and control also have to be asked in the molecular life sciences. Overall, this means that the new localism in the replication crisis debate is even more important than the authors discussed above claim, as it raises important questions for a broader range of disciplines regarding their methodologies and the production of reliable output. At the same time, the discussion in section 4.3 has shown that historicity and plasticity always have to be assessed in the context of actual practice. In themselves they are not a reason for scientists to shun the norm of replicability.

5 Conclusions

The goal of this article has been to introduce what I call the 'new localism' in the replication crisis debate and to address some of the issues this approach faces, in particular the question of demarcation. The new localism claims that issues with replicability might be a local problem for specific sub-fields of science and that replicability itself should not be treated as a universal epistemic norm. I have argued that this is a plausible and important extension of the current debate about replication in science.

Taking this new localism seriously opens the door to a more fine-grained and balanced approach to the debate about a crisis in science. It cautions against the implementation of broad new policy measures, as not all disciplines might need correcting.

The question of how we should identify the fields to which the replicability standard does or does not apply is not fully answered yet. We have seen that there are (at least) three different aspects of scientific practice that can be used to answer this question: the type of questions addressed, the setup used, and the nature of the objects analysed. The discussion in section 4 of this paper, however, has shown that there is still a significant grey zone of research practices where there might not be a clear answer to this question, and where a case-by-case analysis might be the only sensible way forward.

This grey zone importantly includes significant parts of biological research, which has been at the centre of the crisis debate in recent years. The concepts of historicity and plasticity have been identified as key factors that need to be considered in debates about replicability in this field. I have argued that recent developments in the postgenomic life sciences raise important questions about the scope of these concepts, implying that they apply much more broadly than most authors currently assume.

This debate about plasticity and historicity incidentally also highlights important new roles for philosophy of science in the debate about replicability. Philosophers already contribute to the replication crisis debate through reflections about conceptual and epistemic issues. But their insights into and contributions to complex ontological debates might prove equally helpful. Over the last decade or so, the question of how to define life, and of how to understand the plastic nature of biological systems, have been a key focus of philosophy of biology. The diverse literature on these topics could,

indirectly at least, make a crucial contribution to analyses of replicability in science, now that the new localism is opening up a new field of debate.

Last but not least, the authors discussed here have also highlighted that there might be more 'technologies of accountability', or what I referred to as 'trust-establishing practices' (TEPs), that scientists depend on, not just in the observational but potentially also in the experimental sciences. The new localism implies that the current framework within which replication is discussed is too narrow, a point that Bogen (2001) also has made forcefully. Understanding the landscape of TEPs is a further challenge any localist account must face, apart from the demarcation question. It is these practices that can be used to ensure accountability and reliability in the absence of strict replicability. To develop a more nuanced discussion about replicability and trust in science we therefore also have to expand our knowledge of TEPs and of how they are used in scientific practice, an endeavour to which philosophy of science, and in particular philosophy of experimentation, could make significant contributions.

References

Barba, L. A. (2018). Terminologies for reproducible research. *arXiv preprint, arXiv:1802.03311*.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531–533.

Bergman, R. G., & Danheiser, R. L. (2016). Reproducibility in chemical research. *Angewandte Chemie International Edition, 55*(41), 12548–12549.

Bissell, M. (2013). Reproducibility: The risks of the replication drive. *Nature News, 503*(7476), 333–334.

Bogen, J. (2001). Two as good as a hundred': Poorly replicated evidence in some nineteenth-century neuroscientific research. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 32*(3), 491–533.

Camerer, C. F.,et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. Nature Human Behaviour, 2(9), 637–644.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436.

Chesler, E. J., Wilson, S. G., Lariviere, W. R., Rodriguez-Zas, S. L., & Mogil, J. S. (2002). Influences of laboratory environment on behavior. *Nature Neuroscience, 5*(11), 1101–1102.

Coiera, E., Ammenwerth, E., Georgiou, A., & Magrabi, F. (2018). Does health informatics have a replication crisis? *Journal of the American Medical Informatics Association, 25*(8), 963–968.

Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature, 505*(7485), 612–613.

Crandall, C. S., & Sherman, J. F. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology, 66*, 93–99.

De Rijcke, S., & Penders, B. (2018). Resist calls for replicability in the humanities. *Nature, 560*(7716), 29.

eLife editorial. (2017). The challenges of replication. *eLife*, 6, e23693. https://doi.org/10.7554/eLife.23693.

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). Science forum: An open investigation of the reproducibility of cancer biology research. *Elife, 3*, e04333.

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences, 115*(11), 2628–2631.

Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences, 114*(14), 3714–3719.

Festing, M. F. (2004). Refinement and reduction through the control of variation. *Alternatives to Laboratory Animals, 32*(1_suppl), 259–263.

Fidler, F., & Wilcox, J. (2018) "Reproducibility of scientific results", *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility/>. (Accessed 24 May 2019).

Firestein, S. (2015). *Failure: Why science is so successful*. New York: Oxford University Press.

Gilbert, S. (2003). The reactive genome. In G. B. Muller & S. A. Newman (Eds.) *Origination of organismal form: Beyond the gene in developmental and evolutionary biology* (pp. 87–101). MIT Press.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science Translational Medicine, 8*(341), 341ps12.

Guttinger, S., & Dupré, J. (2016). "Genomics and Postgenomics", *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/genomics/>.

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science, 359*(6377), 725–726.

Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7*(6), 645–654.

Keller, E. F. (2014). From gene action to reactive genomes. *The Journal of Physiology, 592*(11), 2423–2429.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr., R. B., Bahník, S., Bernstein, M. J., Bocian, K., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142.

Klein, R. A., et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490.

Lappé, M., & Landecker, H. (2015). How the genome got a life span. *New Genetics and Society, 34*(2), 152–176.

Leonelli, S. (2018). "Re-Thinking Reproducibility as a Criterion for Research Quality." [Preprint]. URL: http://philsci-archive.pitt.edu/id/eprint/14352 (Accessed 2018-10-12).

Lewejohann, L., Reinhard, C., Schrewe, A., Brandewiede, J., Haemisch, A., Görtz, N., Schachner, M., & Sachser, N. (2006). Environmental bias? Effects of housing conditions, laboratory environment and experimenter on behavioral tests. *Genes, Brain and Behavior, 5*(1), 64–72.

Longo, G. (2017). How future depends on past and rare events in systems of life. *Foundations of Science, 23*(3), 443–474.

Montévil, M. (2019). Measurement in biology is methodized by theory. *Biology and Philosophy, 34*, 35–25. https://doi.org/10.1007/s10539-019-9687-x.

Nadin, M. (2018). Rethinking the experiment: Necessary (R) evolution. *AI & SOCIETY, 33*, 467–485. https://doi.org/10.1007/s00146-017-0705-8.

Nosek, B. A., & Errington, T. M. (2017). Reproducibility in cancer biology: Making sense of replications. *Elife, 6*, e23383.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528–530.

Penders, B., Holbrook, J. B., & de Rijcke, S. (2019). Rinse and repeat: Understanding the value of replication across different ways of knowing. *Publications, 7*, 52.

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance, 12*(3), 30–32.

Plesser, H. E. (2018). Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics*, 11, 76.

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), 712–713.

Ramscar, M. (2015). The unspeakable in the pursuit of the unrepeatable. https://ramscar.wordpress.com/2015/08/05/the-unspeakable-in-pursuit-of-the-unrepeatable/ (Accessed 2 June 2019).

Redish, A. D., Kummerfeld, E., Morris, R. L., & Love, A. C. (2018). Opinion: Reproducibility failures are essential to scientific inquiry. *PNAS, 115*(20), 5042–5046.

Richter, S. H. (2017). Systematic heterogenization for better reproducibility in animal experimentation. *Lab Animal, 46*(9), 343–349.

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.

Richter, S. H., Garner, J. P., Auer, C., Kunert, J., & Würbel, H. (2010). Systematic variation improves reproducibility of animal experiments. *Nature Methods, 7*(3), 167–168.

Richter, S. H., Garner, J. P., Zipser, B., Lewejohann, L., Sachser, N., Touma, C., Schindler, B., Chourbaji, S., Brandwein, C., Gass, P., & van Stipdonk, N. (2011). Effect of population heterogenization on the reproducibility of mouse behavior: A multi-laboratory study. *PLoS One, 6*(1), e16461.

Rose, S. (1997). *Lifelines: Biology, freedom, determinism*. London: Allen Lane.

Sarewitz, D. (2016). Saving science. *The New Atlantis, 49*, 4–40.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90–100.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False–positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76–80.

Sorge, R. E., Martin, L. J., Isbester, K. A., Sotocinal, S. G., Rosen, S., Tuttle, A. H., Wieskopf, J. S., Acland, E. L., Dokova, A., Kadoura, B., & Leger, P. (2014). Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nature Methods, 11*(6), 629–632.

Stotz, K. (2006). Molecular epigenesis: Distributed specificity as a break in the central dogma. *History and Philosophy of the Life Sciences, 28*(4), 533–548.

Voelkl, B., & Würbel, H. (2016). Reproducibility crisis: Are we ignoring reaction norms? *Trends in Pharmacological Sciences, 37*(7), 509–510.

Walker, M., Fureix, C., Palme, R., Newman, J. A., Ahloy, J. D., & Mason, G. (2016). Mixed-strain housing for female C57BL/6, DBA/2, and BALB/c mice: Validating a split-plot design that promotes refinement and reduction. *BMC Medical Research Methodology, 16*, 11.

Würbel, H. (2000). Behaviour and the standardization fallacy. *Nature Genetics, 26*(3), 263–263.

Würbel, H. (2017). More than 3Rs: The importance of scientific validity for harm-benefit analysis of animal research. *Lab Animal, 46*(4), 164–166.

Zwaan, R., Etz, A., Lucas, R., & Donnellan, M. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*, E120. https://doi.org/10.1017/S0140525X17001972.