The London School of Economics and Political Science

TOWARDS A MATURE SCIENCE OF OTHER MINDS

Charles Aaron Beasley

A thesis submitted to the Department of Philosophy, Logic and Scientific Method of the London School of Economics for the degree of Doctor of Philosophy

March 2023

Charles Aaron Beasley: *Towards a Mature Science of Other Minds*

DECLARATION:

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work, other than where I have clearly indicated that it is the work of others. The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party.

I declare that this thesis consists of 97923 words (excluding references).

ABSTRACT

From humble beginnings, minds in the natural world have come to take on endless forms. Yet, our best way of studying these minds has come up short. By its own lights, comparative cognitive science has repeatedly struggled to achieve the goals that it has set out for itself. Why is this, and what can be done about it? This dissertation is directed at answering these questions. It does so in two parts.

The first addresses the issue of replication; both the looming threat that a replication crisis holds for comparative cognitive science as well as the broader issue of what a replication is. Here, a deflationary account of replication is defended. This view is argued to hold implications for the practice of replication, interpreting replicability as a demarcation criterion, evaluating the replication crisis across the sciences, and evaluating results of replication experiments. Moreover, this analysis holds direct consequences for the evaluation of evidence in comparative cognitive science. Most pointedly, it is not 'mature' enough to experience a genuine replication crisis. Nevertheless, replication experiments, when properly framed, can still make valuable contributions to the discipline.

The second part builds on the analysis in the first and shows how comparative cognitive science is inhibited by what is termed the 'validity cycle'. Here, the causes of the validity cycle are initially identified (weak theory, underspecified hypotheses, and underdetermined experiments). Then, four ameliorative proposals that have been made are introduced and their shortcomings are analyzed. These are (1) adopting a bottom-up approach, (2) breaking cognitive capacities into dimensions, (3) formalizing the theories, and (4) identifying signatures of cognitive capacities. Modified versions of each are introduced. Taken together, it is argued that these modified proposals represent the most promising way forward for the science particularly when run simultaneously in a complementary manner.

For my parents; Leila Beasley and Glen Beasley

ACKNOWLEDGEMENTS

I am fortunate to have had a wealth of support and guidance over the past four years.

This dissertation would not exist in any form without my advisors Jonathan Birch and Roman Frigg. Jonathan has been a steady, encouraging, and generally brilliant advisor. I fear that I have only begun to learn all the lessons that I would want to from him, which makes concluding my PhD somewhat bittersweet. Roman has provided an essential critical voice to both the content of this project and my approach to it. He has provided insightful comments on every part of this dissertation and has repeatedly steered it in the right direction. I am deeply grateful to both of them for their excellent support and guidance.

I have two other unofficial 'advisors' that I am massively indebted to. The first is Kristin Andrews who has been an enthusiastic champion of my work, in a way that I probably don't always deserve, as well as a consistent source of personal support. The second is Richard Moore who has been a long-term source of academic and personal support. While he hasn't read a word of this dissertation, I would have never made it to the LSE without him.

The LSE is an objectively outstanding place to do philosophy. I am lucky enough to have benefited directly or indirectly from interactions with Ali Boyle, Richard Bradley, Liam Kofi Bright, Laurenz Hudetz, Christian List, Anna Mahtani, Miklós Rédei, Bryan Roberts, and Kate Vredenburgh. So many PhD students have passed through the LSE over the past four years that I have enormously benefited from. I cannot possibly begin to list all of them, but I am grateful for the consistently rich intellectual community that they created. That being said, I owe a special debt to Joe Roussos, who I co-founded the Conjectures and Refutations seminar with.

I've spent a significant amount of time at Cambridge during my PhD and have massively benefited from the members of the Kinds of Intelligence reading group. This would be a fundamentally different and undoubtedly worse project without these interactions.

Without Vaux East, the Castle, and the Arch, I would not be here today writing these sentences with malformed and calloused fingers. Without the HHV, Hog Hill, Lee Valley, and the hills of Kent, my mental health would be in the gutter, and I wouldn't have these bizarre tan lines.

I am nothing without my friends. I can't begin to list some without fear of excluding others, so I won't. Thank you for being there again and again, time after time. I love y'all.

Finally, thank you to my family Leila, Glen, Brandon, Nicole, Nico, and Luca, who have each been a steadfast source of inspiration and support, as well as my chosen family, Lora and Dimitry. Thank you for everything you have given me. I love y'all.

TABLE OF CONTENTS

CONCLUSION

Die Verwirrung und Öde der Psychologie ist nicht damit zu erklären, daß sie eine "junge Wissenschaft" sei; ihr Zustand ist mit dem der Physik z. B. in ihrer Frühzeit nicht zu vergleichen. […] Es bestehen nämlich, in der Psychologie, experimentelle Methoden und Begriffsverwirrung. […] Das Bestehen der experimentellen Methode läßt uns glauben, wir hätten das Mittel, die Probleme, die uns beunruhigen, loszuwerden; obgleich Problem und Methode windschief aneinander vorbei laufen.

Ludwig Wittgenstein (Philosophische Untersuchungen 371)

Psychology's confusion and barrenness is not to be explained by its being a "young science"; its state is not comparable with that of physics, for instance, in its beginnings. […] In psychology, there are experimental methods and conceptual confusion. […] The existence of the experimental method makes us think that we have the means of getting rid of the problems which trouble us; but problem and method pass one another by.

Ludwig Wittgenstein (Philosophical Investigations 371)[1]

[1] G.E.M. Ansombe's translation

Frank hammers
carrots
all day

it works

the earth
can't
leave us

CA Conrad, The Book of Frank

INTRODUCTION

0.1 Two Questions

From humble beginnings, minds in the natural world have come to take on endless forms. Yet, our best way of studying these minds has repeatedly struggled. The history of comparative cognitive science is replete with calls for deep reform (e.g., Beach 1950, Hodos and Campbell 1969). This is not necessarily indicative of a problem. It might even be interpreted as a sign of health in the discipline. However, over the past 20 years, as the vast array of challenges that the conventional practices of the discipline face have come into sharper focus, these calls have become increasingly urgent, concerning, and heated (e.g., Barrett 2010, Shettleworth 2010, Hemelrijk and Bolhuis 2011, Vonk and Shackelford 2013, Mikhalevich et al. 2017, among many others). In certain cases, calls for outright revolution have even gone so far as to argue that the discipline has been fundamentally wrecked (Penn 2011) or that it is not a legitimate science (Farrar and Ostojic 2019). To put it delicately, by many in the discipline's own lights, comparative cognitive science has repeatedly struggled to achieve the goals that it has set out for itself. This gives rise to two crucial questions.

(1) Why is this the case?
(2) Can it be improved?

In recent decades, there has been a marked swell of interest in the science of other minds. This has come with the emergence of research programs targeting a comparatively wide-variety of capacities in a wide-variety of species. The problem is this has not obviously translated into the broad generation of reliable cumulative knowledge as many have hoped.

It does not have to be this way. By enhancing the methodological and theoretical foundations of comparative cognitive science with the goal of diminishing the high degree of uncertainty regarding our knowledge of other minds, and setting appropriate expectations, how the science works can be ameliorated. While there is no reason to think that comparative cognitive science will be launched into the space-age any time soon, there is reason to be optimistic about it achieving and exceeding the goals that it has set out for itself in the long term.

0.2 The Approach

In the best cases, comparative cognitive science is highly interdisciplinary. It aims to draw on evidence from fields such as linguistics, psychology, neuroscience, philosophy, computer science, and anthropology, among others. Given this, multiple approaches to the science will be adopted simultaneously in this dissertation. The first, can be best characterized as straightforward 'philosophy of science' in the broad sense that aims to explain what comparative cognitive science does and make normative claims regarding how it should function. How does it produce the knowledge that it does and what amendments, if any, need to be made to these empirical procedures to improve how they function? The second is congruent with what has been called 'synthetic philosophy', or "a style of philosophy that brings together insights, knowledge, and arguments from the special sciences with the aim offering a coherent account of complex systems and connecting these to a wider culture or other philosophical projects (or both)" (Schliesser 2019; pg. 1). This should be seen as a narrower sense of philosophy of science: what Godfrey-Smith (2016) has called natural philosophy or the philosophy of nature. In this dissertation, these two approaches are tightly connected to one another and stand in a complimentary relationship.

In addition to this, the more traditional tools of analytic philosophy are taken up. This includes conceptual analysis/engineering, thought experiments, and logical analysis, among others. Each of

these have their clear strengths and weaknesses, and their suitability in any given situation is determined by the task at hand. These various methods culminate in a meta-philosophical approach that is rooted broadly in naturalism, interdisciplinarity, pluralism and complementarity.

Finally, a deep and sincere respect for the practitioners of comparative cognitive science is central to this project. Studying minds that are not directly observable is profoundly difficult. The past 50 years have been witness to innumerable borderline heroic attempts to make the research in the discipline as rigorous as possible. While this in no way means that the science is immune to criticism, as will become abundantly clear, it does mean that the starting point of inquiry is that the scientists working in this area are highly skilled and are engaged in serious and very difficult work. If any goal has not been sufficiently achieved, it is in not because comparative cognitive scientists have been sitting on their hands. The target of study has simply proven to be more difficult to grasp than many in the discipline could have possibly imagined.

0.3 The Plan

This dissertation unfolds over two expansive parts and resolves in a short forward-looking conclusion.

The first part defends the deflationary account of replication which deliberately renders the question 'what is a replication?' to be a trivial one. On this account, a replication is an experiment that is stipulated by an epistemic agent to be a test of a hypothesis regarding the causal structure of the effect tested in a target experiment. Focus here is placed on establishing whether a replication experiment is a valid test of the causal structure of the effect tested in the target experiment. This account deals with the issue of replication broadly considered, as well as how it pertains to comparative cognitive science. While comparative cognitive science is yet to experience a replication crisis in a way that many disciplines across the sciences have, this is largely due to the fact that replication experiments have not been widely attempted in the discipline. This has created a looming sense that a crisis could strike the foundations of comparative cognitive sciene at any moment. In adjacent disciplines, the onset of a replication crisis has created a broad sense of panic. This has generally not been productive. Accusation of fraud and bias have been routinely

leveraged against researchers whose work has not been successfully replicated and the fundamental methods that underwrite a variety of disciplines have been repeatedly thrown into question. There is a sense that radical change is needed across these affected sciences, but at the same time, there is no consensus regarding what this should amount to.

Given this, a more fundamental question is taken as the starting point; namely, what is a replication? This is because answering this question is necessarily upstream from any of the very difficult issues and proposals that scientists have been faced with in the wake of the replication crisis. This initially involves probing the types of replication that have featured most centrally in the debates around the replication crisis of the past decade; e.g. exact, direct, and conceptual replication. These accounts, which are broadly labelled 'naïve', are generally inadequate in so far as they lean heavily on an underspecified and intolerably flexible notion of 'relevant similarity.' Given this, the debates that have arisen between these conceptions, and about their inability to yield productive proposals, should be seen as unsurprising.

An analysis is then performed on two recent account accounts of replication that have attempted to get beyond the shortcomings of the 'naïve' accounts: the resampling account and the diagnostic account. The resampling account fails in at least two ways. First, it relies on a type of random sampling of various components of an experiment that in a high number of experimental situations, cannot possibly be implemented in practice. Second, the type of resampling that is proposed requires that every combination of the resampled components count as a replication experiment, and this leads to replication failures for absurd reasons. Taking this account seriously will lead to highly undiagnostic replication failures. The diagnostic account, by contrast fails because it is permissive in a way such that virtually any experiment could count as a replication. This undercuts its plausibility.

Given the inadequacy of these previous accounts of replication, the 'causal approaches', which are far and away the most promising, are then analyzed. There are four accounts that fall under this label; one by Norton (2015), one by Anjum and Mumford (2018), one by Feest (2019), and finally one by Irvine (2021). Norton's account hits a wall in so far as it is demanding in a way that potentially renders the practice of replication inert. Anum and Mumford's account is limited in so

far as undercuts replication as a practice by introducing an extremely demanding take on causality. Feest' account stumbles in so far as it renders replication unable to probe systematic errors. And finally, Irvine's account struggles because interpreting replications as exploratory experiments alone potentially confounds replication as a practice. While each of these accounts have clear limitations that are brought out through analysis, the basic idea of a causal approach to replication remains a promising one.

The causal approaches, moreover, have an interesting upshot. They allow for the introduction of a novel demarcation criterion for replication and the requirement of replicability. Here, an experiment or line of research should be subject to the replicability criterion *iff* there is an adequate understanding of the causal mechanism that is responsible for the effect in the target experiment. Otherwise, there is reason to think that there is too much uncertainty for the replication experiment to be properly diagnostic, and the replicability criterion should not be applied.

Causal accounts have another advantage. They allow for the dispensal of an often-invoked approach to replication; namely a typological one. Replication typologies are broadly intended to make sense of the diversity that is contained within replication as a practice. However, these accounts are broadly inadequate and miss the mark, in so far as they identify the ways in which replications can be similar or relevantly similar to one another. Here, this problem is identified in accounts by Leonelli (2018), Peng (2011), LeBel et al. (2017), and Hüffmeier et al. (2015). In varying ways, each of these authors try to break the practice of replication down into types. However, because they start with naïve conceptions and proceed accordingly, their typologies track potentially irrelevant aspects of the target experiments. Given this, typological accounts largely fail to deliver and are prone to being misguiding.

With the critiques of existing accounts established, the replication crisis at large is turned to with the intention of diagnosing what should be made of the broad replication failures that have spanned a number of scientific disciplines. Across this part of the dissertation, it is argued that the authors who have attempted to give diagnoses of the replication crisis in a general sense are not well situated to do so. Bird (2020) is first addressed, who argues that the replication crisis can be

interpreted as base rate fallacy. Against this, it is argued that he reaches beyond the evidence that is available to him, particularly given his claim concerning the status of theory in the disciplines that have been affected by the crisis. It is then argued that Lavelle (2020), who contests Bird's claim that more replications will be helpful to overcoming the crisis, goes wrong in so far as she has an unconstrained conception of what counts as a replication. Finally, Farrar and Ostojic's (2019) claims that comparative cognitive science is on the verge of a replication crisis is contested and it is argued that there can be either a theory crisis or a replication crisis taking place in the discipline, but that there cannot be in any meaningful sense both at the same time. Again, the overarching argument of this part of the dissertation is that in the most general sense, any claim to the replication crisis should be modest and center the necessarily high degree of uncertainty that comes with such diagnoses. Brash reforms should be avoided.

At this point, the deflationary account of replication is introduced, which is the core of this part of the dissertation. Here, a replication is conceived as an experiment that is a test of a hypothesis regarding the causal structure of a target experiment. It can do this in a way that is valid or invalid or even an unequivocal waste of time, and it should still be called a replication. Just as an experiment that is invalid is still an experiment, a replication experiment that is invalid is still a replication experiment. By adopting this approach, much of their mystery of what replications are dissolves and a host of first order tools of experimental design that can be used to evaluate them are opened up. This also ameliorates the process of interpreting them and clarifies the terms of the debate. The account is deflationary in the sense that it 'deflates' the highly contested task of defining what a replication is by reframing it in the familiar terms of first-order experimentation. Moreover, it is deflationary in the sense that it avoids the introduction of novel terminologies or typologies in favor of deferring to and exploiting existing tried and true features of first-order experimentation.

The core question then becomes the following one: Does the replication experiment measure what it is intended to measure? That is, is the replication experiment a valid one? This has the advantage of placing inquiry broadly within more well-trodden territory. Moreover, this approach makes sense of the relevant and irrelevant differences that replication experiments necessarily have when compared with their targets, without the need to invoke a bloated typology to do so.

Given that validity is central to replication experiments and their evaluation, it is then argued, following Feest's (2020) take on first-order experimentation, that validity should be conceived as something that comes in degrees. Because replication experiments again are first and foremost experiments, this clears way for them to come in degrees. A replication here is graded in at least two senses: (1) it is *genuine* or not, and it is (2) a *successful* or it is not. On the account that is defended, both of these senses should be seen as graded ones.

This validity centered account of replication is then used to show that a number of authors, including Irvine (2020), Feest (2019), Zwaan (2013), and Errington et al. (2021), among others, are mistaken in holding that replications are primarily concerned with assessing the reliability and *not* the validity of an experiment. The argument is then that reliability and validity are either tested concurrently or not at all. This means that replication experiments are very much about testing validity in opposition to the dominant view.

In the final section, an approach to evaluating replications on the deflationary account is introduced. These are captured in the following three 'orienting questions': (1) 'With what degree of certainty is the phenomenon under investigation characterized and individuated?', (2) 'With what degree of certainty is the replication experiment itself characterized and individuated?', and (3) 'What aspect of the target is the replication experiment intended to validate?' This model is then used to evaluate a series of replication experiments that was performed by Amodio et al. (2021) on Eurasian Jays. The overall appeal of doing things in this way is that it preserves the practice of replication while simultaneously making the replication experiments themselves easier to interpret.

Having clarified some of the concerns regarding a potential replication crisis for comparative cognitive science in Part 1 of the dissertation, Part 2 recenters the first-order problems that the discipline faces and assesses the prospects for generating empirical and theoretical progress. This part of the dissertation begins by introducing what is termed 'the validity cycle' or the idea that progress is repeatedly impeded in the discipline because (1) weak theory is tested with (2) underspecified hypotheses, using (3) methods that necessarily result in data that is not clearly

interpretable. This account simultaneously highlights a number of experimental, theoretical, and sociological factors that are intertwined with one another. Importantly, it is meant to be understood as a useful model for thinking through the way in which obstacles have been generated and is in no way meant to be an exhaustive account of all of the problems facing the discipline.

With the 'validity cycle' on the table, a series of research programs that are 'established' to varying degrees will be shown to demonstrate symptoms of being caught in a 'validity cycle': (1) mindreading, (2) episodic memory, and (3) flexible planning. The first two will be given a more detailed analysis than the last, only because it has been the subject of far and away more philosophical and empirical work. Building on this, an instance of what will be labeled an 'emerging' research program will be discussed, which focuses on the (4) non-human animal concept of death research program.

At this point, a diagnosis is given of why the validity cycle arises. It is argued that this is primarily a result of weak theory, underspecified hypotheses, and underdetermined experiments. It is shown here that there are at least five senses in which a theory might be considered to be 'weak' and that proposals that hinge on the wholesale adoption of a model of progress from mature sciences are not obviously promising. It is then argued that the problem of hypothesis under-specification arises when hypotheses are not able to generate experiments that are able to yield data that contribute to their confirmation or disconfirmation. Finally, the problem of experimental underdetermination is analyzed. The problem is initially diagnosed in terms of severe testing and then subsequently made sense of within the context of the Duhem-Quine thesis. Here, it is argued that the underdetermination of theories and hypotheses have been collapsed into one another and that this has potentially confounded the diagnosis of the underlying problem. In turn, this has fed into the repeated emergence of the 'validity cycle.'

Given this, four ameliorative proposals for moving beyond the validity cycle are then analyzed and modified. The first involves taking a bottom-up approach to the study of animal minds. The second involves breaking the target capacity into dimensions. The third involves adopting formalization techniques in the discipline. And finally, the fourth involves shifting the way

experimentation is done towards attempting to identify signatures of cognitive capacities as opposed to evaluating the success or failure of individual experiments alone. Each of these strategies is broadly advocated, albeit with specific modifications.

First, the top-down and bottom-up approaches are disambiguated and four ways in which these can be interpreted are presented: (I) Anthro- v. Zoocentric (starting point) (II) Anthro- v. Zoocentric (goal) (III) Complexity v. Simplicity (IV) Unique v. Widespread (V) Whole- v. Subsystem. In doing so, the ways in which multiple approaches can be adopted simultaneously and in a complementary manner are demonstrated, and two case studies are used to illustrate the point. The first involves C. elegans and sensorimotor integration. The second is centered on path integration and bio-robotics. In each of these cases, which are taken to be examples of progressive research programs, the way in which multiple approaches have been adopted simultaneously as well as in a complementary manner is demonstrated. Conceived at a high level of abstraction, this produces a progress facilitating model for comparative cognitive science.

Second, the proposal that cognitive capacities should be broken down into dimensions is addressed. Two cases are given focus here. The first is the causal cognition research program as discussed by Starzak and Gray (2021). The second is the behavioral innovation research program as discussed by Brown (2022). In each of these cases, it is argued that if dimensional proposals are to be progress generating, they need to meet the following conditions: (1) The goals of dimensional analyses should be made explicit and contestable. (2) Assuming the appropriate goals, the lower-level dimensions that are intended to account for the higher-level capacity should be at least less mysterious than the higher-level capacity. (3) Assuming the appropriate goals, the capacity that is being broken down into dimensions should be retained and properly represented in the lower-level dimensions. (4) Assuming the appropriate goals, absent a defined relation between the dimensions, any possible combinations of values of the dimensions should be seen as capturing the higher-level capacity. (5) Assuming the appropriate goals, dimensional analyses of cognitive capacities need to behave like dimensional analyses in other scientific contexts

Third, the proposal that the implementation of formal and computational methods will alleviate some of the problems that the comparative cognitive science has faced is addressed. While the

formalization of theories in comparative cognition holds the potential to contribute to theory development that is much needed in particular research programs, the blanket application of this strategy to the entire science is not particularly promising. Instead, formalization needs to be performed with a sensitivity to contexts that are amenable to it, and by working in a back-and-forth iterative manner with verbal theories.

Finally, the proposal that how testing is done should be changed so as to focus on signature testing is analyzed. Here it is argued that signature testing can be a promising approach in certain research contexts. However, this can only be done by taking on the challenges that so-called success testing faces. Moreover, given that there is no agreed upon or systematic way to evaluate bodies of evidence, a unique outcome that evades debate should not be expected.

In the conclusion to this dissertation, the overarching arguments are recounted, a cautiously optimistic view of progress in the discipline is reasserted, and a way forward given the high degree of uncertainty that characterizes much of the discipline is alluded to.

In the most general sense, this dissertation is intended to be a contribution to the philosophy of science in general, the philosophy of comparative cognitive science in particular, and to the material practice of comparative cognitive science. In this way, the overarching goal is to make a philosophical contribution that would improve the quality of that science.

1

A DEFLATIONARY ACCOUNT OF REPLICATION

0. What is a Replication?

Over the past decade, as a wide range of replications have systematically failed, the general status of knowledge that is grounded in a variety of scientific disciplines has been increasingly placed into doubt.[2] This has given rise to a multitude of pressing and difficult questions. Are these systematic replication failures indicative of fraud, bias, misaligned incentive structures, or negligence that is widespread in the sciences (Munafò et al. 2017)? Do p-values and other formal features of scientific practice need to be reformed in response to these replication failures (Ioannidis 2018, Benjamin et al. 2018, Wasserstein et al. 2016)? Is there a genuine replication crisis that spans across the sciences, or is it merely apparent or even local to certain disciplines (Fanelli 2017, 2018; Firestein 2015; Redish et al. 2018; Maxwell et al. 2015)? Is replicability a necessary feature of robust scientific practice (Leonelli 2018)? If it is not, which types of research should be subject to the replicability criterion, and which should be exempt (Guttinger 2020)? Are replications even an important part of scientific practice (Buttliere and Wicherts 2018; Feest

[2] Systematic replication failures can be found in molecular biology (Ioannidis et al 2012), experimental psychology (Open Science Collaboration 2015; Baker 2016) clinical medicine (Prinz et al 2011; Begley and Ellis 2012), health informatics (Coiera et al. 2018), Cancer Biology (Morrison, 2014; Errington et al., 2014), quantum computing (Frolov 2021) and artificial intelligence (Hutson 2018), among others.

2016)? These are deep and unsettling questions that urgently need to be addressed. However, they are all posterior to a more fundamental one: What is a replication?[3]

Answers to this question often begin with a qualifier such as, 'roughly' or 'vaguely' and a broadly agreed upon account of replication has been achieved neither in the philosophy of science nor in the scientific disciplines that have been affected by the replication crisis. What is clear is that this lack of a consensus poses an unsettling problem for scientists whose work has been challenged by the so-called replication crisis, as well as for those who are working in adjacent disciplines and research programs. Given that replication attempts are often and understandably perceived to be a threat to existing research rather than a complement to it that aids in advancing the reliability of scientific knowledge, there are looming concerns that the lack of an agreed upon account leaves replication as a practice broadly open to exploitation or manipulation. This issue is particularly salient if definitions of what a replication amounts to are able to be strategically adopted. Failure to control for this could result in biased assessments that allow the values and interests of certain parties to directly shape the interpretation of evidence in a way that would in other contexts be seen as epistemically impermissible. In this way, what counts as a replication is a matter of consequence for both quantitative as well and qualitative aspects of scientific investigation and is in no way a matter of definitional quibbling as has been claimed by some (e.g., Fletcher 2021).

This part of the dissertation presents and defends a deflationary account of replication that attempts to accurately capture both the practice of replication in its generality as well as the wide degree of variation that it contains. On the deflationary account, a replication is as an experiment that is a test of a hypothesis regarding the causal structure of a target experiment. Just as an experiment that is invalid is still an experiment, a replication experiment that is invalid is still a replication experiment. By adopting this approach, much of their mystery of what replications are dissolves and a host of first order tools of experimental design that can be used to evaluate them are opened up. It is deflation in so far as it renders the task of defining what a replication is trivial by focusing on the question of whether the experiment in question is valid and in the sense that it

[3] Replication, replicability, and replicator will be used throughout this dissertation, although these are generally interchangeable with other terms such as reproduction, reproducibility, and reproducer. A more explicit definition of all of these will follow although it will take the majority of this part of the dissertation to home in on precise accounts.

rejects the introduction of additional typologies or terminologies to explain the practice of replication.

Part 1 proceeds as follows:

In section (1), it is argued that the existing accounts of replication fall short. What is termed the naïve accounts of replication are addressed, as are the shortcomings of two recent accounts. An analysis of what is termed the 'causal' approaches to replication is then given with an eye to their various shortcomings before it is shown why typological accounts in general fail to latch onto the relevant aspect of replications. This section is intended to provide a number of guiding lessons that will inform the subsequent positive account. In section (2), this analysis is put to work to assess some recent explanations of the replication crisis. It is broadly argued that they overstep the evidence available to them and that there is either a clear replication crisis or a clear theory crisis, but that both cannot exist simultaneously. In section (3), a deflationary account of replication is given, and it is shown how this makes room for replications to come in degrees. In section (4), it is shown how the deflationary account reveals the extent to which validity and reliability are more closely intertwined than they have been thought to be in the literature and how uncertainty features strongly in this relationship. In section (5), a structure is presented for analyzing replications that are gradeable. It is illustrated through a larger analysis on a recent series of replication experiments by Amodio et al. (2021). Section (6) concludes this part of the dissertation.

1. Why Existing Accounts Fall Short

Given the numerous debates around what counts as a replication that have emerged over the past decade, several scientists and philosophers have attempted to provide a more unified and specified account of replication. In doing so, they aim to (1) adjudicate many of the debates that have arisen out of the so-called replication crisis, (2) maneuver concerns around definitional exploitability, and (3) provide general advice with regards to how to improve scientific knowledge in the most general sense. While it is argued that each of these accounts either explicitly fail to achieve the goals or have serious limitations, they nevertheless manage to provide guiding lessons for moving

forward with the project of giving a general and actionable account of what a replication is which will be introduced in Part 1 Section 3 below.

1.1 Naïve Accounts and Relevant Similarity

Modern discussions of replication and their role in the production of scientific knowledge can be traced back to Popper (1959/2002), who grounds his conception in Kant's (1786) emphasis on regularity as the basis of objectivity.[4] For Popper, a scientifically significant effect is only one that can be recreated by following, "the appropriate experiment in the way prescribed" (Popper 1959/2002, pg. 24).[5] Not only is replicability a way of determining what effects should be taken to be significant, but others, and perhaps controversially Popper himself, have taken replicability to serve as a necessary albeit not sufficient criterion for demarcating genuine- from pseudo-science (Derksen 2019).[6] On this schema, genuine science is replicable while pseudo-science is not. It is within this context that Popper famously argued that, "non-reproducible single occurrences are of no significance to science" (Popper 1959/2002, pg. 66).

Despite this, Popper also knew that there were practical limitations to replication, and that in principle, replicability is a potentially infinitely demanding criterion that is unable to yield so-called 'ultimate statements.' Given this, Popper held that scientific claims must be able to be tested, or replicated, at least in principle, even if it is not practical to do so. Of course, this leaves open several pressing questions about what should count as being capable of being tested in principle in addition to questions around what it means to follow "the appropriate experiment in the way prescribed" (Popper 1959/2002, pg. 66).

These are questions that have never left the philosophy of science and have come to be a part of what will be referred to here as the 'naïve' accounts of replication. It will be argued that these

[4] The intuition that replication is diagnostic of reliability is widespread and relatively old. Some have argued that versions of it can be traced back to the Royal Society in 1660 (Srivastava 2018, Ioannidis 2018)
[5] This take on replication cuts to the heart of Popper's claims about falsificationism as a response to the problem of induction. Many attempts at falsification can be spelled out in terms of replication, and for Popper replicability constitutes his testability requirement for scientific statements.
[6] While falsifiability is Popper's more well-known explicit demarcation criterion, replicability occupied this role, albeit in a more implicit way. For more on this see Braude (1979).

need to be overcome. Naïve accounts take the task of replicating experiments to be an obvious and straightforward one. That is, replication amounts to just doing the same experiment again and seeing if the same thing happens again. If it does, the effect is reliable. If it does not, it is not reliable. End of story. As will be made clear, this view of replication runs the risk of doing active harm to scientific inquiry.

Often, naïve accounts amount to an appeal to what is called direct replication. Direct replication is sometimes introduced in contrast to exact replication.[7] Exact replication entails an identical lossless duplication of its target, and it is impossible to achieve as has been repeatedly highlighted (Stroebe and Stack 2014, Lynch et al. 2015). Instead, direct replication, and all accounts of replication for that matter, require the introduction of a permissible type of variation between the replication and its target, while preserving only the integral features. This has been spelled out in several ways with varying degrees of detail. For example, Schimdt (2009, pg. 91) defines direct replication as the, "[…] repetition of an experimental procedure." This is introduced in contrast to conceptual replication which is the," […] repetition of a test of a hypothesis or a result of earlier research work with different methods." (ibid., pg. 91).[8] This set of distinctions that hinge on the notions of 'sameness' and 'difference', are resonant with many that are made in the various replication typologies.[9] This is a problem because what ends up counting as the 'same' or 'different' is unspecified by these definitions. This kicks the task of providing a guiding definition down the road, where it will need to be addressed by the scientists who are responsible for performing the replication experiments. In a certain sense, this is inevitable in so far as context specific features will always need to be considered. However, if the point of giving an account of replication is to establish a procedure that can be relied upon to prevent the encroachment of bias into the research, these accounts have failed in so far as they leave the practice wide-open to intentional or unintentional bias.[10]

---

[7] Exact replication as a term is sometimes invoked in a way that treats it as equivalent to direct replication. For example, see Doyen et al.'s (2012) discussion of Baragh et al.'s (1996) work on unconscious determinates and replications thereof. However, what they mean by exact replication is unspecified, and appears to just amount to a claim about the appropriate amount of variation between the replication experiment and its target.

[8] The focus here is on Schmidt's (2009) account because it has been particularly influential, but other attempts to define direct replication, and conceptual replication, face a similar problem. See Huffmeier et al. (2016), (Simons 2014), and Cesario (2014), among many others for examples of this.

[9] This will be further discussed later in this part of the dissertation (See Part 1 Section 1.5).

[10] This is not to say that certain parts of physics or chemistry have not arrived at productive means to determine what counts as a replication in these contexts. However, what makes replication in these contexts useful is not an appeal to

In various ways, the apparent difficulty of this definitional task has led recent authors to attempt to bypass the question of what a replication is all together. In doing so, it is assumed that either that the question has been adequately answered, that no answer must be given in order for the debate to proceed, or that answer to the question of what a replication is cannot possibly be obtained. This retains the naïve view of replication and its accompanying problems.

One example of this can be found in a recent article by Guttinger (2020), who argues that despite the apparent differences between accounts of direct replication, there is an underlying consensus that centers on two kinds of sameness that bind a replication to its target: (1) the sameness of experimental protocol and (2) the sameness of the direction of effect. According to Guttinger, if an experiment can be replicated along these criteria, then its results are generally considered to be reliable and trustworthy. However, this 'common denominator' style appeal to direct replication does not bypass the problems that are faced by naïve conception of replication broadly. A blanket appeal to (1) the sameness of experimental protocol has proven to be repeatedly obfuscating.[11] Moreover, it is hard to imagine any account of direct replication holding that merely (2) the sameness in the direction of effect is sufficient to count as a successful replication. Broadly, effects are thought of in statistical terms. A significance threshold or imprecise range seems to be necessary, even if this threshold cannot be precisely stated or agreed upon. For example, take a rare case in which the hypothesis being tested is bound to the direction of the effect; such as one testing for vaccine efficacy. Establishing that the rate of infection is indeed going down with the administration of the vaccine requires an effect that indicates a statistically significant reduction in infection. Merely establishing a miniscule sameness of direction in effect is obviously inadequate.

In addition to this, genuine and significant dispute remains if the effect or observation is multi-dimensional. Such multi-dimensional effects can be found in medicine in the varieties of syndromes that are identified via clusters of symptoms, in addition to research programs on well-being measures in normative economics, among many others. In these types of cases in which

bare similarity. It is rather, the battery of background knowledge of the mechanisms being studied as well as the state of the methods in these disciplines that is responsible for the productivity.

[11] For multiple examples of how complicated establishing what counts as the same experimental protocol can become see Klein et al. (2022).

effects can take on multiple 'directions' at once, the criterion seems to break down as an effective guiding principle.[12] To take another example, a medical intervention X for shoulder-hand syndrome might claim to be able to reduce shoulder and hand pain, reduce shoulder stiffness, but not reduce muscle atrophy. This syndrome is composed of an effect that is in no way unidimensional. If medical intervention X was repeated and was able to reduce shoulder stiffness, reduce muscle atrophy, but not reduce shoulder and hand pain, it might not be clear if this would count as 'the sameness in the direction of effect'.[13]

Another recent example of an implicit endorsement of the naïve view of replication can be found in Fletcher (2021, pg. 57) who writes, "How can one make precise this vague notion of a candidate for direct replication?", other than to appeal to the fact that a replication is "[…] sufficiently similar in the relevant ways […]," to its target. Fletcher then goes on to argue that the project of defining what a replication is can be isolated from other questions regarding replication such as how to measure it, and that these questions are in some ways more pressing. However, this amounts essentially to a refusal to attempt to answer the question when faced with its difficulty. There are multiple reasons for thinking that the notion of 'relevant similarity' will remain inert insofar as it is deeply unconstrained.

Similarity is a term with a difficult history in the philosophy of science, and it is in no way obvious that a blanket approach to similarity will be useful in this context.[14] Despite abundant appeals in philosophy and science to similarity, Goodman (1972) famously argued that similarity is a seemingly useless concept in both philosophy and science in so far as any two things are similar to one another in some way. This difficulty is at least in part because the salient notion of similarity is highly context dependent and is underdetermined by shared properties alone (Lewis 1973). Accounts of relevance do not do much better and run into the same types of difficulties that similarity has repeatedly faced (Cartwright 2009). The point here is not to claim that these

[12] This criterion for evaluating replication has also been discussed and implicitly endorsed by Errington et al. (2021).
[13] Here, Guttinger might argue the result was replicated in some respects but not others, with each effect being individuated under a different 'respect'. This objection would be curious in so far as it denies the way in which syndromes are holistically evaluated in medicine in favor of a decomposed evaluation. What's important in these contexts is not necessarily how the individual effects behave, but how they behave as a cluster, which facilitates diagnosis and progressive treatment.
[14] For further critiques of similarity in the context of assessing external validity, see Cartwright and Hardie (2012).

notions of similarity or relevance are explicitly driving these authors' conceptions of replication. Rather, the intention is to demonstrate how the naïve conceptions that lean heavily on 'relevant similarity' are simply too unconstrained to deal with the complicated debates around what counts as a replication that have emerge in the wake of the replication crisis. That is, absent further specification, there is no reason to think that a bare similarity approach that appeals to something like relevance will do much work in providing a productive account of replication. As Norton (2015, pg. 232) pointedly states the problem, "Just when is a second experiment replicating an earlier experiment as opposed to being a different experiment that looks similar to it?" A way of thinking about how to approach answering this question in a general way is needed, even if at the end of the day this is something that can only be addressed by taking individual instances under close consideration. Put bluntly, appeals to relevant similarity alone simply will not cut it.

Fletcher's (2021) difficulties, however, do not end there. He argues that a bottom-up approach to replication in psychology, that is, one that proceeds via the goals of individual disciplines, is more promising than a top-down one, that is, one that aims to give general principles of replication that cut across scientific disciplines. The core problem with this approach is that this a false choice. Top-down and bottom-up approaches, in the sense described by Fletcher, can and should be simultaneously adopted. That is, a complementary approach is required. Fletcher's opposition to top-down approaches in rooted in Zwaan et al.'s (2018) invocation of Lakatos' (1970) MSRP (Methodology of Scientific Research Programmes) and the apparent problems with both Lakatos' program itself and Zwaan et al.'s invocation of it. However, top-down approaches to replication, which aim to give general principles of replication that cut across scientific disciplines, need not be limited to the particularities and limitations of Lakatos. Instead, they can be rooted in the idea that replication is a general practice of evaluating the trustworthiness of scientific knowledge, even if that is yet to be made precise, and that any local account of replication that departs starkly from this should be approached with extreme suspicion. Moreover, a purely bottom-up approach, which proceeds via the goals of individual disciplines, assuming it is even possible to coherently adopt such an approach, leaves replication as a disjointed practice without any overarching unity. Paradoxically this would make it all the more difficult to assess replications within a highly localized context because there would be nothing more general to guide these modifications.

Against this, the deflationary account that will be defended later in this part of the dissertation (Part 1 Section 3) is complementary in character in so far as it adopts both top-down and bottom-up approaches simultaneously and allows them to place reasonable constraints on one another. Within the context of this account, this means taking seriously general principles about the function of science and allowing them to encounter and constrain the context-specific demands of scientific practice. By allowing these approaches to synergistically interact, a measured and general account of replication can be had. Before turning to and motivating the deflationary account of replication, several other existing approaches to replication will first be discussed and lessons will be extracted from them.

1.2 The Shortcomings of Two Recent Alternative Accounts of Replication

Considering the apparent insufficiency of the naïve understandings of replication, on top of the fact that an agreed upon answer to the question of what a replication is has clearly not been achieved, recently there have been several attempts to give a well-defined and unified account of replication. Instead of remaining within the bounds of the direct or conceptual replication which have dominated the debates over the past decade in the literature around the replication crisis, or making unqualified appeals to 'naïve replication', these accounts attempt to provide novel unified answers to both the question of what a replication is, and what it should be. While they are a step in the right direction in so far as they demand a more specified account, each fails to provide a sufficient account, albiet in its own way.

1.2.1 The Resampling Account

In this section, a recent account of replication put forward by Machery (2020) is analyzed. It is shown that one of the core problems that it encounters is a result of conceiving of replications and their targets in 'components' that are thought to be able to be randomly sampled from and recombined, and that doing so will yield a valid replication experiment. The analysis done in this section on Machery's account will motivate a skepticism regarding the adoption of a proceduralism regarding replications and demonstrate why replications should be evaluated holistically.

The driving intuition behind Machery's (2020) 'resampling account' is that the sampling practices that are typically reserved for the unit or the population should be extended to every aspect of an experiment. On the resampling account, experiments are broken into the following components: treatment, measurement, setting, and experimental units.[15] These components can be either fixed or random, depending on the particular replication. If the component is random, it has a set of more than one member that can be randomly sampled from. If it is fixed the set has one member. This is intended to capture both the various ways in which every experiment is fundamentally unique, while also allowing an identity relation between experiments to exist at a higher level of description. These components, as well as the identity relation that can exist between sets of them, provide the foundation for what counts as a replication on the resampling account. When an experiment is replicated, at least one component that is not fixed must be randomly 'resampled' from for it to count as a replication.

For example, a social psychology experiment that is intended to study 18–25-year-old BIPOC women in Southeast London should be able to be performed on any member of that defined population (i.e. 18–25-year-old BIPOC women). For the purposes of the experiment, there should be nothing special or outstanding about the particular individuals that took part in the target experiment. That is, they should be representative sample of the defined population, In a typical replication of this target experiment, a population of 18–25-year-old BIPOC women in Southeast London (the experimental units) would be sampled from. This much is orthodox. But on the resampling account the exact location and time (the setting), the experimental intervention (the treatment), and the ways of encoding behavior (the measurement), are all sampled from again in exactly the sample way as the population (the experimental units) is in the more orthodox cases. This is the procedure by which replication experiments are arrived at on the resampling account. It succeeds only in so far as it is able to produce a genuinely diagnostic replication experiment. That is, if it in fact tells us what a replication is.

[15] This is similar to Asendorpf et al.'s (2013) account which is rooted in a Brunswickian account of representative design. It also has affinities with Nosek and Errington's (2020) division of experiment into units, treatments, outcomes, and settings, which is common in many psychology studies. It is reasonable to assume that this critique of Machery will apply equally well to Brunswick's original account in so far as it breaks experiments into re-combinable components.

This is illustrated somewhat more abstractly in Figure 1 below. First, the target experiment consists of a sampling of the experimental components. Then, in the same way, the replication experiment samples again from each of the components in order to establish the replication.

Treatments	**T1**	T2	T3	T4
Measurements	M1	**M2**	M3	M4
Experimental units	**U1**	U2	U3	U4
Settings	S1	S2	**S3**	S4

The target experiment (sampled)

Treatments	T1	**T2**	T3	T4
Measurements	**M1**	M2	M3	M4
Experimental units	U1	U2	**U3**	U4
Settings	S1	S2	S3	**S4**

The replication experiment (resampled)

Figure 1. An illustration of the resampling account. The green text in both tables indicates the (re)sampled components. In the target experiment each of the components are sampled from to form the following composition of components (T1, M2, U1, S3). In the replication experiment each of the components are sampled from again to form the following distinct composition of components (T2, M1, U3, S4).

The resampling account is intended to be an improvement on the naïve accounts, and their variations, which were discussed previously in Part 1 Section 1 of the dissertation. It does this in numerous respects. It moves beyond appeals to exact replication. It also breaks out of the seemingly stagnant debates regarding whether conceptual replications should be considered replications at all, or if only direct replications should be considered to be legitimate and/or diagnostic. In addition to this, it sets out a straightforward procedure for establishing what a replication is in a variety of contexts that is seemingly amenable to analysis and clear adjudication. These are features that were notably lacking from the previously discussed accounts.

In addition to this, one of the core virtues of the resampling account can be found in the central and expansive role it gives to random sampling. Random sampling is thought to contribute to the evasion of concerns about bias or cherry picking because, to a certain degree, it places the control out of the hands of the scientist conducting experiment (Johnstone 1989). If a sample is able to be selectively chosen by an experimenter, the results are able to be manipulated according to the values of that experimenter. In so far as this type of prophylactic against intentional or implicit bias can be generalized to every aspect of the experiment; this is at least intuitively an appealing an ameliorative measure. In doing so, the resampling account also implicitly highlights an important link between the structure of experimentation and replication. Despite these virtues, it also encounters several significant problems that are seemingly fatal to the account.

One initial problem that might stick out is that what counts as a fixed or random component is inevitably subject to framing. For example, in a medical randomized controlled trial, a drug could either be fixed (i.e., only one drug is being tested and not another candidate drug) or it could be random in so far as the individual pill is sampled from a population of pills in the bottle. In this way, the very structure of the components is potentially open to an impermissible encroachment of values. If the constitution of a component can be manipulated on the part of the individual(s) performing either the first-order experiment or the replication experiment in accordance with their desired outcome (i.e., their values regarding the outcome of the experiment), there is once again a concerning sense in which the resampling account is open to bias. This is an example of what Douglas (2009) would consider an impermissible direct role for values to play in the scientific process. Such an encroachment would then be impermissible in so far as it impedes the scientific process by allowing bias to shape how experiments are composed.

Given that one of the central goals of the account of replication is to reduce bias, this is a concerning feature of the account. It would perhaps be more useful to say that *all* components should be conceived of as being random, with the differences between them being what type of random sampling is permitted. Doing so avoids the way in which the account leans implicitly on a notion of relevant similarity to determine which components are fixed.[17] On this way of reframing Machery's account, a new and distinct experiment would draw from outside of the set

[17] This was previously critiqued in Part 1 Section 1 of the dissertation.

that defines a given component, while a 'genuine' replication would draw samples from within that set. This would require explicitly negotiating the contents of the set rather than relying on a proceduralism that is built into the account to obtain it. However, even if this solution is a plausible one, problems for the account do not end here.

The more substantial problems have to do with the central role that random sampling plays in the account. This comes in two varieties: (1) non-ideal and (2) ideal.

First, the non-ideal version of the account (1) will be addressed, and an argument will be made regarding the troubled role that random sampling plays in the account. Machery's discussion of the difference between fixed and random components is complicated by the fact that both are framed as 'idealizations', meaning they contain *modifications* to features of experimental inquiry rather than *omissions*.[18] Despite this, he acknowledges the necessity of convenience sampling even when sampling from a population (experimental units), while simultaneously making random sampling fundamental to his account.[19] This means that the contents of components are often chosen simply because they are available. However, even if the ideal of random sampling cannot be achieved in every instance of sampling from a population of subjects, it should still be meaningfully sought after.

However, what might it practically mean to randomly and independently sample from the spaces of treatments, measurements, and settings, especially given that random sampling appears to be insurmountably challenging when it comes to the experimental units or the population? Practically enacting this proposal would require radically altering the face of scientific practice. In many cases, the process of doing so would necessarily face serious obstacles. For example, settings

[18] In the philosophy of science, these are typically contrasted with abstractions. Abstractions here are conceived as a stripping away of features rather than a modification to them. Given this, it is also plausible that Machery's account is engaged in abstraction as well. For more on this distinction and on idealization in particular see Frigg (2020).

[19] Convenience sampling is a problem for a number of reasons. Among others, it potentially leads to potentially non-representative data. For example, Stewart et al. (2015) show that the Amazon Mechanical Turk (MTurk) population, which is widely used for psychology experiments, is made up of ~7300 workers. This, in addition to the well-known problem of psychology undergraduate students not being representative of a larger population as they are sometimes taken to be presents serious issues that disciplines like psychology have only begun to address. For more on this see Heinrich et al.'s (2010) discussion of WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations. Given resource constraints, it is not clear in what sense the population of subjects can better approach being randomly sampled.

can be very difficult to alter even in a non-random sense. This remains the case when the hypothesis being tested for is not bound to a singular setting. Genuinely sampling from a setting space would demand a type of mobility on the part of the treatement, the experimental units, and the measurement that would be astronomically resource intensive. This makes the motivations for adopting an account that builds random sampling into every component of an experiment somewhat unclear given that it cannot possibly be implemented in the way that is being suggested. Moreover, it is not as if the random sampling of every component can be plausibly framed as an ideal that scientists should asymptotically strive towards, in a way that it might be with the experimental units. In this way, it remains entirely unclear what role random sampling could in fact play in the account when it is conceived under non-ideal conditions. The move seems to be, treat this as if it was a random sample, but we all know that random sampling cannot happen. The natural question to ask then is then the following one: what does it mean to randomly sample any component? Machery does not have an answer.

With the critique of the non-ideal version of the account on the table, the ideal version (2) can now be addressed. While the critique of the non-ideal version of the account alone should provide sufficient reason to doubt that the resampling account is a productive account of replication, the more pressing problems arise when this idealization is taken seriously. Two primary critiques will be given of the non-ideal version of the resampling account. The second will be given more substantial treatment than the first.

First, merely resampling an experimental component does not always seem sufficient to generate a replication; at least in the way that they are typically conceived. For example, Inoue and Matsuzawa (2007) compared their chimpanzee Ayumu's (n=1) working memory capability for numerical recollection to that of an adult human (n=12). Both Ayuma and the humans were shown a screen that simultaneously and randomly flashed a series of digits for a couple of hundred milliseconds before they were masked. The participants were then tasked with taping the masked digits in sequential order. They found that Ayuma was as quick and accurate as the average human, which gave way to strong and controversial claims about the working memory of chimpanzees at the species level that were ripe for replication attempts. However, if Ayuma was tested again the next day, that is the setting was randomly resampled from, that would count as a

replication on the resampling account, and one would presumably be justified in claiming that Inoue and Matsuzawa's (2007) experiment in fact replicated.

This is highly unorthodox seeing as the prevailing norms of the discipline in question typically require a sample size of greater than one when the sample space of the experimental units is in fact vast, for the replication experiment to in fact diagnose the target experiment's reliability. One reason for thinking this can be found in the nature of Inoue and Matsuzawa's (2007) claim and the sample space of subjects that is stipulated in their paper. This experiment was not a case study on Ayuma but rather one in which Ayuma was intended as a representative sample of the species chimpanzees (pan troglodytes). Inoue and Matsuzawa (2007; pg. 2) write, "Our study shows that young chimpanzees have an extraordinary working memory capability for numerical recollection better than that of human adults." The norms of comparative cognition, and most other scientific disciplines for that matter, typically hold that replications require genuinely representative samples. This is not to say that radical changes to scientific practice cannot be made. However, such radical changes require substantial justification.

Second, a core novel feature of the resampling account of replication lies in the way in which it breaks experiments into independent components that are able to be randomly sampled from. This entails that any randomly selected combination of the experimental components, assuming they are not fixed, can output an experiment that will count as a diagnostic replication of the target. However, in the following it will be argued that the independence of these components, as well as their non-hierarchical nature, is a source of problems. This is an issue that will arise in any context in which there are dependencies between the experimental components. It occurs frequently enough to pose substantial challenges to the resampling account.

For example, Lo et al. (2019) used multi-modal microscopy to achieve a structural and chemical mapping of an Allende meteorite particle. Part of their analysis made use of transmission electron microscopy, which was performed with a Titan 60-300 equipped with a HAADF detector. Of course, there is nothing special about this particular piece of equipment. They could have just as easily used any other functionally equivalent TEM (transmission electron microscope) that was manufactured by another company. On Machery's account, the collection of all the functionally

equivalent TEMs would plausibly constitute the measurement sample space. However, the operation of any given TEM is something that must be learned. While a given technician who is a member of the 'situation' sample space might be able to operate the Titan 60-300, they might not be able to operate another TEM such as the FEI Tecnai F20, which a technician at another lab is able to operate. This means that if the 'setting' space and the 'measurement' space are randomly sampled from there is the potential to land in a scenario in which a replication will fail, or not be possible to carry out, for merely practical reasons that have nothing to do with the trustworthiness of the data or knowledge it produces. In certain scenarios, randomly sampling experimental components would clearly be a misleading strategy (see Figure 2 below).

Treatments	T1	**T2**	T3	T4
Measurements	**M1**	M2	M3	M4
Experimental units	U1	U2	**U3**	U4
Settings	**S1**	S2	S3	S4

The replication experiment (resampled)

Figure 2. An illustration of the resampling account applied to the Lo et al. (2019) example discussed above. The colored text indicates the randomly resampled components. The red text above is intended to indicate a situation in which the measurement and the setting spaces were randomly sampled from and were not able to be recombined to yield a diagnostic replication. In the example discussed above M1= Titan 60-300 and S1= a technician that would not be able to operate it. In this imagined case there are 256 or 4^4 possible samplings of the components. 4^2 of these however, or 16/256 would yield replication experiments that would not be able to evaluate the target.

Another example can be found in a hypothetical case in which populations and settings are entwined in a tightly connected dependency. If an experimental population contains a mixture of English speakers and Spanish speakers, the vignettes given to them must be in a language they understand, so the English language vignettes must be given to the English speakers and Spanish language vignettes to the Spanish speakers. Independently resampling both the population and the language of the vignettes can again result in failed replications for reasons that have nothing to do with the actual trustworthiness of the research, which on Machery's account is what replication is intended to assess. This means that in this case the resampling account would output an experiment that would almost uncontraversially not be considered to be a replication. Once again,

if the resampling account is an answer to the question "what is a replication?", that answer seems to have come up short.

In sum, while there are experiments in which the components are able to come apart in a way that would allow them to be randomly sampled so as to yield a diagnostic replication, there are also many experiments that do not allow for this. Given that the random resampling of the non-fixed components is intended to provide a method for arriving at diagnostic replication experiments in a way that at least partially evades bias, these types of results pose a serious challenge to the account.

One plausible solution to this problem that would retain the basic features of the resampling account would entail taking on a domain restriction to not allow for such cases to arise. This would in essence exclude members from each of the experimental components that were not able to be freely combined with the others. However, doing this would exclude cases that would seem to count as genuine replications. Building on the Allende meteorite particle example, a restriction of the measurement space to only include equipment that *all* the technicians in the setting space can operate would exclude potential sources of replications, as would only including technicians in the setting space that are able to operate all of the relevant equipment. Moreover, such an *ad hoc* restriction comes at a high cost and betrays core commitments and appeals of the account that are intended to exploit the random sampling of the experimental components to yield a less biased diagnostic replication experiment. This solution both removes the randomization of components, which is basic to the account, and it excludes potential sources of replication in a way that is unprincipled.

At this point, the resampling account is in a bind. Either it can exclude cases that it should include, include cases that it should not include, or potentially both. Conflicts like this one are particularly likely to arise given that there are no restrictions placed on establishing membership in a component, or guidance for how this should be done, which by itself is a source of difficulties. However, given the scope of the account, this might not be surprising.

Machery (personal communication) has proposed a solution on which components of an experiment that are not able to be freely recombined need to be treated as a combined population. On this version of the account, if a treatment and a measurement are dependent on one another they would be combined to form a singular component that would in turn be able to be sampled from. Machery (personal communication) highlights here the way in which the definition of component membership demarcates the space within which the experiment is theorized as being able to be generalized.

This constitutes a significant alteration to the account. The components of a given experiment would change according to the way in which they are dependent on one another, and in some cases, they would need to be collapsed into one another all together. This solution presents additional problems, and again partially removes the random sampling aspect of the account in a way that might open the door to bias.

Return to the THF microscope-scientist dependency case discussed above. One thing that might be done is to create a component that binds the scientist X to the microscope they are trained on. This seems straightforward enough. But assume that there are 10 microscopes, and that scientist X is trained on 5 of them. In this case a multi-step sampling procedure would need to be implemented. First, the space of [scientist-microscope] components would be sampled from. This would allow for the base of components to be established that would then be sampled from in a second step. Moreover, this would create path dependencies and would require treating these component as fixed in a way that is not principled on the resampling account's own lights. Absent this two-step procedure, microscope-scientist pairings would be randomly established, and part of the previous problem would remerge: cases that should be included would be excluded for the purposes of making sure that the account does not output results that would be unconstraversially considered to not be replication experiments.

Given this, one might think that the account could be more simply restated so that there is a set of experiments that could be carried out that constitute the space of potential replications, and that this singular set that considers experiments holistically can in some non-ideal sense be randomly sampled from. However, framing the account in this way only brings out the unclear

role that randomly sampling plays in the selection of individual components. Moreover, it amounts to abandoning the core novel and seemingly appealing feature of the account; breaking experiments into components and then random sampling each of them independently so as to evade bias. Given this, and the practical problems discussed previously on the non-ideal interpretation, either the random sampling aspect of the account should be subject to more detailed practice-based analysis, or it should be dispensed with all together.

Another concern that might be had with Machery's solution to the problems addressed on the ideal version of the account lies again in its high degree of flexibility and the potential bias that might arise from this. For example, the familiar concern that any failed replication might be able to be explained away in a *post hoc* manner as a non-replication because there is a dependency between the components that is responsible for the production of the effect that was previously not accounted for. What is needed then is a theory of component dependencies so that potentially biased justifications can be separated out from legitimate ones.[20] Absent such a theory, this relation is left implicit and intuitive, and given how the replication literature has played out over the past 15 years, this remains a serious challenge. In this respect at least, it is unclear how the account consitutes an improvement upon the ones that were discussed in the previous section.

Recovering the resampling account seems to require substantially altering it and abandoning many of its core features. This can only be done in so far as experiments are conceived of holistically and not as a randomly sampled collection of components that can be recombined with one another. That is, there are certain combinations of components that should be permitted and certain combinations that should be prohibited. But this amounts to saying that some experiments count as replications in a particular context and others do not, when considered holistically.

In light of these issues and given that the resampling account is normative in nature, adopting it will seemingly not bring clarity to the current practice of replication, or necessarily lead to the more reliable trustworthiness assessments of the sciences that have been affected by the replication

[20] This might seem to resonate with Collins' (1994) experimenters regress. In cutting-edge science, whether the instruments being used are themselves novel and non-standardized, there is disagreement about both the facts and the reliability of the instruments, leading to seemingly intractable disputes. Collins famously uses the example of gravity waves.

crisis. These are reasonable demands to place on an account of replication, particularly of one that claims to be able to improve upon the naïve accounts previously presented and provide an answer to the question "what is a replication?" Instead, the resampling account will potentially introduce even more confusion and uncertainty.

1.2.2 The Diagnostic Account

Replications are commonly thought to perform a diagnostic function regarding the experiment that they target. However, performing replications is not the only way that this can be done and there is a tendency on the part of many accounts of replication to overburden the concept. By framing replication in a way so that it includes potentially every possible way of assessing the trustworthiness of scientific knowledge. By requiring the concept to include such a wide variety of methods, both the project of describing the practice of replication and the normative project of enhancing the practice of replication are convoluted. In this section, a recent account of replication that fails precisely because it adopts an overblown understanding of the role that replication experiments should play is analyzed.

On Nosek and Errington's (2020) diagnostic account, one experiment is a replication of another *iff* it can both increase *and* decrease one's credence in evidence taken from a prior experiment.[21] It is essential on the diagnostic account that the replication is able to both increase and decrease one's confidence in a prior hypothesis. If it cannot do this, then it is a generalizability test; or a test of the scope of the hypothesis that is being tested in the study.[22] This means that every replication is a generalizability test because even an attempt to exactly reproduce an experiment will inevitably introduce some novel feature that was not present in the original. These generalizability tests, however, need not be replications.

[21] Nosek and Errington frame their account in terms of 'claims'. It is unclear whether these should be understood as conjectures, hypotheses, or something else entirely. For this reason, in reconstructing their account, I have used hypotheses in place of 'claims' for the sake of clarity.

[22] The diagnostic account remains silent on cases in which an experiment is only able to reduce our confidence in the results from a target experiment. This might arise in cases that have been so extensively confirmed by a given experiment that one's credence could not possible be brought any higher by that experiment. Rather, in this case it could only possibly decrease. Depending on one's credence function, the faster-than-light light neutrino experiments could be interpreted as a plausible instance of this (Adam et al. 2012).

The diagnostic account is intended to capture the evaluative role that replication plays in research across the sciences, which makes it particularly appealing in its simplicity and scope. It provides a fully general account of replication by attempting to plainly state the criteria for determining when one experiment is a replication of another. Moreover, it relegates replications to a theoretical rather than a procedural role. This aspect of the diagnostic account is appealing. Despite these virtues, the diagnostic account, like the resampling account, ultimately generates more confusion than clarity. More specifically, it falters in asking too much of replications.

The account initially goes awry in so far as it captures cases that are far outside of what would normally count as replications. In doing so, it warps the practice of replication beyond recognition. For example, an experiment done by Weaver et al. (2004) found an alteration in the epigenome of rat pups whose mothers licked, groomed, and nursed them at a glucocorticoid receptor gene promoter in the hippocampus. This is a test of a hypothesis that is steeped in background theory and auxiliary hypotheses from a variety of disciplines.[23] Among many others, these include hypotheses about the function of the hippocampus and its larger role within the brain, hypotheses about the role of glucocorticoid and other hormones in the body, and hypotheses concerning the functioning of methods like sodium bisulfite mapping. Any experiment that can potentially both raise and lower one's credence in any one of these theories or hypotheses that are not foregrounded, will also be able to potentially raise and lower our credence in Weaver et al.'s target hypothesis. This means that experiments that are external to environmental and behavioral epigenetics can potentially count as replications. For example, an experiment testing the validity of the method of sodium bisulfite mapping, or bisulfite sequencing, can count as replication of experiments in which this method was used.

This raises the issue of what types of evidence should count as being relevant in the diagnostic account, and whether the account can in fact answer this. There are two distinct concerns here. If it is extremely responsive, then potentially any experiment might count as a replication of another. If it is generally non-responsive, almost any experiment might not be counted. These are concerns that are somewhat analogous to the paradox of the ravens that Hempel encountered in the formation of his theory of confirmation (Hempel 1945), on which any an observation of any

[23] For more on the role of background theory and auxiliary hypotheses in experimentation see Duhem (1954).

red pencil or a white shoe, came out confirming the hypothesis that all ravens are black. While many solutions have been proposed to Hempel's paradox (e.g., see Good 1960), Nosek and Errington's account does not set a principled threshold. Moreover, varying assessments of one experiment's ability to bear on another often give rise to seemingly important disputes about what should count as a genuine case of replication, what is a generalization, and what is neither of the two.

In addition to this, an increasingly common desideratum for an account of replication, as seen in the discussion of the two recent novel accounts, is that it can distinguish itself in a principled way from generalizations; the application of a hypothesis to a population that is distinct from that of the target experiment.[24] This desideratum emerges within the context of the replication crisis in general, and within the debate between direct and conceptual replication in particular. Many skeptics of the use of conceptual replication have claimed that rather than being directly diagnostic of the target experiment, conceptual replications are instead tests of generalization. This distinction between replications and generalizations is intended to dispense with the direct v. conceptual distinction and bring ease to the task of sorting genuine instances of replication from other research practices that it might easily be confused with. Doing so is intended to make clear what the scope of the claim of the target experiment in fact is, and to make evaluating the implications of replications and their failures into a more straightforward task. The appeal of this concept is seen most clearly in replication failures. Whereas a failed replication is generally thought to cause the trustworthiness of its target to be placed into doubt, a failed generalization is generally thought to do no such thing (see Table 1. below).

Type of experiment	**Effect of success or failure on target**
Replications	Places trustworthiness of the target into doubt.
Generalizations	Does not bear directly on the target.

Table 1. Two types of experiments that are commonly bound together in the replication literature.

In distinguishing these two types of experiment, it is important to recognize that they are in fact testing two distinct hypotheses that will require distinct control conditions to be tested in a valid

[24] Both Machery (2020) and well as Nosek and Errington (2020), among others, set this as a desideratum.

way. This means that there is a claim implicit in generalizations that requires further explication that goes over and beyond the original experiment. Moreover, the deep difficulties that come with the concept of generalizations can be seen most clearly in cases of success rather than failure. Assume that both the original experiment and the experiment that tested a 'generalized' population were successfull. A generalized population might contain the original population, but it might also not. How exactly should the evidence from both of these experiment bear on one another? It is not at all clear in any sort of general sense, and it is even less clear how tying this underspecified concept in a deep way to replications is going to be helpful.

Given this, there is no particularly compelling reason to bind an account of replication to an account of generalizations, other than to demarcate genuine from non-genuine instances and types of replication. In addition to the reasons provided, there are many ways for something to not be a replication of its target that are not exhausted by the concept of a generalization. Given this, a positive account of replications is seemingly sufficient, particularly seeing as there is no clear or agreed upon theory of generalizations to fall back on. While spelling out the precise purpose and function of generalizations is a worthwhile project that has been largely neglected, particularly in cases involving successful experiments there is no clear reason that a theory of replication needs to take on this task. Setting it as a desideratum for an account of replication requires further justification. Currently, it is unmotivated.

Furthermore, assuming a particular credence function, there are reasons to think that if an experiment is only able to confirm a hypothesis, and not disconfirm it, we should not think of it only as a generalization and not a replication as Nosek and Errington here claim. While replicating experiments that have only failed and are not supported by previous evidence are not common, there is no principled reason to think that they should not be carried out. Here, it does seem that in certain cases the obtainment of a positive result would only be able to increase a credence in the hypothesis, assuming that it would be low to start with, again, assuming a particular credence function. Given this, cases in which additional negative results are obtained should on this framework be seen as replications and not generalizations, contrary to Nosek and Errington. This in combination with concerns regarding the relationship between generalizations

and replications should be sufficient reason to cast doubt on the utility of the account and its ability to move productively beyond the naïve conceptions of replication.

In summary, the diagnostic account goes awry in so far as it overburdens the role of replication in scientific practice, where overburdening means that the functional role of the concept has become too broad. That is, it potentially includes everything from meta-analyses to experiments in adjacent research programs to experiments in other disciplines, By making it cover anything that is able to increase or decrease ones credence in a target experiment, it holds the potential of obfuscating both the descriptive project of capturing replication as it is currently practiced, and the normative project of improving the way in which replications are performed.[25] This provides support for the idea that replication experiments perform a particular type of assessment of the trustworthiness of their target and should not constitute the entirety of available trust-inducing practices.

1.2.3 Lessons from Two Recent Accounts

Accounts	**Appeals**	**Limitations**
Machery (2020) – The resampling account	• Evades the direct v. conceptual debate • Introduces random sampling	• Includes cases that should not count as replications • Excludes cases that should count as replication
Nosek and Errington (2020)- The diagnostic account	• Simple, straightforward, and maximal in scope	• Potentially anything counts (or does not count) as a replication • Overburdens the practice

Table 2. An analysis of two recent accounts of replication

In this section, two recent attempts to give an answer to the question of what a replication is were analyzed (See Table 2 above for a summary). Both accounts were shown to face series of significant limitations. However, the shortcomings of these accounts are instructive in so far as

[25] Nosek and Errington's project is clearly an ameliorative one, however, this project will necessarily lean on the descriptive project of capturing replication as it is currently practiced. If this is entirely abandoned there is good reason to think that the ameliorative project will result in confusion. For an example of this in a social context see Mikkola (2009).

they render some the challenges facing an account of replication clear. The importance of holistically evaluating replications was defended. It was also argued that when replications are overburdened the science suffers on two fronts: First, it potentially results in a convoluted and confused practice of replication and second it leads to the potential neglect of other trust inducing practices that are available. In the next section, more promising accounts that are unified in their focus on the causal structure of the effect in question will be focused on. These will serve as the foundation for the positive deflationary account that will follow later in this part or the dissertation.

1.3 Causal Accounts

While not an explicitly articulated approach to replication, in the following, a cluster of views will be unified here under the term, 'causal accounts.' These accounts of replication typically demand an extensive understanding of the target phenomenon in addition to an extensive understanding of the target experiment. Broadly, for a replication to have any sort of diagnostic property on these accounts, an understanding of the causal relationship that is being investigated in the experiment is required as is having an understanding regarding how the causal relationship that produces the effect of interest is sensitive to specific interventions, or something functionally equivalent to it. While not every hypothesis that is tested in an experiment can be characterized as being explicitly causal in nature, it will be assumed that empirical investigation generally is centered on the manipulation and control of these causal relations (Feest and Steinle 2016) and that this characterization of experiments is central to broadly understanding and evaluating replications.

1.3.1 The Material Analysis

The first example of a causal account comes from Norton (2015), who defends a material, as opposed to a formal, analysis of replication, and argues that determining whether the results of a replication are significant is not something that can be settled in the abstract. On this account, a formal solution to what makes a replication significant, will necessarily come up short. This is because there are relevant background conditions in each case that determine and characterize the

precise significance of a given replication, and how it should be characterized in relation to the target experiment. These conditions can be broken into two classes: the (A) experimental conditions and the (B) confounding conditions.

First, the (A) experimental conditions are constituted by a set of facts that specify when an effect or process X will produce what he calls a veridical experimental outcome VO, or an outcome that obtains the effect that the experimental protocol was designed to demonstrate.[26] In a footnote to this definition, Norton states that this is equivalent to 'construct validity', or the ability of a theoretical construct to explain the results of a given test. This is a concept that can be traced back to Cronbach and Meehl (1955), which is often contrasted with predictive validity, or the ability of the results to predict a criterion variable.[27] In essence then, the (A) experimental conditions are meant to demonstrate that X *can* cause V.[28]

By contrast, the (B) confounding conditions are constituted by a set of facts that specify when an effect or process X will produce what Norton calls a spurious experimental outcome SO, or an outcome that obtains the effect that the experimental protocol was not designed to demonstrate. This can follow one of two paths. First, X can be sufficiently disrupted so that S does not occur. Second, it can be the case that X is not present, which will prevent SO from occurring. In essence, the (B) confounding conditions are meant to demonstrate that something other than X does not cause SO.

Taken together, background facts from both (A) and (B) are taken to be sufficient to demonstrate that X causes Y. However, on Norton's account, successful replications are only able to test whether facts from class (B) are present. This takes two forms. First, by replicating an experiment

[26] While Norton never uses the word causation, and he is explicitly critical of the usage of the concept in science (see Norton 2007) there is reason to think that something like causation is implicit is his account of both (A) and (B) conditions. One reason one might want to avoid causal language can be found in the massive literature on the metaphysics of causation, and the looming pressure to specify an account of causation when making sense of replications. This requirement will be rejected in this part of the dissertation and a pragmatically pluralist account of causation will be adopted going forward.

[27] A more extensive discussion of the concept of validity is given later in this part of the dissertation (see Part 1 Section 3.1).

[28] Ascent to this aspect of the account can be found in Romero (2020), who has briefly asserted that a replication mirrors its target in all the ways that are responsible for producing the effect. However, absent facts from (B), this position has little to no traction.

multiple times and varying the factors in an experiment that permit variation, confounding factors that obscure the causal link between X and Y can be practically eliminated. This means that one off replications on this account are insufficient. Second, by replicating an experiment with no relevant change to these factors, random error can be excluded as being responsible for the spurious outcome SO.[29]

However, for a replication to license an evaluative inference regarding the target experiment, facts from both (A) and (B) must be supported. This is a problem because replications are only able to test facts from (B). In many ways, this distinction maps onto the common claim in the literature that replications can test reliability but not validity.[30] Recognizing the distinction between these two types of facts importantly holds the potential to bring clarity to disputes around replications.

For example, in his discussion of intercessionary prayer, successful replications lack evaluative import in so far as they are missing facts from class (A). This means that although Leibovici (2001) was able to 'replicate' earlier findings that showed intercessionary prayer to be effective, by using the standard methods implemented in the biomedical sciences such as RCTs (randomized control trials), the replication was not diagnostic because the target study lacked construct validity. Another way to frame this, and one that I will defend later in the paper, is that not only did the target experiment lack construct validity, but so did the replication.[31]

This case is contrasted with the successful replication of studies that showed that the H. Pylori bacterium was the cause of stomach ulcers even though it was long thought that bacteria could not survive in stomach acid. While the initial findings of Marshall and Warren (1983) struggled to get published, the study eventually afforded strong evaluative import, in so far as facts from both (A) and (B) were present. Facts from class (A) in this case included the demonstration of a 100% correlation between the presence of duodenal ulcers and H. Pylori bacterium and facts from class (B) included the multiple replications of in labs all over the world which were purportedly able to exclude confounding factors.

[29] Notice that despite the introduction of other constraints, the account does not evade an appeal to relevance.
[30] This assumption will be explicitly challenged later in this part of the dissertation (see Part 1 Section 4).
[31] This point was not made by Norton.

Norton's approach to replication also offers a way of interpreting failed replications. For example, in research on cold fusion, there were both successful and failed replications of the same experiment. However, these experiments were performed in non-ideal epistemic scenarios. As replication attempts continued it became increasingly clear that experimenters differed in their background assumptions regarding facts from class (A). Therefore, the types of inferences that could be drawn from the replications, as well as the interpretation of the replications themselves were understandably and predictably various. In this instance, making sense of the failed and successful replications required first settling the background assumptions regarding facts from class (A).

While this is not highlighted by Norton, this lack of consensus regarding background assumptions has direct implications for settling the necessary facts in class (B) as well. If there is no consensus regarding the construct validity of the target experiment, or facts in class (A), we cannot possibly settle what the confounding conditions might amount to in so far as they are dependent on a specific characterization of facts in class (A). This is a point that will be repeatedly returned to over the course of this part of the dissertation.

Norton's account is a powerful and potentially elucidating one when it comes to interpreting replications. However, it is also extremely demanding; and perhaps excessively so. This is because it requires a thorough understanding the causal structure of the effect of interest and its sensitivity to confounding conditions in order to get off the ground or have any evaluative import. Absent this, the replication will remain inert. While in a fully ideal scenario this account holds a strong appeal, in practice it presents at least two difficulties that are rooted in two closely connected idealizations. These are addressed in the following two sub-sections on idealized causality and idealized validity.

1.3.2 Idealized Causality Leads to Idealized Replications

Anjum and Mumford (2018) locate their critique of causally centered accounts of replication in the causal interpretations of experimentation more broadly. While they refrain from specifying

who is the exact target of their critique, they do introduce two target definitions: first, they hold that an experiment amounts to the introduction of a cause, under a particular condition, producing an effect. With this context, they, second, argue that a replication consists of introducing the same cause, under the same condition, which, in turn, produces the same effect.[32]

Anjum and Mumford (2018) contest this way of framing replication considering the four assumptions it makes about causation: (1) the same cause gives same effect, (2) causal necessitation, (3) total cause, and (4) that causation is a deterministic and closed system. In their account of causation, none of these assumptions hold because scientific inquiry often targets open systems, with non-linear processes, that contain hypersensitive elements that are generally characterized by high degrees of uncertainty. This means that on their account, how causation is conceived of in the sciences broadly construed must be reconsidered. This is the larger project of their book. As a part of rejecting the above four assumptions about causation, Anjum and Mumford go on to argue that a less idealistic conception of replication than the one that is presented above must be adopted; namely one that is less outcome or results centered.

Again, the general idea on causal accounts is to isolate the causal features of a study as the ones that must be preserved in the replication, while also allowing for variation in all the other features that do not introduce interference. After all, it is variation that allows for confounding conditions to emerge and for the precise causal structure of the effect to be investigated. This was discussed in more detail in the account put forward by Norton discussed above, on which facts from both class (A) and class (B) must be known for the replication experiment to be a diagnostic one.

Nevertheless, given that the epistemic stance that one often comes to when carrying out replications is characterized by a significant degree of uncertainty (otherwise why would anyone bother to perform the replication experiment), there is the need specify how one might know that they are performing a replication at all. One common proposal advocated by some in the open science movement (see Nosek et al. 2015) is to perform extensive documentation of experimental protocols and to be maximally transparent about that process. Anjum and Mumford interpret this as attempting to get 'as close as possible' to the original experiment, or as an attempt

[32] One might wonder if 'the same condition' here is redundant if the relevant cause and effect are preserved.

to perform an exact replication, which was reviewed in the section above on 'naïve replication.' This 'as close as possible' heuristic can be a useful one but if it does not capture the causal mechanism being studied in the experiment, it can give rise to an ungrounded sense of certainty in the diagnostic value of the replication. This is seemingly a common overisght.

Given this, Anjum and Mumford argue that replications that attempt to be as close to their target as possible have a minimal diagnostic value and that instead replications that intentionally introduce changes, is what scientists genuinely care about. This is because they are generally assumed to be interested in generalizing their results beyond a particular experiment or a laboratory context, narrowly construed, as well as confirming the robustness of their theories. However, generalizing findings beyond the lab requires an understanding of the causal structure that is at play in the target experiment, and there is no sense in which one can just jump to something conceptual replications or robustness analyses without having this requisite background knowledge. Doing so runs the risk of bypassing the relevant causal structure all together.

There is a sense in which Anjum and Mumford's broader argument concerning causality is confusing their account of replication. Understanding causal structures and the range of confounders to these structures in experiments is not an easy task, as Norton and many others in the replication literature have highlighted. That replications target experiments on open systems, with non-linear processes, that contain hypersensitive elements is simply a testament to this. However, this fact says little if anything about what replications are or how they should work. It does nothing to shift what replications should be targeting. Difficult tasks require modesty and the adoption of an appropriate level of uncertainty.

Given this, at base, this proposal is not obviously an ameliorative one. If anything, it might very well lead research further astray insofar as it implicitly proposes stepping away from the project of identifying causal structures both in replications and target experiments. While it is plausible that a more complicated account of causality as it features in the practice of replication could be helpful in certain contexts, particularly in those in which replication failures have been hastily interpreted, Anjum and Mumford's arguments struggle to motivate this. Still, the difficulties that an account of replication that might be characterized as causal faces with respect to determining the causal

structure of a given experiment is worth holding on to, and being realistic about, even if the solution remains momentarily out of reach.

While Anjum and Mumford highlight some of the important problems facing replication, particularly for the scientific disciplines that have been most effected by the replication crisis, they do little in the way of providing plausibly ameliorative proposals that might bring the practice forward.[33] Perhaps most pressingly, the problem of idealization remains a particularly difficult one. While Anjum and Mumford repeatedly invoke the need for a more realistic treatment of scientific phenomena and the practice of science they fall well below the bar they have set for others and fail to provide a sufficient positive proposal.

1.3.3 Idealized Validity

Several recent papers have placed the general value of replications in question on the basis of the epistemic limitations that the target experiments face. In this section, two of these will be discussed and critiqued; one from Irvine (2021) and another from Feest (2019). While these articles share a number of commonalities, they importantly diverge from one another. This will be highlighted.

Against the backdrop of the replication crisis that at times has placed the whole of scientific knowledge into doubt, Feest's (2019) provocatively stated thesis, is that replications are overrated. She starts by rejecting possibility of exact replication, and in doing so argues that all replications require what she calls 'individuation judgements.' A consequence of this is that, according to Feest, all replications, including direct replications, are conceptual in character. That is because as soon as the possibility of exact replication is rejected, conceptual assumptions necessisarily emerge regarding (1) the relevant features of the dependent and independent variables in addition to how they are described and (2) the construct validity of the experiment in question, which determines

[33] One might object here and claim that this has nothing to do with the intended project and that the idea is to give an accurate descriptive account of replication as a practice. Anjum and Mumford, as well as each of the other non-naïve accounts, however, are attempting to give ameliorative analyses. A clear account of replication and the role it plays in scientific practice should be a clear enhancement given the amount of debate around what to make of failed replications, both in particular cases, as well as when conceived systematically.

whether the experiment in fact measures what it is claiming to measure. In most cases, particularly in fields like psychology and medicine which have been directly subjected to the replication crisis, these assumptions and judgements take the form of including unstated auxiliary hypotheses about which there is a high degree of uncertainty. It is in these contexts that Feest's argument has a particularly strong bite in so far as the importance of probing the description of the relevant effect is arguably more acute when the causal structure under question is characterized by a high degree of uncertainty.

Given this, Feest argues that direct replication is concerned with controlling for random error and is generally directed at assessing the reliability of an experiment. However, it does nothing to address systematic error, which boils down to assessing the validity of an experiment. Random errors are typically thought to be able to be caught by replications because if the effect under investigation is due to confounding factor, varying iterations of that experiment will presumably, under ideal conditions, cause the effect to surface as being unreliable. By contrast systematic error, typically does not present the type of variation that will be caught by multiple replication attempts in so far as they are caused by aspects of the experiment that are thought to be essential to producing the effect. If a systematic error is made both in the target and in the replication experiment, the effect will, at least in theory, remain stable.

However, this aspect of Feest's account is in tension with the idea that direct replications are in fact necessarily conceptual, in the sense just discussed. If this is accepted, then it is plausible that there will be variation in the types of individuation judgements that are made. In this way, and in taking on a number of assumptions, direct replication would be potentially well situated to test the kind of effect that is being investigated. However, this would require a shift in focus towards individuation judgments rather than a shift in the 'type' of replication. Instead of attempting to test the generalizability of claims as a test of the description of the effects under question, as Feest argues for, or attempting to test the external validity of the claim, focus can instead be placed on internal validity while remaining within Feest's description of the problem.[34] Moreover, doing so would allow for the practice of replication to be retained in something resembling its current form, with the appropriate caveats.

[34] This is an aspect of a position that will be defended later in this part of the dissertation (see Part 1 Section 3.2).

Given that systematic errors must be tested for in addition to questioning the description of the phenomenon under question, Feest argues that it is perhaps better to think about replications that involve a high degree of uncertainty as exploratory experiments, rather than as experiments that are directed at hypothesis testing. This would push them towards fully departing from the traditional function of replication, which might make it seem like anything worth being concerned about with regards the replication crisis would be dispensed with. This, however, is not the case.

Feest's strategy can be framed in part as a response to certain aspects of the replication crisis. Given the weight and apparent certainty that often comes with replication failures and the sometimes devasting consequences for the scientists involved in the target studies, in certain contexts at least, it is perhaps sensible to consider reframing the task at hand (e.g., see Dominus 2017). Still, it is not clear that this is necessary assuming that the appropriate degree of uncertainty is made explicit and the ability to probe for systematic errors is built into the practice of direct replication.

This skepticism around direct replication might make it seem like Feest advocates conceptual replication at certain points. This is not the case. Conceptual replications, under their typical framing, are even more demanding that direct replications. This is because they require a grounded conceptualization or characterization of the targeted experiment that allows them to be variously operationalized. Feest then remains skeptical of the value of direct replications, without arguing that they should be abandoned in favor of conceptual replications.[35]

Feest's account of replication and its limitations might again seem to heavily overlap with those of other causal accounts. However, it is not clear that this is so. Her focus throughout is on the conceptual description of the effect that is being studied. While this clearly falls under the causal approach to replication that has been captured in this section more broadly, it is also distinct in its scope and in its recommendations. Feest argues that uncertainty regarding the operationalization of a variable implores research to take up an explicitly exploratory approach to replications, which implies that there would not be an attempt to explain or control the phenomenon under

[35] This again was a mistake made by Anjum and Mumford that Feest manages to avoid.

investigation. In this way, Feest's analysis is important in so far as it places an important aspect of the practice of replication into focus; uncertainty. This remains the case regardless of any shortcomings that her recommendations might have.

1.3.4 Idealized Invariance

As previously mentioned, Irvine's (2021) approach to replication is in many ways similar to Feest's, and again falls under what has been termed the causal approach to replication. Irvine starts with the distinction between what she calls 'good' and 'bad' replications. A 'good' replication is one that can test the core claim of the original study. Within the context of psychology, on Irvine's account, this implies achieving psychometric invariance regarding the target phenomenon. In cognitive science, psychometric invariance implies that the measurement and operationalization of psychological properties can be stabilized across experiments. Just in the way that a light meter on a camera can 'invariantly' measure the amount of light hitting a sensor, in an ideal situation, a psychological property would be able to be measured with the same type of stability and rigor. Being able to do so, would count towards qualifying a replication as being a 'good' one on Irvine's account. These are replications that cannot be easily dismissed in so far as they are able to plausibly diagnose the trustworthiness of the target experiment. By contrast, a 'bad' replication is one that is straightforwardly unable to do this.

Like Feest (2019), and following Stroebe and Stack (2014), Irvine (2021) argues in a certain way that all replications are conceptual in the conventional sense. Again, this is because no two experiments are identical, differences between the replication and its target will necessarily be introduced. Nevertheless, on Irvine's account, direct replications must be able to assess the reliability of the core claim of the target study. Being able to do this requires being able to reliably identify confounders. By contrast, conceptual replications are more demanding and require an understanding of the causal profiles of the measurement procedure as well the phenomenon of interest, in so far as they are attempting to intentionally investigate the target phenomenon from a perspective that is independent of the target experiment. Regarding the target phenomenon,

Irvine (2021, pg. 8) writes that it needs to be known, "what reliably causes it and what it reliably causes in turn, relative to key situational factors".[36]

From this, it follows that even if a purportedly direct replication can assess the reliability of a given experiment, there is nothing that guarantees that it will be able to assess the validity thereof. In this way, the power of replication is at times overstated, which is a position that she shares with Feest (2019). This means that while repeatable experiments show something, it is unclear what that something is, primarily because the target experiment might be measuring the phenomenon in question in a way that is in some way inaccurate, or it might not even be measuring the phenomenon at all. This results in a series of questions about the diagnostic power of direct replications when it comes to theory confirmation.

Can direct replications diagnose the validity of theories? Irvine argues no, primarily because of the state of the psychological literature. If it was clear that the experimental procedures were valid, then direct replications would be able to diagnose something about the truth of the experiments that bear on these theories as Zwaan et al. (2018) claim. However, this is almost never the case in psychology, and it is debatably never the case in scientific inquiry writ large. Given this, the bar is simply set very high, and expectations regarding theory evaluation or confirmation in this context should be set accordingly.

Assume that diagnostic replications require substantial background theoretical knowledge in additional to an extensive understanding of the phenomenon of interest and the experimental methods that are used. It can then be concluded that in many research programs in the disciplines that have been directly hit by the replication crisis, including psychology and the biomedical sciences, replications will not be obviously useful in their current form, and they could very well be profoundly misleading. Given this, Irvine argues, like others who have come before her, that replications should be reframed as exploratory studies that should be put to the service of theory development. Implicit here is the idea that in performing a direct replication, testable claims about

[36] Irvine interestingly holds here that replications are a type of robustness analysis. This is a break from other authors who hold that robustness analysis is not a type of replication (e.g., see Nosek et al. 2022). Of course, what this ends up meaning comes down to the understanding of both of these terms.

the causal structure in the target hypothesis must be made. The integral aspects of an experiment must be distinguished from the disposable ones. In this way, the process of attempting replications as exploratory experiments could be ones in which uncertainty potentially comes to the surface. Carrying them out could be an opportunity to make theoretical claims more precise and to break down complicated theories into more manageable ones that can eventually be tested in hypotheses.

Replication here becomes an iterative process and is in no way a one-shot procedure. It cannot take place in a single round, and it cannot result in conclusive diagnoses of the target experiment, given the nature of exploratory experiments. This is not obviously or necessarily a problem, but it does represent a significant departure from replication as it is currently practiced, and how evidence from replication experiments is validated. Here, like Feest, Irvine invokes Chang's (2004) conceptions of iterative scientific progress as a process of enrichment and self-correction. This is proposed as an overarching guide to how replications can lead to progress in developing sciences like psychology. But again, iterative enrichment is not synonymous with exploratory experimentation for the ends of theory development. The worry here is that the subject seems to have been changed to something other than replication entirely.

Still, there is good reason to be sympathetic to Irvine's interpretation of replication, and the causal interpretations of replications more generally. The move away from one-shot, all-or-nothing style replication has advantages given the high degree of uncertainty that pervades experiments in many of the disciplines that have been affected by the replication crisis. However, again, it is not clear in what sense exploratory studies have anything to do with the practice of replication, or how this does not amount to a wholesale abandonment of replication practices. This is accentuated when Irvine writes, "aiming at replication is deeply misplaced anyway, because it will always be possible to come up with an alternative explanation of why a particular replication succeeded or failed: replication attempts would just never be informative" (Irvine 2021, pg. 23). If the argument is that we should back off replications for the time being, that is one thing, but it is something else entirely if replication experiments should be interpreted exclusively as exploratory experiments

because of the possibility of giving an alternative explanation.[37] Importantly, these are distinct proposals that the account struggles to take a stand on.

This tangle of issues is made more complicated by the fact that Irvine does not specify what an exploratory experiment is, or how she is using the term, which has been given various treatments in the philosophy of science. Exploratory experiments have been historically defined by the fact that they are not designed to test a hypothesis or evaluate a particular theory (Hempel 1966). That is, they are negatively defined as not being confirmatory experiments. However, more recently their role in theory development has received limited attention (e.g., see Feest 2012). One interpretation of this can be found in Franklin (2005), who makes sense of exploratory experiments in terms of local versus background theories. Exploratory experiments on this account will be directed by background theories while leaving local theories unspecified. This allows for a domain of inquiry to be coarsely specified and to do a limited amount of guiding work while simultaneously allowing for a multitude of parameters within the target system to be tested in an unconstrained way. By metaphorically allowing 'a thousand flowers to bloom', features of the target system can ideally be slowly homed in on and eventually captured by testable hypotheses.

If this is the way that the exploratory strategy is being adopted a number of issues remain. First of all, it seems to end up having very little to do with replication as it is currently practiced. Rather, the strategy would be to fully pull back from the practice of replication and to instead focus on getting to the point where hypotheses can be rigorously tested that bear on more full-blooded theories. Maybe, at some point down the line, replications can resurface, although given the persistence of underdetermination it does not seem likely on the terms of this account. While this interpretation has a clear appeal, it is once again in tension with the way that Irvine discusses exploratory replication experiments. It is not clear how it can be known that exploratory experiments are identifying stable phenomena at all. There seems to be faith that this process will

[37] A similar proposal and the problems that come with it is made by Boyle (2021) in the context of comparative psychology. Namely that the discipline should not aim at replications and instead should widen its scope. However, in this context as well it ends up being unclear what role replications can play in scientific practice that is entrenched with uncertainty regarding background theories, target phenomena, and measurement techniques.

progress incrementally in the way that other sciences have, but how specifically this will go is repeatedly left unspecified by both Feest and Irvine. Prolifertaitng exploratory experiments alone will not solve the problems laid out for replication by Irvine. Only hypothesis testing could possibly allow for the type of substantial background knowledge in additional to an extensive understanding of the phenomenon of interest and the experimental methods that are used. Given the history of the progression of the sciences of the mind, a more concrete way forward is required.

1.3.5 Recounting the Causal Accounts

Accounts	**Appeals**	**Limitations**
Norton (2016)	Sets clear criteria for when a replication would be diagnostic.	Demanding in a way that ceases the practice of replication in many contexts.
Anjum and Mumford (2018)	Correctly emphasizes causal structure, albeit an unconstrained view thereof.	Pivoting to conceptual replications potentially makes diagnosis more convoluted.
Feest (2021)	Highlights the way in which replications are description dependent.	Replication is reduced to testing the description of the effects, and external validity.
Irvine (2021)	Handles replications with an appropriate degree of caution.	Claims to preserve replications while simultaneously wholly abandonding the practice.

Table 3. Summary of the causal accounts discussed in this section.

In this section, four causal accounts of replication have been analyzed (see Table 3 above for a summary). Although the respective limitations of these accounts have been emphasized, their respective appeals have also been given attention. This is because they have gone a long way towards specifying what a replication is, and towards improving the practice of replication. It is also important to emphasize, however, that in the sub-optimal epistemic scenarios that characterize many of the disciplines that have been affected by the replication crisis they invariably leave much to be desired. This is a problem that will continue to be addressed and grappled with throughout the remainder of this part of the dissertation.

1.4 A Novel Demarcation Criterion

One of the core virtues of causal accounts is that they attempt to take a step towards clarifying what features a replication should be targeting. In the following, this will be used to weigh in on a debate about which types of research should be held to the replicability criterion. In doing so, a novel and actionable solution to the replicability demarcation problem will be introduced. The argument of this section builds directly on the previous section. However, one could be sympathetic to the causal accounts of replication (as a contribution to an answer to the question: what is a replication?) without thinking that they *also* explain when require research should be required to be replicable.

The use of the term demarcation within the context of the literature on replication is polysemous. In one sense, replicability has been framed as a means of demarcating science from pseudoscience (Popper 2002, Hirvonen and Karisto 2022). This means that if an experiment does not conform to the replicability criterion on this account, then it cannot also be a genuine science.[38] Another sense, however, abandons this wide version of the replicability criterion and asks what types of research should be required to be held to the replicability criterion, while remaining agonistic with regards to the stronger claim that any research that is not replicable should be considered to be pseudo-science. On this approach to the problem, ethnographic research for example, might not turn out to be replicable in principle, and given this it might be held that it should not have to be replicable, but that it still can remain within the realm of what is considered to be genuine science.

In light of the so-called replication crisis, there has been a renewed interest in specifying exactly which areas of scientific research should be held to the replicability criterion. Here, some version of the following question is asked: if replicability is an epistemic norm for scientific research, as has often been claimed, should it be applied across the sciences or just in specific research contexts? Answering this question will help draw boundaries around the scope of the replication crisis and will have direct implications for how research is evaluated. That is, it might not make

[38] This was discussed previously in this part of the dissertation (see Part 1 Section 1.1).

sense to talk about there being a replication crisis in something like ethnographic research. Even if the replicability criterion does not match up exactly with the demarcation criterion between science and pseudoscience, and it turns out that a genuine science might not be replicable, there is reason to think that being able to know what research should be subject to the replicability criterion is useful and desirable. This line has been drawn in a number of ways.

In highlighting the strong amount of variation that exists within scientific inquiry generally and in turn the strong amount of variation that exists within practices of replication, Leonelli (2018) argues that not all research should be subject to the replicability criterion and that in non-standard types of inquiry, that is, types of inquiry that do not allow a scientist to exert a high degree of control over their target of research or to implement standardization, holding this type of research to the replicability criterion is possibly epistemically damaging.[39] On this account, the main determinate of how the replicability criterion is demarcated is the nature of the target phenomenon, construed coarsely on the disciplinary level. While many of the humanities and the social sciences might not be usefully held to the replicability criterion, it is thought that other sciences such as physics and chemistry should be. Leonelli's focus here is on the way in which research in certain fields can be context sensitive and dynamic and using that to draw the line appropriately.

This position is in line with an argument put forward by Rijcke and Penders (2018), who distinguish the type of inquiry done in the humanities from the natural sciences and the type of knowledge that is in turn produced. While the natural sciences are purportedly in the business of producing 'single statement factual outcomes', the humanities allow for multiple valid and at times even clashing answers to the same question that are rooted in social and historical context. In this way, this makes them a poor target for the replicability criterion according to Rijcke and Penders because inquiry in these disciplines can be characterized as being generally more coarse and intrinsically pluralistic. On this account, the humanities are set up by design to invite multiple perspectives on the same subject matter that are not necessarily conflicting with one another. While not stated in precisely this way, this seemingly has to do with the necessarily high degree of uncertainty that comes with work in the humanities. Given this, Rijcke and Penders present a

[39] How 'standard' this type of research in fact is, is highly contestable.

distinction between indifferent and interactive kinds and argue that only indifferent kinds should be subject to the replicability criterion.

Another attempt to lay down the demarcation criterion can be seen in the distinction between living and non-living systems, which maps loosely onto the indifferent and interactive kinds set out by Rijcke and Penders (2018). Both Schmidt (2009) as well as Crandall and Sherman (2016) advocate this view. Under this approach, non-living systems are thought to be uniform and unchanging in a way that living systems are not. Humans undergo new experiences constantly that presumably can change them in some way that makes experimental data dynamic. This therefore situates them as a difficult and potentially inappropriate targets for replication because their history can affect their status as a stable object of inquiry. Crandall and Sherman (2016) therefore argue that there is some sort of distinction between the objects of inquiry in the natural and the social sciences. This distinction is also loosely made by Nadin (2018).

Guttinger (2020) has argued against interpreting these distinctions as strict ones and provides several counterexamples to show that any attempt to draw a sharp line between disciplines will ultimately fail. For example, he discusses the way in which the literature on model organisms has attempted to move away from standardization to increase replicability (see Voelkl and Würbel 2016). In this way, using different strains of animal under the same condition, such as different strains of mice in the same cage, has led to a significant increase in replicability. Guttinger (2020, pg. 10) builds on this and writes, "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." He goes on to argue that the demarcation that the above authors advocate is simply not supported by the science. This is because there are a number of cases that broach every attempt at providing a hard and fast coarse-grained demarcation, and this is primarily because many non-living physical systems are no longer seen as passive.[40] Given this grey-zone of practices that resist previous attempts at defining the scope of the replicability criterion, Guttinger argues in general for a more fine grained view

[40] Another case in which the historicity condition would be violated can be found in the material sciences that are interested in the changes that occur to an object over time like the heritage sciences.

on which a case-by-case analysis needs to be implemented for there to be progress on this question.

A case-by-case analysis will likely not introduce confusion when dealing with empirical research, and in a certain sense it is unavoidable. However, it is not clear that the project of providing a more general demarcation criterion needs to be abandoned, or that Guttinger's approach to the problem is on the right track. Rather than attempting to demarcate the replicability criterion in terms of the nature of the target of inquiry, which is plausibly too coarse to provide a sufficient analysis, an *epistemic* criterion should be implemented that is rooted in a detailed theory of what a replication is. This holds clear normative implications.

While it may be descriptively the case that a certain subset of scientists embrace the replicability criterion and other do not, where this line is drawn should in many cases be moved, and it should be clear that the current line and one being prescribed here are not co-extensive. Moreover, adopting an epistemic criterion requires highly local analysis, which in a certain very abstract way brings it in line with Guttinger's account. However, the overlap ends there, in so far as this account possesses an overarching guiding principle, which Guttinger resists.

On the epistemically directed account of demarcation being put forward, an experiment or line of research should be subject to the replicability criterion *iff* there is an adequate understanding of the causal structure being tested in the hypothesis in the target experiment.[41] An adequate understanding of when an effect or process will produce an outcome that obtains the effect that the experimental protocol was designed to demonstrate, and when an effect or process X will produce an outcome that obtains the effect that the experimental protocol was not designed to demonstrate. This has nothing to do with the nature of the target inquiry per say, although it can obviously be a relevant factor. On this account there will plausibly be islands of experiments in any given discipline that will be subject to the replicability criterion rather than there being a clear line between disciplines, and these islands will loosely track the maturity of a given research program. Moreover, there are areas of research within mature scientific disciplines like chemistry,

[41] This is meant to be coarsely derived from the characterization of replication that is put forward by causal accounts. The particular target of the replicability criterion is going to be determined by the members of the target class.

which are thought generally to be unstable in a way that would allow them to resist being clearly subject to the replicability criterion, simply because the requisite knowledge is not available.[42] In these contexts, the replicability criterion should not be enforced.

Again, while this is a local analysis, the epistemic criterion can make sharper distinctions than the vague and misleading concepts like historicity, living, dynamic, or active will be able to, while also not wholly abandoning the project. Part of the problem is that many of these concepts are highly philosophically disputed. For example, in disputes regarding the active/passive distinction, active and condensed matter physics (Popkin 2016; Walther 2018) will provide a difficult case in which there is potentially no fact of the matter given the current state of the available concepts and the empirical research. Instead, an epistemic criterion that requires a robust understanding of the causal structure of the effect in the target experiment will not clearly pose the same types of conceptual issues although it will surely come with its own context specific difficulties.

1.5 Typological Accounts

Another common and popular way of approaching the problem of defining a replication has been to break the problem down into parts. In the replication literature, this has repeatedly involved specifying several ways in which replications can devolve into a variety of types. Intuitively this might make way for a more detailed analysis of a complicated phenomenon that has proven particularly difficult to get a grasp on. However, there has repeatedly been a focus on features of replication that are potentially misleading. Part of this seems to stem from the vague language that is used to talk about direct and conceptual replications, and the general confusion around the question of what a replication is. Many of these approaches end up focusing on the varieties of bare similarity that can be had between the target and the replication experiment, that should be

[42] One way of conceiving of the maturity of science is in terms of its level of progression. The philosophy of science is loaded with theories of scientific progress, and much of this work is done at an extremely high level of abstraction (Dellsén et al. 2021). One can be found in Popper (1963), who advocated a notion of scientific progress that was centered on his concept of verisimilitude or truth-likeness. This is typically framed as a realist account on which progress in a given science is achieved in so far as it is closer to the truth. Another can be found in Kuhn (1970), Laudan (1977), and Lakatos (1980) who have argued for paradigmatically anti-realist accounts that are centered on solving as problems, be they conceptual or empirical, that are put forward in a given research tradition. To adopt Kuhn's terms, it might be the case that comparative cognitive science is in a pre-paradigmatic phase. To use Lakatos' (1980) language, it might be the case that comparative cognitive science is a budding, or even declining, research program.

familiar from the discussion of the naïve accounts earlier in this part of the dissertation (Section 1.1). The overarching lesson from these typological accounts of replication is that tracking an unspecified notion of similarity and scaling that with context will almost surely be a poor guide to evaluating trustworthiness and supporting the practice of replication. In the following, a small selection of examples that have adopted this approach are analyzed and their shortcomings are highlighted.

1.5.1 A Replication Spectrum

In addressing the role of replication in the computational sciences, Peng (2011) introduces a replication spectrum that scales according to the amount of information available (See Figure 3 below). This analysis is intended to address multiple issues. First, it sets out the information that needs to be provided for a replication to be performed. Second, it sets out a variety of forms that a replication can take given the various levels of information that the research makes, or is able to make, available.

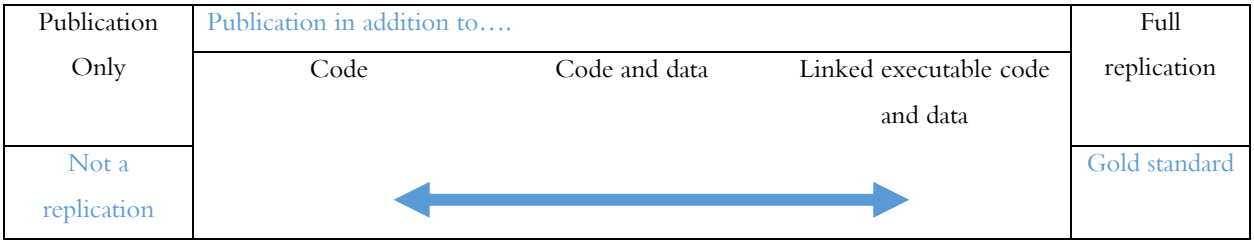


Figure 3. A model of a replication spectrum adapted from Peng (2011, pg. 1226)

On one extreme, scientists provide linked executable code and data, which allows for a full replication to be performed. On the other, only the publication is made available, which implies a situation in which a replication cannot be performed; it falls outside of the typology. Peng's spectrum is a way of describing the variety of replications in computational research. The degree to which opacity permeates computational research on this account is cause for concern. While Peng concedes that some research is not practical to replicate, such as the Sloan Digital Sky Survey, which makes use of massive complex data structures to allow astronomers who do not have telescopes to do computational research, nevertheless, computational science should perform

replications and this spectrum is a means for describing a gap that needs to be closed in the discipline. Peng's spectrum stems from a review of microarray gene spectrum analyses which included studies that fell along different points on this spectrum in accordance with the amount of information that was provided in the published paper.

Peng makes three recommendations for establishing norms of inquiry. First, code should be required to be published in some form on an online platform such as Github for all computation based scientific research. While according to Peng's typology, this alone would not allow a full replication to be performed, it would allow for another less 'complete' type of replication to be performed. Second, cleaned code along with data sets should be published and made widely available. And finally, a centralized database should be created for computational based research that will allow for code to be linked. If all three conditions are met a full replication will be able to be performed.[43]

Peng's account of replication importantly highlights the way in which issues around transparency can be central to the practice of replication and how this can affect various fields and lines of research in distinct ways. This aspect of the account is worth taking seriously. However, it is not clear that the spectrum Peng introduces is doing the work that he envisions.

Peng posits that an experiment becomes gradually more replicable as the number of components available for replication increases, and that given that the information provided will be various, that there will be a spectrum of types of replication that will exist in computation based scientific research. However, there is reason to think that replications function more like a step function. To illustrate this, take a bicycle analogy under consideration. Under condition (1) you might have the frame and the groupset (i.e., the front and rear derailleur, the brakes, the calipers, etc.), and under another (2) you will have this this plus a chain. However, absent a set of wheels you are not going anywhere in either condition (1) or (2). Relative to this function, both are equivalent.

[43] One curious feature of Peng's spectrum seems to be that it does not allow for the possibility of only the data and the publication being provided, which is certainly conceivable. This is one of the many limitations of this one-dimensional spectrum.

In the same way, a replication will only be useful if the right set of information is available that will allow it to be diagnostic of the target experiment. Peng argues that a condition in which (1) a study that provides the published the paper and one in which (2) the code alone is less replicable and constitute different types of replication. However, just as in the bicycle example, it seems very plausible that both cases are functionally equivalent in their ability to assess the trustworthiness of the target experiment; this could be fully diagnostic or not at all depending on the case at hand. The larger point is that Peng's spectrum is not able to pick up on this imporant difference, and it is an essential one for carrying out replications. Instead, it tracks the type and amount of information that is made available in order to establish a spectrum of replications that have dubious diagnostic capacities.

1.5.2 A Replication Continuum

Another example of this approach to replication can be found in LeBel et al. (2017), who have put forward a typology of replication that scales according to how similar a replication's methodology is to its target. These vary according to the following:

- the number of design facets
- the operationalization of the independent variable
- the operationalization of the dependent variable
- the independent variable stimuli
- the dependent variable stimuli
- procedural details
- physical setting
- contextual variables.[44]

Additionally, there are five types of replication on this account. These types scale according to a similarity relation between the replication and its target.

[44] It is worth noting that there is no standardized or agreed upon way of counting what is included in the methodology and what is not. For example, while Guttinger (2020) broke variable operationalization out from the methodology, LeBel et al. (2017) do not do this.

- exact replication (all facets are the same)
- very close replication (the procedure and the context are different)
- close replication (the independent or the dependent stimuli are different)
- far replication (the independent or the dependent variable operationalization are different)
- very far replication (everything is potentially different).

Generally, LeBel et al.'s (2017) continuum gets its pull from the idea that a replication experiment and its target can be relevantly similar or different from each other in several ways. Although there are five-types of replication that LeBel et al. lay out, the claim is not that there are not hard boundaries between these five types. Similarity is instead conceived of as a continuous relation. Each of these types are thought to have a distinct function. For example, exact replications are thought to be able to confirm the existence of the phenomenon of interest, while instances where there is a varying amount of difference between the replication experiment and its target (such as in very close and close replications) are thought to contribute to assessing both the validity of the experiment as well as the generalizability of the claim that is being made. Part of the upshot of LeBel et al.'s typology is that exclusively favoring what they call exact replication is thought to be obfuscating in so far as it blocks these other factors from being assessed. Therefore, on their account, only by employing a variety of replication types along this continuum can the trustworthiness of an experiment be analyzed, because each distinct function is purported to have a distinct epistemic role to serve.

While LeBel et al.'s account is appealing in so far as it attempts to give a more fine-grained analysis of the variety that exists within the space of potential replication, it also encounters difficulties. The first has been identified by Irvine (2020, pg. 9), who writes that typologies such as LeBel et al.'s," […] certainly tracks one notion of similarity, but it may have little to do with causal independence, and this is the crucial feature. In principle, a procedure that incorporates changes to all the design facets of the original study may still be fairly similar in terms of basic causal structure." This highlights both the limitation of this typology, and the broader limitations of a kind of typological approach to the question of what a replication is. While LeBel et al.'s account associates particular types of changes that can be made with particular types of replication

(e.g., very close replication on this account allows for the procedure and the context to be different), it lacks the resources to account for which difference are the relevant ones. The procedure and the context can vary between the replication and its target in an unbounded manner. Instead, the causal structure of the effect tested in the target experiment must be preserved. Otherwise, the replication, regardless of the type, will fail to be a diagnostic one.

LeBel et al.'s account lacks the resources to determine what the right type of difference should be, or why changes along these dimensions at all will be relevant to assessing the target experiment. In addition to this, more pressingly, their account lacks a theory of what counts as a difference *full stop*. Absent these, it will potentially track the wrong types of changes, which will lead to a misleading assessment of the target of the replication. Consider LeBel et al.'s account of exact replication, which again requires that all facets of the experiment be held stable. This is impossible when interpreted literally, so choices need to be made about what to count as being *close enough* to admit into the space. While LeBel et al. concede in a footnote that their account is going to admit borderline cases, this misses the point. Absent a theory of why two experiments are the same or different in the right way, even so-called 'exact' replications, which have been repeatedly rejected, are not going to be diagnostic.

1.5.3 Replication as a Sequence of Different Studies

Another typological account of replication within the context of social psychology can be found in Hüffmeier et al. (2015), who argue that replications should be reconceptualized as a sequence of different studies and offers five hierarchically organized distinct types of replication study: (1) exact replication, (2) close replication, (3) constructive replication, (4) conceptual replication in the lab, and (5) conceptual replication in the field. These five types then vary according to the following criteria.

- Authorship
- Aspired comparability to the target
- Typical situation for replication
- Contribution to theory building

- Less likely explanation if the replication is successful

Like LeBel et al.'s (2017) continuum, each of Hüffmeier et al.'s types of replication have a distinct function. For example, 'exact replications' are meant to protect against false positives, while 'conceptual replications in the lab' are meant to test the extent to which the findings are contingent upon contextual factors in the lab that the target experiment was conducted in. Hüffmeier et al. claim that their typology is comprehensive and that it should be thought of as a sort of hierarchical procedure. This means that replication attempts should proceed successively from types (1) to (5). The idea here is that if there is a replication failure in type (1) exact replication, then replication attempts should not proceed from (2)-(5). This is because (3) constructive replication is thought to be more theoretically demanding on this account than a (1) exact replication, in so far as it requires more theoretical knowledge about how the effect generalizes. Given this, Hüffmeier et al. argue that replication experiments should take place in steps.[45] If an experiment passes type (1), then it can proceed to type (2), and so on. This is both a way to manage limited resources and to preserve diagnostic clarity. Moreover, an experiment that is replicated under (4) conceptual replication alone, has a purportedly distinct meaning from an experiment that has been replicated under types (1)-(4), in so far as it has been more extensively probed, although it is not clear how this situation would arise if their reccomendations are taken seriously.

I take it that there is something important in the hierarchical aspect of this approach to replication.[46] Once again, given the account's reliance on similarity, it remains a question what is required for each type of replication to obtain. In each case, their answer to this question is inadequate. For example, in explaining type (1) exact replication, they argue that the replication must mirror its target, "in all aspects", (Hüffmeier et al. 2015, pg. 83). However, once again, changes between a replication and its target are inevitable; exact replication is impossible. For this reason, these changes and differences must be theorized. In its current state, Hüffmeier et al.'s account lacks the conceptual resources to show why some changes are relevant and others are not.

[45] It is worth mentioning that this is a departure from the more orthodox one-shot replication experiment model.
[46] The virtues of this aspect of the account will be revisited later in this part of the dissertation in giving an analysis of Amodio et al.'s (2021) replication experiments (see Part 1 Section 5).

Instead, like many other typological accounts, they are left appealing to bare similarity, which is significantly unconstrained. This places their account of replication as a sequence of experiments in a difficult spot, without a clear path towards practical implementation. If the account potentially tracks the wrong aspect of an experiment at any given type in the series, then the analysis as a whole will potentially cease to be a diagnostic one. In this way, absent a richer theory of what a replication is, and what makes the relationship between a replication experiment and its target salient, the account encounters serious difficulties for reasons that are shared with LeBel et al.'s (2017) continuum which was analyzed above.

1.5.4 Replication Pluralism

Another replication typology can be found in Leonelli (2018), who argues that replication is a practice that admits both a variety of interpretations as well as a variety of functions across the sciences. This means that if replicability is pursued as a general acontextual benchmark for scientific inquiry, the trustworthiness of research program and even entire disciplines will be potentially misdiagnosed. On this account, which intended to be primarily descriptive, there are a variety of ways in which replicability is pursued that are context relative. In addition to this, there are many high-quality research contexts in which replicability should not be pursued as an end, that are nevertheless genuinely scientific. These considerations lead Leonelli to introduce a replication typology in which the degree of control over the environment, and the reliance on statistical methods that scientists have in a context, determine the appropriate type of replication that should be implemented. For example, computational engineering and informatics are held up as paradigm cases under which control is absolute and the reliance on statistics is high and made explicit. She calls attempts at replication in this context 'computational replicability', which is thought to obtain when the same data is used to obtain the same results.[47]

Leonelli's account runs into two main problems; it (1) overextends the practice of replication and the (2) the proposed typology tracks the wrong features of research. The first problem can be brought out by returning to Leonelli's account of 'computational replicability.' There are two

[47] Leonelli (2018) uses the word reproducibility. However, I will again use replicability for the sake of consistency. I see no reason to think that this will make a relevant difference.

ways of interpreting this type of replication. The first amounts to essentially checking that the calculations were performed correctly, and do not have an obvious connection to replication as it is currently practiced. In a paper that is heavy in logic, formal computation or mathematics, this might involve checking the proofs. While this is at least partially what Leonelli has in mind, this type of replication seems to target any study that makes use of quantitative methods that straightforwardly involve numerical or symbolic calculations. The second interpretation involves broadening the scope to computational methods more broadly. On this interpretation, it is unclear that the degree of control that Leonelli claims is characteristic of computational methods is present in every or even most cases, particularly those that make use of AI, ML, and deep learning methods which have notedly been subject to the replication crisis (see e.g., Hutson 2018).

If the second interpretation does not hold, and the first interpretation is endorsed, another question arises; namely, should this be considered a replication at all or is this an entirely distinct type of reliability assessment that falls clearly outside of the practice of replication. It seems uncontroversial that calculations should be checked for errors and that not doing so poses a risk to the quality of research. In fact, something similar has been argued by Nuijten et al. (2018) who hold that prior to performing replications, it is important to verify that the target study does not contain errors in analysis and that doing so can save time, labor, and resources more broadly. However, importantly, what Nuijten et al. do not claim is that this is a type of replication, and why would they? The domain of replication, under Leonelli's account, has ballooned. In attempting to capture the variety of types of research across the sciences, and the way in which this licenses a seemingly legitimate diversity within the practice of replication across a variety of scientific disciplines, Leonelli's construction of an ever more sprawling account of replication risks overburdening replication with tasks that it has nothing to do with. Doing so risks confusing important tasks that are not directly or currently tied to replication, such as checking calculations. The strategy, moreover, amplifies the risk of mischaracterizing distinct practices that have been lumped under the header of replication. In this way, Leonelli's pluralist approach to replication is surprisingly singular and constrained in its ability to account for the variety of trust-inducing practices that are potentially available to scientists and epistemic agents broadly construed; in particular to a scientist making use of computational methods. Similar questions regarding the utility of an inflated account of replication can be asked of Leonelli's other types such as scoping

replication, indirect replication, and hypothetical replication (see Table 4 below for a brief description of these types). In contrast to computational replication, these types entail situations in which there is a limited degrees of control over the environment.

Type of replication	**Description**	**Degree of control on environment**	**Reliance on statistics as inferential tool**
Computational replication	Obtain same results from the same data.	Total	High
Direct replication	Obtain same results from different runs of the same experiment.	Very high	High
Scoping replication	Use differences in results to identify relevant variation.	Limited	Variable
Indirect replication	Obtain same results from different experiments.		
Hypothetical replication	Corroborate results implied by previous findings.		
Reproducible expertise	Any skilled experimenter working with same methods and materials would produce similar results.	Low	Low
Irreproducible observation	Any skilled observer would pick out similar patterns.	None	Low

Table 4. Table adapted from Leonelli (2017; pg. 12)

Here, the experiments/observations are repeated under these various types in different ways, for different reasons. Given this, it is not clear that the practice is still being held together under the header of replication at all, and again there is a potentially high cost for attempting to lump these seemingly various practices together. For example, repeating experiments to identify relevant variation in results, as is the case in scoping replication, seems to have little to do with reliability assessment, which many in the replication literature hold to be central. As Guttinger (2020) briefly highlights, failures of scoping replication would not seem to warrant the type of concern that seems to come with the replication crisis, if they warrant concern at all. Moreover, it is not even clear how scoping replication failures would be able to arise, given that they are not even

explicitly aimed at testing the reliability or trustworthiness of an experiment. Again, this should lead us to think that replication as a practice is being stretch beyond its limits. Too much is being asked of it; it is being overburdened.[48] This means that both the practice of replication and the results that arise from it become increasingly difficult to interpret and navigate.

The more serious problem for Leonelli's account lies in her selection of parameters, which determine how the various types of replication are categorized and the content they take on; the (1) degree of control over the environment and (2) the reliance on statistics as an inferential tool. Broadly, Leonelli's proposal is that how a type of research scores on (1) and (2) determine the type of replication that should be used to analyze it, and that not doing so will lead to a distorted view of the science. While there might be a sense in which these parameters are indirectly indicative of how easy it will be to settle on what counts as a replication in any given context, this will not necessarily be the case. This is because these variables can come apart from the richness of background knowledge about the causal structure that is being tested. Again, it is this that determines how diagnostic a replication will be.

In this way, one can plausibly have (1) a very high degree of control over the environment and (2) a very high reliance on statistics as an inferential tool, without being subject to computational replication that requires that the same results be obtained from the same data. This, for example, might occur in a computational biomedical context in which a new research program is being established and the causal mechanism that is being studied is poorly understood if at all. In these cases, statistics feature about as prominently as it possibly could, and the environment is highly controlled, but this research should not be expected to perform a 'computational replication' in Leonelli's sense, in which the same results are necessarily obtained from the data.[49]

Likewise, one can conceivably have (1) a low degree of control over the environment and (2) a low reliance on statistics as an inferential tool, while still being rightly subjected to a high standard of replication that requires that the same results to be obtained from the same data. This, for

[48] The risk of doing so has been briefly touched upon in the analysis of Nosek and Errington (2020) previously in this part of the dissertation (see Part 1 Section 1.2.2).
[49] It is being assumed here that the pool of data is being resampled from. This might also occur in an ML context in which the training set is drawn from a larger relevant pool of data.

example, could occur when there is a high degree of richness of background knowledge about the causal mechanism that is being tested for, in something like certain studies on the mechanisms of perception in cognitive science. While these studies typically have a low degree of variance, the environmental conditions are not necessarily well controlled, and they do not always rely on statistical methods to a high degree. This parametrical misalignment runs Leonelli's typology into difficulties. This is because it predicts that the function of replication and the ability to replicate experiments will correspond with variation along these two parameters. While these might sometimes serve as a proxy for the degree of background knowledge about the causal structure that is being tested for, in many cases it is very plausible that these will come apart.

In sum, Leonelli's typological approach goes wrong in two respects; it (1) overextends the practice of replication and (2) the proposed typology tracks the wrong features of relevant research. Given this, it has the potential to confound the practice of replication: it should not be adopted.

1.5.5 Why Typological Approaches Fail

Three typological accounts of replication have been discussed (See Table 5 below for a summary). Each account was shown to make a similar mistake.

LeBel et al. (2017)	**Hüffmeier et al. (2015)**	**Leonelli (2018)**
Exact replication	Exact replication	Computational replication
Very close replication	Close replication	Direct replication
Close replication	Constructive replication	Scoping replication
Far replication	Conceptual replication in the lab	Reproducible expertise
Very far replication	Conceptual replication in the field	Irreproducible observation

Table 5. The three typological accounts that have been discussed.

If there are any lessons to be drawn from the literature around the definition of replication that has emerged over the past 10-15 years, it is that (1) replication as it is currently practiced contains an enormous diversity and that (2) replication successes or failures are almost never interpreted in a straightforward way. In the early days of the so-called 'replication crisis', Gómez at al. (2010) documented over 79 types of replication that have been proposed within abundant typologies across various scientific literatures. This included both instances in which the same term was used

in different ways as well as instances in which different terms were used in the same way. This variation was identified both within and across several disciplines. In the 11 years since this study, as interest in the replication crisis has exploded, the number of typologies and definitions has increased exponentially and has resulted in a bloated and deeply disjointed literature. At this point, the introduction of even more typologies or types of replication, or even more terminologies threatens to be more obfuscating than clarifying. This can be seen clearly in the three cases above, in which the wrong features of replications are being tracked.

It might seem that there are at least three different projects on the table. The first is to give an account of what a replication is. The second is to give an account of the varieties that are contained within that practice. And the third is to account for what makes an experiment replicable and the type of variety that exists within that practice. While these are admittedly distinct projects, the issue of what counts as a replication, or the question of what a replication is, is fundamental to each of them. Its relevance in the first case is obvious. In the second case, giving an account of the varieties of replication that can exist is entirely contingent upon what a replication is.

In all cases discussed in this section, if replications vary along a continuum that has nothing to do with the relevant features of what makes an experiment a replication, then the account has gone awry. Moreover, this has direct implications for the third project of demarcating replicability, is so far as you cannot even get off the ground with this project unless you have a sense of what counts as a replication. In this way, while these projects might seem entirely distinct, they are held together by a fundamental question: What is a replication?

As a general strategy, breaking down a complex problem into parts is a reasonable way to make it more tractable. However, typological approaches to replication are only as good as the criteria that are used to establish the typology. Too often, these criteria have targeted the wrong or all too vague features of the target experiment. Any typological account of replication that makes this mistake must be rejected.

1.6 Guiding Lessons from Prior Accounts

In section 1 of the first part of this dissertation a variety of takes on replication were analyzed and critiqued. By bringing out both the virtues and vices of these accounts, the outlines of the positive account that will come in section 3 have been introduced. In this section, the guiding lessons taken from these accounts will be briefly summarized. The discussion will broadly address the following: (1) naïve accounts, (2) the resampling account, (3) the diagnostic account, (4) causal accounts, and (5) typological accounts. For the sake of brevity, not every account replication that has been discussed will recounted. Rather, the broader lessons will be extracted that will be taken forward to the positive account. These are summarized in Table 6 below. While the lessons are unable to wholly determined the content of the deflationary account that will be introduced in Part 1 Section 3, they go a long way towards pointing it in the right direction.

Account	**Core Lessons**
(1) Näive Accounts	a. Appeals to relevant similarity or identity cannot sufficiently shore up an account of replication. A successful account must invoke other criteria.
(2) The Resampling Account	b. The replication and its target should avoid procedural evaluations and breaking replications into components. A successful account should rely on holistic evaluation.
(3) The Diagnostic Account	c. A successful account of replication should avoid standing in for all trust inducing practices.
(4) Causal Accounts	d. A replication should test the causal structure of the effect tested in the target experiment, while also not being demanding in a way that makes replication experiments impossible to carry out.
(5) Typological Accounts	e. Replications can test various aspects of the effect tested in the target experiment, but similarity is not a reliable guides to accounting for this type of variation.

Table 6. A summary of the virtues, vices, and lessons taken from part 1 section 1 that will come to shape the positive account that will be introduced in part 1 section 3. These discussion of these accounts will provide

What has been termed (1) naïve accounts in Part 1 Section 1 includes direct and conceptual replication and was generally meant to capture the idea that no additional analysis of replication was needed at all. The problem with this approach to replication was brought out and the extent to which such accounts leaned heavily on conceptions of relevant similarity and identity. It was

argued that these conceptions were underspecified in a way that was potentially harmful to the practice of replication. The core lessons from these analyses going forward is that appeals to relevant similarity or sameness are insufficient guides to an account of replication.

In introducing the (2) Resampling Account, Machery attempted to give a clarified account of replication by breaking the target experiment into components that can be then randomly sampled from in order to compose the replication experiment. However, the analysis of the account demonstrated the way in which this particular type of proceduralism is unable to derive valid replications from sets of components. In place of proceduralism, a holistic approach to establishing and evaluating replication experiments against their targets should be taken.

Nosek and Errington's (3) Diagnostic Account was appealing in the simplicity of its framework for establishing and evaluating replications. However, its radical approach to the question of what a replication is, left it overextended. A positive account of replication should be maximal in scope; however, this should not come at the cost of making replication a stand in for every trust inducing practice. Where possible, a successful account of replication should resemble the current shape of the practice, in a very general sense.

The analysis of the (4) Causal Accounts provided a focal point for a positive account. A replication should test the causal structure of the effect tested in the target experiment, while also not being demanding in a way that makes replications impossible to perform or requiring that the practice be wholly demanded. Instead, the limitations of any particular replication experiment should be built into the practive of understanding the trust worthiness of the target experiment.

Finally, it was shown that although the (5) Typological Accounts were pervasive in the replication literature, they often failed to track the correct respect in which replication experiments are in fact able to vary. Rather than varying in terms of relevant similarity with relation to their target, a successful account of replication will instead capture the variety of ways in which the causal structure of the effect tested in the target experiment. Such an approach will both include the variety of type of replication experiment while also allowing for focus to be established upon the relevant features of the target experiment.

In sum, Part 1 Section 1 has made a number of broad contributions to the task of answering the core question of this part of the dissertation: What is a replication? First, it demonstrates the extent to which the dominant accounts have proven to be insufficient. This affirms the extent to which an account is in fact needed at all. Second, it makes the missteps of the previous accounts clear. This serves the dual function of both making it easier to avoid the miststeps of previous accounts while simultaneously directing a positive account. The lessons from this section will be taken forward, particularly in Part 1 Section 3 in which the deflationary account will be introduced in more detail.

2. Theory and Evaluating the Replication Crisis

In the following section, three interpretations of the replication crisis are analyzed. The first is conceived more broadly and will include an evaluation of the replication crisis across several sciences. By contrast, the second and third will target research specifically in the psychological and cognitive sciences; developmental psychology and comparative cognitive science respectively. In performing these analyses, the previous sections are explicitly drawn upon in order to give a fine-grained analysis of the nature of the crisis in these contexts. Most directly, how the question, what is a replication is answered holds direct implications for how the replication crisis, conceived broadly or narrowly, should be evaluated. Given this, across each of these accounts, it will be argued that there can either be a replication crisis or a theory crisis, but that both cannot exist simultaneously.

2.1 On Base-Rate Fallacies and the Replication Crisis

Bird (2020) has argued that the interpretation of replication failures as being indicative of bias and fraud is mistaken and that the replication crisis can instead be explained as a base-rate fallacy. The claims is that if most hypotheses that are being tested are not in fact true and the NHST (null hypothesis significance testing) significance level is low but non-negligible, then there is reason to expect that a high degree of successful experiments will in reality be false positives.

As has been studied extensively by the heuristics-and-biases research programs, our intuitions about probability are very often misleading (Kahneman and Tversky 1973). Specifically, the base-rate fallacy is thought to occur in particular instances in which a general phenomenon is inferred from a particular occurrence. However, there is much debate regarding under what conditions base-rate information is neglected and exactly what the mechanism is that causes this type of neglect to occur (Koehler 1996, Barbey and Sloman 2007).

A well-known example of the base-rate fallacy comes from a medical context. Imagine that disease X is found in 1 in 1000 patients and that the available test for disease X has an accuracy of 95%. When Casscells et al. (1978) asked the medical students in their experiment what the probability was that a patient P who tested positive in fact has disease X, only eleven of the sixty students (18%) gave the correct answer (that is, 1.9%).[50]

Calculating the correct answer involves taking the following steps (see Table 7 below for reference).

S…	**Has the disease**	**Does not have the disease**
Tests positive	πr	$(1- \pi)(1-r)$
Tests negative	$\pi(1-r)$	$(1- \pi)r$

Table 7. Adapted from Bird (2018).

- Identify the probability that S has the disease without considering their results. Call this π.
- From this derive the probability that S does not have the disease $(1-\pi)$.
- Call the accuracy of the test r.
- Then divide the probability that S has the disease and tests positive by the probability that S tests positive. $\pi r / \pi r + (1- \pi)(1-r)$ is then equal to the probability that S has the disease given that they have received a positive test.
- So, $(0.001x0.95)/ (0.001x0.95)+(1-0.001)(1-0.95) = 1.9\%$

[50] The following explanation of the base-rate fallacy is loosely derived from an explanation given by Bird (2018).

The most common answer given by the students in Casscells et al.'s (1978) study (27 of 60 or 45% of the students) was 95% and the answers ranged from .095% to 95%. Presumably, the students who got the most common wrong answer simply equated the accuracy of the test with the probability that patient P in fact has disease X.

Bird's claim is that existing analyses of the replication crisis commit the base-rate fallacy in the same way that the students did when they calculated the probability that S in fact has disease X given that they received a positive test.

To make this more concrete, Bird asks his readers to assume (1) that in a given field afflicted by the replication crisis such as psychology, 10% of hypotheses that are tested are true, and (2) the testing is 95% accurate in fields that use NHST, then it turns out that when chosen at random, the probability that a given hypothesis is true given that it has passed the test is 68% or around two-thirds, implying that around a third of hypotheses that were successfully tested in fact pass the test. This is a provactive argument in so far as it claims to show the way in which NHST taken together with a seemingly reasonable assumption regarding the accuracy of hypotheses in a discipline alone can yield the results that are able to set off a replication crisis. This argument interestingly demonstrates the way in which this can happen even when every aspect of the science is functioning as intended. That is, there need not be the presence of something like fraud or a fundamentally defective statistical method for a crisis to arise.

While this arugment holds a number of appeals, its foundation is less solid than it might initially appear. The focus in the following will be on assumption (1) of the argument for the base rate fallacy model of the replication crisis that Bird puts forward.

Bird's defense of assumption (1) is purportedly derived from a study by Johnson et al. (2017). However, this study in no way shows what he claims that it does. Bird (2018, pg. 8) writes," only 10% of the hypotheses tested are true, is in fact the case in psychology, according to an analysis by Johnson et al. (2017)". But Johnson et al. (2017) do not argue this point. The closest they come to saying anything like this can be found in the following: "Although it is difficult to assess the proportion of all tested null hypotheses that are actually true, if one assumes that this proportion is

approximately one-half, then these results suggest that between 17% and 25% of marginally significant scientific findings are false" (Johnson et al. 2017, pg. 19316). Therefore, assumption (1) is very much an assumption and not robust data as Bird misleadingly implies with this citation. Moreover, there is an a-priori reason to be skeptical of this position given that it is unclear how reliable data at this level of assessment could possibly be obtained.

Bird's defense of assumption (1), however, does not end there. His second argument backs off from direct claims regarding the truth of the hypotheses that are tested in a discipline such as psychology, and is rooted in the claim that sciences with a low π, or a low prior probability of being true, are ones that are, "[…]dominated by a well-established and well-confirmed theory and the hypotheses of the field test specific aspects of this theory or applications of the theory in particular domains, or they extend the theory in plausible ways" (Bird 2018, pg. 10-11). Bird here contrasts disciplines like physics, with ones like biomedicine and psychology. To get an example on the table, the experimental discovery of the Higgs boson particle in 2012 was originally theorized to exist in the 1960's. The Higgs mechanism is an extension of the standard model which is well empirically confirmed. Fields like biomedicine or psychology, however, differ from this in that they do not have an equivalent or even comparable overarching theory like the standard model. For example, in biomedicine, the typical systems that are studied are extremely complex, and in many ways under theorized. While our understanding of them is seemingly improving, the connection between overarching theories in these disciplines, if they exist at all, and the hypotheses being tested in them cannot clearly be described as an extension of overarching theories in the way that they can be in physics. This is because there is deep uncertainty both with regards to the hypothesis being tested and the overarching theories that are proposed, again, assuming they exist at all.

If the situation is somewhat dire in medicine, in the psychological and cognitive sciences, the situation is only worse in so far as there is virtually universal agreement that the discipline lacks an overarching theory. This has meant that hypotheses that have been tested in the discipline have frequently been formed analogically or intuitively, and that this has tended to leave the science in

a state of disjunction. From this, some such as Muthukrishna and Henrich (2019), have claimed that there is a high probability that the hypotheses being tested are not true.[51]

Bird, moreover, appeals to the effect that falsity breeds falsity. The idea here is that if false hypotheses are confirmed, they are in turn taken as a model or inspiration for the formation of further hypotheses, which is a problem even if the relationship between the two hypotheses is unfounded. Given this diagnosis Bird suggests three ways forward: (1) doing nothing and accepting this is part of normal science, (2) generating higher quality hypotheses, and (3) increasing significance thresholds (requiring a rather lower α than the 0.05).[52] The focus here will be on (2).

In the context of biomedicine, Bird holds that higher quality hypotheses can be obtained by doing more preliminary research which will presumably be able to fill out the background theory. In many ways, this is resonant with analyses above given by Irvine (2021) and Feest (2019), however, again, seeing that in psychology there is no clear or generally agreed upon background theory, another remedy is in order. Bird hypothesizes that more replications will be able to weed out the low-quality hypotheses thereby resulting in a sort of positive feedback loop.

However, this assessment of the problem again brings out the way in which Bird's base-rate fallacy account of the replication crisis is more ambitious than the available evidence can make good on. This is because it relies on meta-philosophical claims about the veracity of hypotheses in certain disciplines such as psychology that have been affected by the replication crisis that it is in principle unable to confirm. First, there is an important distinction to be made between the truth of hypotheses and the ability to detect the truth of these hypotheses. This distinction becomes particularly salient in fields like psychology where the validity of experiments is often under question. Absent an overarching theory, the problem is less that most theories in psychology are

[51] Muthukrishna and Henrich (2019) then go on to propose the adoption of dual inheritance theory as an overarching model for psychology. However, staking the discipline on an imprecise and speculative theory might very well make the problem worse. Certainly, rushing into the adoption of an overarching theory without proper evidential support is ill-advised. This issue will be further addressed in the next part of this dissertation (see Part 2 Section 2.1).

[52] There has been much discussion in the literature on the replication crisis of strategy (3). For example, Benjamin et al. (2018) have proposed decreasing significance levels by an order of magnitude. For criticisms of this proposal see Mayo (2017) and Lakens et al. 2018. For a defense of this proposal, see Machery (2021). Adopting this strategy would relegate comparative cognitive science almost exclusively to the realm of exploratory experimentation.

determinately wrong, and more that it is a struggle to show whether these theories are cable of being truth tracking at all. That is, many hypotheses in psychology are formulated in such a way that makes a clear evaluation of them fundamentally indeterminate.[53] This becomes even more clear when replications attempts are made. As will be argued later in this section, if the target experiment is invalid, a replication failure of it is not going to be a diagnostic test of the target hypothesis. In this way, Bird's inferences are not evidentially grounded, which results in the base rate fallacy model of the replication crisis losing its footing. Put more succinctly, Bird has no way of knowing the base rate of true hypotheses in psychology and building this into the core of his model is ultimately detrimental to the broader account.

2.2 Fragility, Theory, and Replication

Lavelle (2020) argues that the role of theory in replication has been neglected in the literature and that this has led to a misdiagnosis of the problems underlying the replication crisis. She illustrates this point through an analysis of a series false belief tasks in the developmental psychology literature. On the basis of a set of commentaries, most notably by Gil Harman (1978), Daniel Dennett (1978), and Jonathan Bennett (1978), on experiments done by Premack and Woodruff (1978), Wimmer and Perner (1983) developed what would come to be known as the false belief task, the Maxi task, or the Sally Anne task. This experimental task was designed to be one that isolated ToM (theory of mind) as a variable so that one could not possibly solve the task without possessing this cognitive capacity. These early false belief tasks typically involved verbally narrating a story to a child and then asking them to explicitly solve a task for the experimenter that involved tracking the mental states of others.

For example, in the setup a child will be set up in front of a screen that plays a video.[54] In the video, two characters named Jaden and Willow are sitting in the kitchen eating steamed buns. Willow likes hot sauce on her bun, so she goes to the refrigerator, gets the bottle, pours it all over her bun, and then returns it to the fridge. Willow eats some of her bun and then gets up and leaves the kitchen to go to the bathroom. While she is out of the room, Jaden gets the hot sauce

[53] This is a point that will be returned to in Part 2 Section 2.2 of this dissertation.
[54] This is a fictional example.

out of the fridge and hides it in the pantry. Willow then returns to the kitchen and sits with her bun. At this point the child is asked where Willow will go to get the hot sauce. Children who are three and younger typically will say, falsely, that Willow will search for the hot sauce in the pantry. Children who are four and older will typically say, correctly, that Willow will search for the hot sauce in the fridge.

From this, Wimmer and Perner (1983), and subsequently many others, claimed this was because children under the age of four were unable to reliably attribute mental states to others. While this style of task was very influential and there are hundreds of variations on it in the literature (see Wellman et al. 2001 for a meta-analysis of these), it quickly became clear that the experimental paradigm was afflicted with several confounding factors. A subset of these became known as pragmatic explanations of task failures, which most notably were focused on the explicit, elicited, or verbal components of the task. Generally, these were thought to make the task demanding in a way that would prevent children under the age of four from passing it, but for reasons that might have nothing to do with their mindreading capacities. In light of this, it was hypothesized that children under the age of four were in fact able to mindread, and several implicit, spontaneous, or non-verbal false belief tasks were developed that were intended to forego this problem.

One example of this can be found in Onishi and Baillargeon's (2005) 'expectation violation paradigm'.[55] First, there was a training phase of the experiment, in which the 15-month-old infants observed an experimenter playing with a toy and then placing it in one of two boxes; a yellow one or a green one. This was intended to accustom the infants to the setup. Next, the infants observed one of four conditions. In the one of the false-belief conditions, the toy moved independently from the yellow to the green box where the experimenter left it, or vice versa. The experimenter was not present during this condition. In the other, the experimenter watched the toy move from the yellow to the green box. After being occluded by a screen, the toy then moved independently back to the green box from the yellow box. Finally, at this point, in each of

[55] Interestingly, directly prior to these mindreading experiments, Baillargeon (2004) used this paradigm as a test for physical cognition. This in some ways reflects the type of shift on the level of the hypothesis without the shift on the level of the experiment that can be found in Hare et al. (2000) and Hare et al. (2001). This point will be returned to in Part 2 Section 2.2 of this dissertation.

these trials, the experimenter would either reach into the green or the yellow box. During this, the infant's looking time was recorded.

In these experiments, Onishi and Baillargeon found that the infant's looking time would be higher if the experimenter acted in a belief-discordant manner. That is, if the experimenter reached in the box that they should not have reached into given the knowledge that was available to them. Here, Onishi and Baillargeon assumed that the experimenter who acted in conflict with their belief would surprise the infant, and that a surprised infant would have a higher looking time. This was thought to indicate that the 15-month-old infants understood that others have beliefs that, "[…] may or may not mirror reality", (Onishi and Baillargeon 2005, pg. 257). That is, it was thought to indicate that they were able to mindread.

Like their explicit counterparts, these implicit experiments were likewise very influential and came in many variations (e.g., expectation violation paradigm, anticipatory looking task, etc.). However, also like their explicit counterparts, these tasks came with problems of their own, most of which were focused on a debate around the validity of the experiments, or as to whether successful completion of these tasks involved a theory of mind at all. While a meta-analysis performed by Barone et al. (2019) found that the results from the violation of expectation paradigm in particular to be robust, Kulke and Rakoczy (2018) performed a series of nine replications of Onishi and Baillargeon's (2005) implicit false belief tasks and found that four were successful, three of which were done internal to Baillargeon's lab, and the other performed by Dörrenberg et al. (2018) was only able to partially replicate the original task and found that, "infants can correctly recognize when someone's behavior does not match her false-belief, but not when that behavior does not match her true belief!" (Lavelle 2021, pg. 9). This created a situation in which replications attempts were difficult to interpret and a replication crisis in the research program was thought to be on the horizon.

Lavelle (2021) rightly draws attention to the fact that there is no consensus around the theory that is being tested in the implicit theory of mind research program.[56] Psychologists neither agree on how to precisely characterize belief attribution, expectation, or looking time, all of which are

[56] While this will not be argued here, it will be assumed that the same is also true of the explicit research program.

central features for interpreting evidence taken from the expectation violation paradigm. Importantly, this has direct implications for how the phenomenon of interest is operationalized. In highlighting this feature of the experimental paradigm, Lavelle references Collins' (1984) claim that methodological consensus must come together both with (1) consensus regarding the nature of the phenomenon in question as well as with (2) consensus regarding what constitutes a good test of that phenomenon.[57] Using this analysis, Lavelle argues that consensus on neither of these fronts has been achieved in the research program.

In doing so, she highlights the role of that 'fragile effects' play in obtaining such a consensus and entertains the claim that infants' ability to attribute false beliefs is a fragile effect and that Baillargeon's implicit knowledge played a significant role in realizing this effect. Lavelle then argues that fragile effects require the consideration of (1) the effect's scope and (2) its ecological validity. In her discussion of (1) scoping, Lavelle argues that an effect is fragile if the theory is able to state under what conditions it will arise, and under which it won't, or if the theory supports future scoping that would lead to this knowledge being obtained. If it doesn't do this then the effect is non-existant on Lavelle's account. In her discussion of (2) 'ecological validity', Lavelle argues that an effect is fragile if small changes to an experiment can interrupt it. This is because if the effect is highly sensistive to apparently small contextual factors in the lab, on Lavelle's account, there is reason to doubt that it will in fact manifest in the real world.

However, two different senses in which an effect might be seen as 'fragile' are blurred on Lavelle's account. This has direct implications for how such a claim should be interpreted. The first is **ontological fragility**. This can be seen in Lavelle's discussion of ecological validity. On this conception, an effect is understood to be delimited to a certain context and will only be realized under a certain set of conditions. However, in this sense, every cognitive effect is 'fragile.' This is because no cognitive capacity is unbounded or realized a-contextually. Every cognitive capacity requires delineating the contexts in which it will be able to be realized and doing so is part of what it will mean to understand and scope that capacity both ontogenetically and

[57] It is worth noting that it is not clear in what sense phenomena are able to be tested. Typically, hypotheses are tested which can produce evidence that is used to confirm theories regarding target phenomena.

phylogenetically. If this is the case, it is not clear what work the introduction of the term is doing. This sense of fragile should be dispensed of in so far as it is currently too vague to be of use.[58]

The second sense of the term is **epistemological fragility**. This can be seen in Lavelle's discussion of scoping. On this conception, the contexts that support the realization of the capacity are poorly understood. Fragility in this sense reflects the relative uncertainty with which the phenomenon of interest can be invoked. This will necessarily come in degrees and generally require further analysis.

In blurring these two senses of fragility, Lavelle goes on to make a stronger claim regarding the status of the effect than the evidence available to her can support. Namely, she argues that if false belief attribution breaks down with 'tweaks' to the experiment, then this provides support in favor of the hypothesis that in the real world, which is messy and not controlled in the way that an experimental context is, infants are not using a theory of mind to understand others' behaviors. This is a claim about the **ontological fragility** of the effect. That is, it is a claim about the nature of the causal structure of the effect that was exhibited in Onishi and Baillargeon's experiment; that is, what will allow it to manifest and what will block it from doing so. However, this claim is unsupported even on her own lights. Namely, it is in tension with the fact that her argument places an emphasis on there being a deep lack of a consensus regarding how to characterize the phenomenon of interest in a theory. At many points she is clear that this has strong implications for evaluating replications. This arises most pointedly in the summary claim where Lavelle (2021, pg. 22) writes," However, we do not yet know (a) whether VoE [violation of expectation] experiments reveal that infants can attribute false beliefs to others, or (b) whether VoE studies of IFB [implicit false belief tasks] replicate." This means that the precise explanatory and diagnostic roles of replications and how they can relate to validity and reliability assessments remain unclear on Lavelle's account.

[58] Ameliorating this sense of the concept would require indexing fragility to the number of contexts in which the capacity could be realized and then determining a way to compare these across capacities. However, given the current epistemological limitations in this context, it is unclear how this would be done.

This comes together to support the epistemological interpretation of fragility; the one that should be adopted regarding the violation of expectation infant false belief task research paradigm, and more broadly. However, in taking on the epistemological interpretation of fragility the ability to evaluate the status of infant false belief attribution in replication experiments needs to be given up. This applies to both failed as well as successful replication experiments. Lavelle is optimistic about the ability of replication failures to be used for theory development. While this might be the case, successful replications should inspire a symmetrical response, both of which are directed towards understanding causal structure of the effect tested in a target experiment.

The difficulties that Lavelle's account encounters do not end there. For example, the precondition set out for replications having any sort of evaluative import is that consensus to be formed around the theory that is being tested. But this is not exactly the issue at hand. Many research programs across psychology have seemingly established consensus regarding the theories they are testing, and yet they still struggle to progress on their own lights. The core problem is rather that they are underspecified, underdetermined, and admit too many interpretations. This can remain the case even when consensus appears to be present. There are many instances of this, particularly in smaller research programs in the psychological and cognitive sciences.

In addition to this, Lavelle (2021) holds that Bird's (2018) base-rate fallacy model of the replication crisis is the correct one. However, she contests Bird's claim that more replications will be able to resolve the crisis given the theory crisis that disciplines like developmental psychology are facing. Here, there are at least two problems with Lavelle's position.

The first involves a tension between two of her claims; (1) that there is a replication crisis regardless of the source and (2) that the status of theory in the developmental psychology is weak in a way that constitutes a crisis of its own. Lavelle, however, cannot have it both ways; there is either a clear theory crisis or a clear replication crisis. If both are claimed to exist simultaneously, then the replications that have been carried out, or the hypothetical replications that would be carried out, are not clearly diagnostic and are therefore not clearly valid replication experiments at all. If this is the case, then talk of a replication crisis, regardless of its source is ill-founded.

The second involves the role of replication experiments in disciplines like developmental psychology. Lavelle's argument here is on the right track. Replications alone are not sufficient to take on the concerns facing the psychological and cognitive sciences considering the emergence of the theory crisis. This critique, however, is limited. While more replications are neither necessary nor sufficient for dealing with the types of problems that have been created by the replication crisis, there is no *a-priori* reason that when conceived of as part of a larger strategy, that it would not be possible for them to play a productive albeit indirect role. Part of this will come down to what a replication is considered to be. However, Lavelle refrains from answering this question, and this sort of abstention creates further problems for the account. For example, in her discussion of experimenter skill (Collins 1985), Lavelle highlights the seemingly limitless number of variables that are involved in any target experiment and how this makes carrying out replications extremely difficult. This, however, is a missed opportunity to guide or direct research in disciplines like developmental psychology. The important aspects of the target experiment are the ones that are causally responsible for the effect in the experiment. Such an analysis can bring clarity to what counts as a well-formed experiment, how the phenomenon of interest is characterized, and what we should expect from both the target and replication experiments.

2.3 A Theory Crisis or a Replication Crisis; Not Both

Farrar and Ostojic (2019) have argued that comparative cognitive science displays multiple signs of degeneration. Given this, they claim that the science is operating broadly under an illusion of scientific rigor. In making their case, they lay out the following six issues that they see in the discipline, which have blocked it from developing (Farrar and Ostojic 2019; pg. 1):

i) The field is biased towards confirming more exceptional abilities in animals.
ii) There is likely to be a high rate of false positive discovery.
iii) There is a persistence bias towards studying more exceptional abilities, even in the presence of strong methodological criticism.
iv) There is an absence of a formal method to assess evidence of absence of a cognitive ability.
v) There is ambiguity in definitions used to make claims.

vi) The field of research is small.

While there are several arguments here that could be addressed, in particular the authors' view of scientific progress, the discussion will be limited to a particular line of critique that is focused on the relationship between status and use of theory and replication in the discipline. Central to Farrar and Ostojic's argument is the idea that comparative cognition displays all the markers of a discipline that is on the verge of a replication crisis. Their argument is supported by referencing the emergence of the replication crisis in adjacent disciplines as well as in other parts of psychology that have very similar profiles (i.e., ones that have small sample sizes and noisy measurements), which are also theorized to have a high rate of false positives.

One factor that might set comparative cognitive science apart, which is highlighted by Farrar and Ostojic, is the discipline's tendency to favor within-subject designs over between-subject designs, although this is not obvious. One might think that employing within-subject designs, will be able to remove noise between individual groups or subjects (Stevens 2017, Open Science Collaboration 2015). Typically, within-subject designs are employed in research scenarios that require a small sample size and are seen as advantageous because of their ability to wash out more stable aspects of individual variation that one might encounter in a between-subject design. However, the virtues of within-subject design are based on several premises. First, that the effects that are being tested in comparative cognition are not time-sensitive and in no way carry over. And second, it needs to be assumed that the within-subject samples are representative. Given the very particular life histories and even genetic profiles of the animals that are used in many comparative studies, there is good reason to think they might not be. This could very well be reason to think that replicability might be a problem for the discipline. That is, given broader considerations about the status of research in the discipline, there is no reason to think that within-subject designs will provide shelter for the concerns around the replication crisis that have affected other disciplines that employ between-subject designs.

Despite this moment of optimism about the use of within-subject designs in comparative cognitive science, Farrar and Ostojic are broadly pessimistic and contend that although there is not currently a replication crisis in the discipline, there is good reason to think that one is coming.

Part of their reason for thinking this is based on the claim that, in addition to the impending replication crisis, comparative cognition is also undergoing a theory crisis. They think this for the following two reasons.

First, they argue that theory in comparative cognition is definitionally ambiguous and that this creates a situation in which claims can be flexibly adjusted. According to Farrar and Ostojic, part of the problem is that the discipline routinely makes use of verbal theories and that there is often a significant gap between the statistical test that is done and the substantive claim that is made on the basis of it (Meehl 1990). This is a significant problem because the same data can be used to support a variety of theories. That is, the task of theory selection is underdetermined by the data.

On Farrar and Ostojic's diagnosis of the discipline then, both the replication crisis and the theory crisis combine, together with other problems they see in the discipline, to support the claim that the methodological and theoretical status of the discipline at large is one in which an "illusion of science" is being carried out in the discipline.[61] However, there is a clear tension in their view that should be familiar from the discussion of Lavelle (2021) above: There is either a genuine replication crisis in which the general status of research in comparative psychology is able to be shown to be unreliable **or** the status of theory in the discipline is inadequate in a way that would make it difficult, if not impossible, to assess whether the research in the discipline is reliable. Farrar and Ostojic cannot have it both ways. The problem arises as a confusion between uncertainty and unreliability. To illustrate this, briefly consider the mirror self-recognition research program, which Farrar and Ostojic hold up as an example of a case in which evidence of the absence of a capacity is able to be obtained.

The first non-human animal self-recognition studies were performed by Gallup (1970). In these, he initially attempted to study this capacity by placing mirrors in an enclosure with the chimpanzees in his lab for 10 days. After this period of exposure, his chimpanzees spontaneously displayed an increase in behaviors that were purportedly self-directed and were meditated by the

[61] There is reason to think that comparative cognitive science is probably better off thought of as a nascent science that is in the process of development. Even if a number of the arguments that Farrar and Ostojic put forward are accurate, there is no a priori reason to think that the discipline is off course in this respect or that this is indicative of anything other than the normal developments that occur in scientific practice.

mirrors. Two of these behaviors were characterized as grooming and self-observation. This provided the motivation for an experiment in which, after putting the chimpanzees under general anesthesia, he marked an ear and the ridge of their eyebrows with a red dye. Upon regaining consciousness, a subset of the chimpanzees was then placed in front of a mirror. The chimpanzees that were exposed to a mirror responded to the marks significantly more than the ones that were not. Rather than attempting to examine the strange marks in the reflection of the mirror, the chimpanzees examined the marks on their body that were being reflected in the mirror. On the basis of this, it was claimed that they were able to figure out that they were the source of the image in the mirror. From this result, and the initial spontaneous behaviors that were observed prior to the test, Gallup argued that chimpanzees possessed the capacity of self-recognition.

Gallup's chimpanzees are assumed to be intentional systems and attributing self-recognition to them seems to support certain inductive inferences even if it is not exactly clear which ones. The hypothesis seems to be that if the chimpanzees did not have a concept of self that they were capable of recognizing in the mirror, then they would continue responding socially to the image they encountered. In many contrast cases, such as those involving cats, the social responses are aggressive ones. In addition to this, self-recognition is conceived in intuitive folk psychological terms, and a test for a self-concept, which is also intuitively conceived, is conceptually derived from the task in the experimental setup (Gallup 1970, pg. 87).

Since these original tests, the mirror test has been performed on a variety of species including dogs, elephants, dolphins, magpies, horses, manta rays, squid, and ants, among others.[62] While members of all these species have passed some version of the mirror test, the claims in these experiments have been disputed for a variety of reasons. Some of these have to do with the specifics of the marking medium and its relationship to the physiology of the animal being tested. For example, Rajala et al. (2010), surgically implanted a block into their monkeys' heads after they failed the more conventional mark test. However, due to the highly invasive nature of the block, it appears to have severely confounded the experiments.[63] More recently, upon reviewing

[62] This is a common strategy in comparative cognition. Rather than probing a cognitive capacity more deeply in an individual species, tests are generalized and applied to other species as tests of the capacity in them.
[63] It is also very plausibly that these experiments were severely unethical.

evidence from a variety of species, Gallup and Anderson (2017) claimed that great apes are the only primates that are able to pass the mirror test in a way that is indicative of the presence of a theory of mind. However, even this conclusion has been challenged because of the relationship between the task and the capacity that is being tested for (e.g., see Heyes 1994).

In this way, further experiments have seemed to be unable to clearly demonstrate evidence of self-recognition in a way that would lead either to the further understanding of the conception of self-recognition or how the sophisticated cognitive capacity is distributed. While most of the challenges have had to do with the underdetermined nature of experiments that have been carried out, and the fact that the data yielded from these experiments are open to many interpretations, other challenges have centered on questioning the exact capacity that is being tested for. For example, visual self-recognition (Gallup and Anderson 2017) is potentially relevantly distinct from mirror self-recognition (Krachun et al. 2019), both of which are distinct from self-recognition full stop, and even how to think of the self-concept and self-recognition have also been highly disputed (Mitchell 2013, Boyle 2018). In addition to this, the very concept of self that is underlying the capacity of self-recognition has been challenged, and Heyes (1994, 1995) has argued that Gallup's mirror studies provide no convincing evidence of great apes, other than humans, being able to use mirrors to gather information about their own bodies. More pointedly, she argues that these experiments provide no evidence of the presence of self-recognition. This is due to the number of confounding factors included in the experiments that have not been controlled for such as the anesthesia that was used on Gallup's original chimpanzees, or a lack of conceptual criteria for what counts as a self-recognition-based interaction with the marks, among many other factors.

This is all to say that the status of theory in the non-human animal self-recognition research program is troubled at best. Returning to Gallup's (1970) seminal experiment, what would need to be known to carry out a replication of it? Coarsely, according to the causal accounts of replication presented above, an understanding of the causal structre of the effect in the experiment and how it is sensitive to confounders would need to be had in order for a replication study that is clearly diagnostic to be obtained. There is no reason to think that anything like this is present in the self-recognition research program. In fact, the theory is not even stated in terms that resemble

anything like what would be required. This creates a situation in which if replications were to be performed, the diagnostic properties of these replications in the very best of circumstances would be highly uncertain, which means that in any situation in which a replication is performed, there will always be a dispute about how to interpret it.

But if this is the case, and if this situation is representative of the status of research in comparative cognitive sciences conceived broadly, how can the discipline ever possibly be subject to a genuine replication crisis in the near future? The answer is that it cannot. This is not to say that an **apparent** crisis cannot take place; one in which there is a broad sense of panic about the status of research in the discipline. There is even reason to think that an apparent replication crisis is already underway.[64] However, given the current status of research in comparative cognition, the discipline cannot possibly undergo a **genuine** replication crisis. This is a problem that can only possibly affect the discipline once it has undergone further theoretical development.

Building on the account touched upon above, a discipline is facing a 'genuine replication crisis' to the degree to which replication failures are able to justifiably draw the status of the knowledge in that discipline into question. The degree to which this occurs will then in turn serve as justification for the consideration or implementation of reforms that can be made in response to it. There is a lot at stake here. Over the past 10-15 years, replication failures have resulted in virtually every aspect of the sciences that have been affected by these failures being placed up for debate. For example, it has been claimed that the replication crisis demonstrates the presence of wide-spread fraud and/or bias, and some have held that the fundamental methods, including the statistical methods, that are used in disciplines like social psychology need to be completely overhauled.

While there is nothing wrong with reconsidering the way that any given science is working, these types of proposals are often made from a place panic. After all, if most of the scientific knowledge in various disciplines had been shown to be wrong, a rational response to this would be to start making the necessary reforms as soon as possible. However, when an apparent replication crisis is

[64] For evidence that this is in fact the case, see Farrar et al. (2021) regarding contemporary attitudes of comparative cognitive scientists regarding the reliability of their work and the work of others in the discipline.

mistaken for a genuine one in the sense described above, this can be significantly harmful both to individual researchers who are unjustly framed as being fraudulent or incompetent, as well as to entire disciplines which are written off as engaging in 'pseudo-science.'

What is at stake in the claim that comparative cognition cannot face a genuine replication is exactly this. Assuming the presence of a theory crisis, replication failures in comparative cognitive science are not able to justifiably draw the status of the knowledge in the discipline broadly into question in a way that would justify the brash adoption of reforms. Moreoever, replication failures are neither clearly indicative of the presence of fraud or incompetence, or the illegitimacy of the science writ large. Given the status of theory in the discipline, a genuine replication crisis and the many proposals that are likely to fall out of it would be currently unjustified.

2.4 Multiple Crises and their Interactions

In this section, three attempts to analyze the replication crisis have been analyzed. This has been done both in a general sense and with respect to particular disciplines (see Table 8 below for a diagram of these).

EVALUATING THE REPLICATION CRISIS

Author	**Scope**	**Difficulties**
Bird (2020)	Disciplines broadly subject to the replication crisis	• Account of the source of the crisis not evidentially supported • Performing more replications absent theory development will not solve the crisis • Coexisting theory and replication crises
Lavelle (2021)	Developmental psychology	• Flat conception of fragility • Limited conception of what replications are and their utility • Coexisting theory and replication crises
Farrar and Ostojic (2019)	Comparative cognitive science	• Excessive pessimism regarding the broad state of comparative cognition

		• Misleading assumptions about the science • Coexisting theory and replication crises

Table 8. A depiction of three attempts at evaluating replication crises in varying contexts that have been discussed in this section.

Although these accounts are distinct in scope and conclusion, they share a common thread; attempts to both evaluate the replication crisis as well as give a proposal for resolving it need to be sensitive to the state of the theory in the target research program and even discipline. As has been argued in the discussion of all three of these accounts, if there is a theory crisis in the research context that is being evaluated, one cannot simultaneously claim that there is also a replication crisis. This point, once again, trades on a distinction between genuine and apparent replication crises, with a genuine crisis only being possible in the absence of a theory crisis. While these analyses have focused primarily on the possibility of evaluating the replication crisis, the next section will focus more directly on the prospects of evaluating replications themselves, given that such evaluations situated in epistemically various ways. There the positive deflationary account of replication will be more explicitly introduced and developed.

3. There is No Special Problem of Replication

This section will give a preview of the positive 'deflationary' account of replication that will be gradually developed. Thus far, it has been argued that the existing accounts of replication (naïve, resampling, diagnostic, casual, and typological) all face significant limitations. These limitations not only have direct consequences for evaluating individual experiments, research programs, and the replication crisis broadly construed, but they also have direct consequences for demarcating what scientific research should be held to the replicability criterion. In the remainder of this part of the dissertation, the question 'what is a replication?', particularly when performed under non-ideal conditions, and/or those involving high degrees of uncertainty, will be answered; a deflationary account of replication will be introduced and defended. The argument has four key steps, which are briefly summarized below in Table 9.

First, the central question, 'what is a replication?', is deflated (i.e., deliberately rendered trivial). On the 'deflationary account', a replication is an experiment that is stipulated by an epistemic agent to be a test of a hypothesis regarding the causal structure of the effect tested in a target experiment. The question upon which the utility of the practice hangs is reframed to become the following one: Is the replication experiment a valid test of the causal structure of the effect tested in the target experiment? Here, the concept of validity is discussed and the advantages of the 'deflationary account' over a selection of the accounts that have been discussed in this part of the dissertation are defended (in particular Part 1 Section 1).

Second, given that a replication experiment's validity is pivotal on the deflationary account, by making use of an analysis of validity given by Feest (2020), it is argued that if first-order experimental validity comes in degrees, then the validity of replication experiments must as well.

Third, the deflationary account provides the grounds for rejecting the widely held idea that replications are about reliability and not validity. Background knowledge of the relevant effect secures the assessment of the validity of the replication experiment as well as its target. In this way, replication experiments cannot assess the reliability of their targets without first being able to assess their validity.

Fourth, given that the validity of replication experiments is graded, a framework for evaluating them it is introduced that is centered around the following three questions: (1) With what degree of certainty is the phenomenon under investigation characterized and individuated? (2) With what degree of certainty is the investigation itself characterized and individuated? (3) What aspect of the target experiment is the replication experiment intended to evaluate? A representative example of a recent replication study from comparative cognitive science is run through this structure and the results are used to support the broader claim that replications in comparative cognitive science are not clearly diagnostic, and therefore any talk of a replication crisis is premature.

A SUMMARY OF THE KEY PROPOSALS OF THE DEFLATIONARY ACCOUNT OF REPLICATION	
1. What is a replication? (Section 3.1)	An experiment that is stipulated by an epistemic agent to be a test of a hypothesis regarding the causal structure of the effect tested in a target experiment.
2. What makes a replication experiment valid? (Section 3.1)	If it is in fact a test of a hypothesis regarding the causal structure hypothesized in the target experiment.
3. Does the validity of a replication experiment and its success come in degrees? (Section 3.2)	Yes, if experimental validity comes in degrees, then the validity of replication experiments come in degrees.
4. Do replications assess validity? (Section 4)	Yes, background knowledge of the causal structure of an effect secures the assessment of the validity, both in the replication experiment as well as the target.
5. If the validity of a replication experiment is graded, how should it be assessed? (Section 5)	With the following questions: • With what degree of certainty is the phenomenon under investigation characterized and individuated? • With what degree of certainty is the investigation itself characterized and individuated? • What aspect of the target is the replication intended to validate?

Table 9. A summary of the key takeaway proposals from the deflationary account of replication that are defended below.

3.1 Deflating the Central Question

In the most abstract sense, an experiment is a test of a hypothesis; typically concerning the causal structure of a phenomenon. To initially keep things simple, it can do this in a way that is valid or invalid. Its validity, however, does nothing to affect its status as experiment. That is why the following sentence in no way seems paradoxical: *That experiment is invalid*. In this way, what counts as an experiment can be thought of as depending primarily on a stipulation made by an epistemic agent. As long as experiments are stipulated by an epistemic agent, in the broadest sense,

as being a test of a hypothesis, they count as experiments. Little is at stake in this demarcation, and in many ways this definition is a rejection of the project of demarcating experiments in any sort of deep way.[65] Experiments can be invalid, have nothing to do with the causal structure of the target phenomenon, be completely misguided, and a waste of time and/or other resources, and still count as experiments. There is no motivated reason to think otherwise.

In exactly the same way, a replication is an experiment that is a test of a hypothesis; typically, about the causal structure of a target experiment. It can likewise do this in a way that is valid or invalid or even an unequivocal waste of time, and it should still be called a replication. Just as an experiment that is invalid is still an experiment, a replication experiment that is invalid is still a replication.

So, what constitutes the relation between a replication and the target? This is what will be referred to as the 'constitution question.' The deflationary account's answer to it is the following: *stipulation by an epistemic agent as being a test of the causal structure of the effect tested in the target experiment.* Questions about the utility of these experiments are not questions about their status as replications *per se*. Taking this account on board entails a somewhat substantial shift in the way that replications are typically talked about or debated, but it is a helpful one in so far as it will ultimately aid in the projects of both evaluating the results of replication experiments and determining precisely what a replication is. Rather than debating whether an experiment is a replication or not, which is on this account is essentially the 'wrong' question to be asking, focus and energy should instead be place on determining whether the replication experiment is valid. Doing so allows experiments and replications to be treated equivalently. This approach also opens a number of familiar conceptual tools that are already employed to evaluate first-order experiments, that can be taken up in their current form to evaluate replication experiments. In addition to this, it yields a better definition of replicability. That is, an experiment is replicable if a valid replication experiment can be carried out on it.[66] This definition demystifies the causes of replicability and allows for the clear identification of the factors that enable it.

[65] It does seem plausible that this approach to defining experiments could provide a way of separating them out, albeit coarsely, from things like observations as well as exploratory research, which are generally not directed at testing hypotheses.
[66] One could also just entirely reject the way in which the question is formulated.

Moreover, the deflationary account brings out the sense in which there is a tripartite distinction to be made between theories of replication (see Table 10 below), following a similar one that has been made between scientific theories of explanation by Solomon (1989). This way of framing theories of replication has existed implicitly in the literature but has not been explicitly addressed.

A TRIPARTITE DISTINCTION BETWEEN THEORIES OF REPLICATION	
Ontic theories of replication	A theory R of replication is **ontic** *iff* according to R, there exists no statement of the form 'x is a replication of y' that is true relative to a person S_1 and not true relative to another person S_2.
Epistemic theories of replication	A theory R of replication is **epistemic** *iff* according to R, there exists a statement of the form 'x is a replication of y' that is true relative to a body of knowledge K_1 and not true relative to another body of knowledge K_2
Pragmatic theories of replication	A theory R of replication is **pragmatic** *iff* according to R, there exists a statement of the form 'x is a replication of y' that is true relative to a person S_1 and not true relative to another person S_2.

Table 10. A tripartite distinction between theories of replication.

For some, these distinctions might seem to throw a wrench into the argument as it has been carried out up until this point. The accounts of replication that have been addressed so far were all evaluated against 'ontic' criteria. Namely, it has repeatedly been argued that facts about what an epistemic agent stipulates or facts about what an experiment was able to epistemically demonstrate, were both unable to address the question of whether an experiment is a replication. Instead, many of these accounts failed to attempt to probe the ontic facts concerning the causal structure of the effect being investigated in the target experiment. This was the cause of their shortcoming.

It might seem curious or surprising then, that the deflationary account is an explicitly pragmatic theory of replication in so far as it turns on stipulations made by an epistemic agent rather than ontic or epistemic facts. To make this more explicit, on a pragmatic theory of replication, of which the deflationary account is an instance, there are cases in which a statement of the form 'x is a replication of y' is true relative to a person S_1 and not true relative to another S_2. Perhaps, S_1 stipulated that x is a replication of y and S_2 did not. The problem seems to be that this creates floating criteria of evaluation on which previous accounts were evaluated against ontic standards while the current account is evaluated against pragmatic standards. The call would then be for equal criteria for evaluation to be applied across the board, and were this to occur, then the deflationary account would fail in so far as it has nothing to do with ontic criteria.

This is only an 'apparent' problem. The deflationary account has a clearly pragmatic approach to the question of what counts as a replication; again, stipulation by an epistemic agent. However, in adopting the strategy of treating replications as experiments as has been argued, this question should not be given much energy. Instead, focus should be redirected to the question of whether a replication experiment is a valid one. In addressing this question, ontic facts are front and center. This allows the worry that the evaluative criteria are somehow floating in a way that unduly favors the deflationary account to be set aside.

This, however, is not where the apparent obstacles end. In so far as the deflationary account is a pragmatic theory of replication, one might worry about 'the problem of unintentional replications.' That is, if what constitutes a replication is the stipulation by an epistemic agent, can an experiment X be a replication of experiment Y if it has not been stipulated to be so.[67] On the deflationary account, the answer is no. But if that is the case, how can situations be accommodated in which scientists unwittingly perform the same experiment and obtain the same results (e.g., see Merton 1963)? Multiple independent discoveries provide a salient case in which one would be inclined to view a series of experiments as replications of one another, even though they were not intended to be so by the original experimenters.

[67] A first-order example of this can be found in natural experiments approach to economics. Angrist, Imbens and Card received the Sveriges Riksbank Prize in Economic Sciences in 2021 for their contribution to the development of this approach.

This problem is not as threatening as it might seem to be. Again, the deflationary account of replication is squarely pragmatic in character. However, a stipulation is not a baptism. There is no reason to think that once a stipulation is made, that it is fixed for all time in the way that something like a rigid designator is (see Kripke 1980). This means that even though one epistemic agent did not stipulate an experiment to be a replication of another, there is no reason to think another epistemic agent is not able to do so after the fact and then subsequently ask whether that replication is a valid test of the causal structure of the effect in the target experiment. This applies to ordinary experiments in the same way that it applies to replication experiments. That is, experiments can be valid tests of hypotheses even if they were not stipulated to be so.

In this way, the account provides a way to address what Gelman (2020) in the context of a discussion of the replication crisis has called a 'temporal bias'; the idea that because one experiment X was done prior to another Y, there is tendency to claim that X cannot be a replication of Y, because replications are commonly thought to necessarily occur subsequently to their target experiment. As Gelman argues this is a systematic error; a bias. In addition to this, Peels and Bouter (2021) argue that there is no a-priori reason to take an intentional replication experiment more seriously than unintentional replication experiment, assuming it is set up correctly. However, what the deflationary account contributes to such a view are clear criteria for what being 'set up correctly' will mean in any given context. And again, while what counts as a replication shifts in concurrence with the stipulations of the epistemic agent, the criteria for determining whether a replication is valid are not so liberal.

Yet another objection that might be made to the deflationary account, is that is that it simply evades the question. That is, in answering the constitution question with an appeal to stipulations made by epistemic agents, the problem of how replication experiments epistemically function is simply avoided. Given the account's parallels with Callender and Cohen's (2006) take on scientific representation, and the similar criticisms that it has faced, this issue is worth addressing head-on.[68]

[68] While Callender and Cohen's (2006) paper is implicitly addressed here, no position will be taken with regards to its plausibility as an account of scientific representation. Moreover, there is good reason to think that scientific

However, the answer to this objection is essentially the same as the answer to the first. If the deflationary account merely consisted of a stipulative criterion, then this concern might be warranted. However, as has been repeatedly emphasized in this section, this is in no way what the account amounts to. Instead, the deflationary account is intended to deflect attention away from the 'constitution question' and place focus on the 'validity question'; i.e., the question of whether a replication experiment is valid. Doing so places inquiry in highly familiar territory. Core to this move is the claim that what allows replications to fulfill their epistemic function is their validity as experiments. In this way, rather than refusing to take on the task at hand, the deflationary account provides a much-needed clarification thereof. By laying bare the determinates of a replication experiment's epistemic utility the practice can be enhanced. This objection is therefore unwarranted.

The basic features of this account are on the table and some of the initial concerns have been addressed. Now, a more detailed analysis at what makes experiments valid can be given. However, the move towards emphasizing the way in which validity is central to replication experiments again encounters more stumbling blocks than might initially be expected. This is because conceptions of validity are numerous.[69] This abundance of accounts has come about despite attempts on the part of several scientific institutions to establish standards that could be agreed upon and referred to. One example of how convoluted the landscape has become can be found in Newton and Shaw (2013) who highlight the over 151 types of validity that have been discussed in the psychological literature alone, many of which are clashing and overlapping (121 of them are directed at measurement validity). As in the case of the replication literature, the fracturing of validity into types with distinct labels has undoubtedly been deeply confounding.

While concerns around what would be eventually be referred to as validity go back to the late nineteenth and early twentieth century if not much earlier, Buckingham (1921) notably attempted to make the concept more precise and in doing so focused in rather narrowly on the predictive

representation and replication are significantly distinct objects of inquiry. In this way, a valid criticism of one will not automatically entail a valid criticism of the other.

[69] In many ways conceptions of validity mirror the proliferation of conceptions of replication.

capacity of a given test. This, however, changed with Cronbach and Meehl (1955), whose wider conception of validity was rooted in the discipline of psychology and broke it up into four types; predictive validity, concurrent validity, content validity, and construct validity (see Table 11 below).

CRONBACH AND MEEHL'S FOUR TYPES OF VALIDITY	
1. Predictive validity	A measurement's ability to predict a future result.
2. Concurrent validity	The correspondence between a novel and a validated measurement.
3. Content validity	A measurement's ability to cover a construct.
4. Construct validity	A test's ability to measure the concept it is intended to.

Table 11. An abbreviated list of types of validity from Cronbach and Meehl.

While the physical sciences seem to have converged on ways to calibrate their measurements, there is no internationally recognized standard for psychological measurement that resembles anything like kilograms, meters, and seconds. This poses challenges for measuring the target of inquiry in the sciences of the mind because cognitive capacities such as declarative memory or level-one perspective taking are theories, or constructs in Cronbach and Meehl's language, that must be measured via proxies; in comparative cognitive science, typically via behavior. Given this, Cronbach and Meehl proposed that a measure of one construct, must relate to the measure of another construct in a predictable way (sometimes also called convergent validity). In doing so, it was thought that a construct could be examined as part of what they referred to as a 'nomological net.' This coherentist style approach to the measurement of unobservable entities yields a high degree of underdetermination in cognitive tests. Given this, constructs and measurements are ideally incrementally adjusted and improved upon throughout an iterative and holistic process of experimentation.[70] This is because for any given situation (1) a construct/theory could be wrong relative to the phenomenon, (2) the measurement of it could be wrong relative to the phenomenon, or (3) both could be wrong. These questions remain relevant even if a hypothesis

[70] See Poston (2021) for an account of coherentism more broadly and some limitations that it might encounter.

has been supported by the results of a particular experiment. While concerns are ideally reduced through a process of iteration, the looming threat of underdetermination never truly exits the scene. Importantly, Cronbach and Meehl's approach signals a method for approaching the type of phenomena that are typically studied in the sciences of the mind.

The debate around validity in psychology has largely been centered on psychometrics, and at times has veered in the question of whether the scope of validity should include the interpretation and use of a test. Many concerns around validity are couched within certain disciplines such as medicine or are local to particular research programs like intelligence testing. These typically do not capture a broader conception of validity that can span across experimentation in a variety of disciplines, including the physical sciences such as chemistry and physics. Given this, the focus here will be placed on a more wide-ranging form of validity following Kelley (1927; pg. 14) who writes, "The problem of validity is that of whether a test really measures what it purports to measure […]." Building on this, Borsboom et al. (2004, pg. 1061) write, "[…] a test is valid for measuring an attribute *iff* (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure." The second part of Borsboom et al.'s definition (b) functions as a more detailed account of what it means to really measure what is being intended to measure, and it is plausible that having a causal understanding of the measurement should in many instances allow for the intervention, manipulation, and control of the phenomenon in question.

However, following Cronbach and Meehl's distinctions, measurement is typically an aspect of an experiment and not identical to it, and the use of the term 'test' by both authors is somewhat ambiguous. For this reason, the following distinction is perhaps an important one to make. **First**, a measurement is a valid one *iff* it in fact measures what it purports to measure. And **second**, an experiment is a valid one iff it tests the hypothesis that it purports to test.[71] Going forward, these will be the senses of experimental validity and measurements validity that will be adopted, particularly in the context of determining what makes a replication experiment valid.

[71] There is clearly a lot more to be said about what makes an experiment a valid test of a hypothesis, however, the details of such an account are orthogonal to purposed at hand.

With this conception of validity now on the table, the following question can be posed: Does the replication experiment test the hypothesis about the causal structure of the effect in the target experiment? That is, is it valid? Again, the approach entails treating the investigation of target experiments just like any other phenomenon being investigated.

Given this, the deflationary account of replication holds at least four initial advantages over the accounts discussed in the previous sections. These are as follows:

(1) It hinges on the use of existing conceptual resources; enhances the practice
(2) It explains why some replication experiments are indeterminate; enhances the understanding of the practice
(3) It is heuristically parsimonious; enhances the understanding of the practice
(4) It requires holistic evaluations; enhances the practice

The first advantage of the deflationary account holds (1), lies in the fact that it makes use of the existing conceptual resources that are used to formulate, perform, and evaluate ordinary experiments. This clarifies the practice of carrying out replication experiments by reducing the space of conceptual features that require explanation. Take a now canonical example of a research program that faced replication failures under consideration: the ego depletion effect. The theory holds that self-control draws on, and has the potential to drain, on our limited cognitive resources. In a study done by Baumeister et al. (1998), they tested and confirmed the hypothesis that cognitive load either decreased their subjects' performance on puzzles, or their willingness to attempt to solve them. In a multi-lab replication of these experiments done by Hagger et al. (2016), this effect was not able to be recreated, thereby resulting in a purported replication failure. This led to claims that the ego-depletion did not exist (Osgood 2017).

However, this replication experiment's failure is not the end of the story. As Stroebe and Strack (2014; pg. 59) write, what replication experiments are able to say about their targets will generally be underdetermined because,"[…] studies conducted at different times and with different subject populations might test different theoretical constructs." This is not to say that HARKing, cherry-picking, P-hacking, fishing, and/or data dredging or mining could not have occurred in these

experiments. Questionable research practices, or QRPs (unintentional but institutionalized), and fraud (intentional) could be behind Baumeister et al.'s claim. However, a failed replication experiment of type investigated by Hagger et al. will be unable to determine that.

Given that the precise causal structure of the ego-depletion effect as studied by Baumeister is ultimately mysterious, the replication experiments as they were attempted by Haggar et al., various types of similarity, methods for decomposing experiments as seen in the resampling account (Machery 2021), and other novel epistemic technologies, are used as guides. However, these guides are unreliable and can unwittingly render a replication experiment undiagnostic in so far as they do not necessarily direct the experimenter towards the causal structure of the effect in the target experiment (i.e., in this case the ego-depletion effect). Substantial background knowledge about the effect would be required to be able to do so. This is simply out of reach in many research programs like this one, and many others that have been affected by the replication crisis. Experimenters are in the dark with regards to which aspects of the target experiment are essential to producing the effect and which are superfluous. For these reasons Hagger et al.'s meta-analysis is also undiagnostic, in so far as it relies on data from replication experiments that are undiagnostic. Perhaps more importantly this necessarily places heavy restrictions on any substantive claim that Baumeister et al. had about ego-depletion theory in the first place and even had the replication experiments been successful, any substantive claims about the effect would need to come with serious and substantial caveats.

Instead of relying on unconstrained and unreliable 'guides' like similarity or relevance that have the potential of distorting inquiry if they latch onto the wrong features of an experiment, focus should be redirected towards determining how to perform a valid experiment that is directed at the causal structure of the effect of interest in the target experiment. In the context of Baumeister et al.'s (1998) study on ego-depletion this means performing an experiment that targets the purported causal structure of the effect of ego-depletion in that study. This constitutes an advantageous shift because it renders the task exactly the same as it is when scientists aim to achieve validity in any other experimental context. Of course, this is not to say that valid experiments are easy or in any way straightforward to carry out, particularly in the sciences of the mind. A task being clear is importantly distinct from it being easy. In particular, establishing valid

experiments has proven to be extremely challenging in comparative cognitive science. However, this approach deflates the practice of replication into a familiar and tractable one; one that scientists already have the tools to be able to take on. There is no need to appeal to a host of other 'novel' epistemic tools like those seen in the resampling and the diagnostic accounts.

This brings out a second advantage that the deflationary account holds (2); namely it offers an explanation as to why determinately successful/failed replications require substantial background knowledge about the causal structure responsible for experimental results. This also demonstrates why the possibility of a replication crisis is out of reach for research programs and disciplines that lack such knowledge. For example, imagine a chemistry experiment on the effect of a catalyst on the rate of a chemical reaction. In the target study, it was claimed that adding a certain catalyst to the reaction increased the rate at which the reaction occurred. In this study, the scientists propose a mechanism for the catalyzed reaction, suggesting that the catalyst facilitated the formation of an intermediate compound that was more reactive than the starting materials. The causal structure of the effect in question, in addition to the substantive background knowledge in which it is entrenched, have direct implications for what a replication experiment can target in this type of experiment in this particular discipline. This means that when scientists, here chemists, go to carry out a replication experiment, given the substantial background knowledge about the causal structure of the effect that is present in this context, its validity can be assessed with a high degree of certainty, when compared to other more impoverished epistemic contexts. Of course, uncertainty will remain in any empirical context. However, the replication experiment in this abstract case is plasubily able to be diagnostic of the mechanism of the catalyzed reaction that is tested in the target experiment.[72]

Contrast this with another now canonical experiment in the replication crisis literature on the so-called facial feedback effect (Strack et al. 1988); the hypothesis that facial expressions have an effect of affective response. In one of the experiments, for example, the subjects were instructed to hold a pen between their teeth which caused them to smile. This was claimed to cause them to view the cartoons as funnier than if they were not smiling. The facial feedback effect is a theory

[72] Again, all empirical inquiry is subject to uncertainty and underdetermination. The point of contrasting this case with one is psychology is simply to highlight the degree to which the epistemic situations differ.

that had gained support from multiple independent studies. However, Wagenmakers et al.'s (2016) pre-registered replication or Strack et al.'s (1988) experiment, among others, failed to replicate. This is a context, however, in which the success of failure of a replication experiment is highly indeterminate. This is primarily because substantive background knowledge regarding the causal structure of the effect in the target experiment is absent in this context. The replication experiments in this case are therefore highly undiagnostic of the claimed effect in the target experiment. This would moreover remain the case even if the experiments were found to replicate; that is, if they were successful.

A third advantage that the deflationary account holds (3), has to do with a kind of heuristic parsimony that it provides in the face of proliferating typological accounts. By striking a measured balance between a monism and pluralism, the deflationary account makes sense of the wide diversity of cases that fall under the practice of replication while simultaneously unifying them under a single explanation. Here, replication experiments are 'one thing' and by binding what counts as a replication to the stipulation of an epistemic agent, the account can wholly bypass some of the more entrenched, intractable, and progress inhibiting debates regarding the difference between direct and conceptual replications, or the question of what a replication is simpliciter.[73] There is no incentive to invoke a typology to make sense of the practice in a general sense.

A fourth advantage of the deflationary account (4) can be found in the fact that it requires holistic evaluation. The abandonment of holistic evaluation was a significant problem for Machery's (2021) resampling account as discussed above. However, the deflationary account centers holistic evaluation in two senses. The first has to do with the holistic evaluation of individual replication experiments being anti-proceduralist and necessarily conceived of as being theoretically laden. The second has to do with the evaluation of evidence from across replication experiments and with a larger context of background knowledge. However, just as any experiment can target a particular aspect of the phenomenon in question, a replication experiment can, but need not necessarily, target a particular aspect of the relevant phenomenon (here the target experiment). This makes it continuous with how most other scientific phenomenon are studied and can be one

[73] In addition to this, validity is also 'one thing.' It can be established in a variety of ways across a variety of contexts; however, it is unified under one heading or goal.

of the primary resources for identifying cases of HARKing, cherry-picking, P-hacking, fishing, and/or data dredging or mining. Although there are a variety of ways of parsing up an experiment for analysis such as Brunswik's (1955) proposal (participants, situations, operationalizations, and time points), how this should precisely occur will be determined by the causal structure of the effect in the target experiment; no formula will be able to determine the target. In this way, validating distinct aspects of an experiment independently should not be substituted for a holistic approach to the phenomenon. In many cases, this will require aggregating evidence from multiple experiments on the target experiment, which should be carried out in accordance with the 'principle of total evidence' (Carnap 1947). In this way, the account resists one-shot evaluations of experiments which were common during the emergence of the replication crisis. This follows from the assumption that the study of any given scientific phenomenon should not be subject to a so-called 'crucial experiment' alone (Bacon 1620). Admittedly, this involves a shift in how replications are typically carried out and evaluated. However, by framing replication as a series of experiments that in many cases that probe different aspects of the same phenomenon, even while targeting an experiment holistically, a stronger evidential basis for the evaluation of a target experiment can be established.

3.2 A Replication's Degree of Validity

Replication experiments are constituted by the stipulation of an epistemic agent, and they are evaluated in terms of their ability to perform a valid test of a hypothesis regarding the causal structure of the effect tested in the target experiment. However, just as in an experimental context, the validity of a replication should not be seen as an all or nothing affair. This might initially seem counterintuitive. Validity is a concept that has its origins in formal logic where a deductive inference is typically framed as being either exclusively valid or not, in accordance with the law of excluded middle. However, in an experimental context validity is never this black and white, and whether an experiment is in fact a valid test of a hypothesis it is often an unsettled matter. This section draws on an analysis of degrees of validity put forward by Feest (2020), to make the case that a given replication experiment can be more or less valid relative to its target

experiment.[74] While Feest's argument centers on disagreement at the level of theory, the analysis given in this section will instead be grounded at the level of experiment. The overarching claim is that if experimental validity comes in degrees, then the validity of replication experiments must as well. This is an extension of the deflationary account's broader strategy of making use of the existing and established epistemic tools to explain replication experiments.

In a general sense, implicit bias is the idea that people's actions are unwittingly shaped by prejudice (Brownstein 2019). On Feest's account, implicit bias is a lumpy construct or theory, meaning there is no agreement, and therefore uncertainty regarding how the causal structure of implicit bias should be characterized and/or measured.[75] The construct's 'lumpiness' implies that there are multiple distinct, overlapping, and potentially clashing mechanisms and/or processes that are able to account for the phenomenon of interest. In implicit bias research, this leaves open the possibilities of testing the construct in partiality, completely missing the relevant construct, testing the whole construct in combination with confounders, as well as testing only a part of the construct with confounders. If this is the case, how can one be certain that an experiment is in fact testing what it is intend to? On Feest's account, if the construct can explain the variance that is found in the test, then it should be seen without controversy as having some degree of validity. However, key to this, and a point that Feest does not sufficiently emphasize, is that the construct must be characterized appropriately; namely as a lumpy one and not one that is highly constrained in the mechanistic sense.

Lumping strategies and lumpy constructs are best characterized in terms of uncertainty. If there is a high degree of uncertainty with regard to how to characterize a given phenomenon, and there is the possibility of testing multiple characterizations of that phenomenon simultaneously, particularly in a case where these various characterizations are nested in one another, then they can at least initially be tested in a lumped fashion. This can be a strategy for slowly and iteratively

[74] While Feest's focus is on construct validity, the graded conception applies to the general conceptions of validity discussed in the previous section (3.1).

[75] The terminology associated with studying a phenomenon with a 'lumping' approach goes back Nosek et al.'s (2011) strategy for testing implicit social cognition which hinges on making use of a variety of techniques that are thought to be measuring the same thing. This is contrasted with 'splitting' strategies that they hold are characteristic of mechanistic approaches, and are focused instead on refining, constraining, and isolating the relevant mechanisms that constitute the phenomena of interest. This type of approach is familiar from the vast literature on natural kinds. For more on this see (Bird and Tobin 2022).

homing in on a more constrained theory. There is of course risk in this strategy, and there should at least be reason to think that the construct exists if the strategy is to be adopted, meaning that there is reason to think that there is in fact something to test.

The hope is that any errors that might emerge would wash out through the process of iterative experimentation. However, there is not always a clear timeline along which this might occur and there is always the risk that a research programs that adopts a lumpy construct will end up going nowhere. This problem becomes particularly worrying when the mechanism that one is in fact probing turns out not to be in any way contained in the lumpy construct, or through the process of iteration the construct is eliminated, or validity ends up decreasing over time (Tal 2017).

As Feest highlights, this all takes place against a background in which the mental processes that her example on implicit bias deals with are likely only being captured in partiality. This can go at least two ways. One involves a case in which a theory is fine grained and developed and is not successfully captured in its entirety by the experiment. This situation is distinct from another in which a theory is coarse and not developed in any sort of fine-grained sense. Here, it should be clear that some aspect of the phenomenon in question is not captured with any sort of certainty by the experiment in question due to the level at which inquiry is taking place. These concerns bring to the fore the role that theory can play in limiting the establishment of validity, and the ways in which attempting to validate experiments is able to contribute to the development of theory; that is, each can develop the other. By iteratively increasing the grain of specification and the precise scope of the theory that is being tested, the process of empirically establishing validity can contribute to the process of understanding the construct. The upshot of Feest's argument then is that the actual epistemic state from which inquiry is performed must be accepted and built into that inquiry. In addition to this, limited knowledge is not grounds for the rejection of a theory or construct outright, nor is it reason to forego experimentation all together. Rather, achieving validity is a process that takes place iteratively over time, and requires fully owning a less than ideal epistemic position.

If constructs that are tested by ordinary experiments can be plausibly seen as having degrees of validity as Feest has argued, and it is accepted that replication experiments should also be

evaluated in terms of validity using ordinary and established epistemic tools, then it is likewise plausible to think that replications experiments should also be seen having degrees of validity.[76] In this way, limited knowledge regarding the causal structure of the effect in the target experiment is not grounds for foregoing replication experiments all together. But how might this be made sense of? Shifting to a notion of replication experiments that can be characterized by degrees of validity, requires changing the way they are developed, performed, and evaluated. Replication experiments can be approached in a binary fashion in at least two senses; either the experiment under consideration is a *valid* replication or not, or it is a *successful* replication, or it is not (see Table 12 below).

TWO SENSES OF GRADING IN REPLICATION EXPERIMENTS	
A replication experiment's validity/invalidity…	… takes a graded evaluation as a consequence of a lack of understanding of the causal structure of the effect in the target experiment.
A replication experiment's success/failure…	… takes a graded evaluation as a consequence of the graded validity of the replication experiment. This is primarily a pragmatic feature if validity has not reached a critical degree.

Table 12. Two ways in which replication experiments take graded evaluations.

This type of binary thinking can be identified in the literature around Carney, Cuddy & Yap (2010), who claimed to show that brief displays of embodied expansiveness (i.e., what they called "power poses") influence neuroendocrine levels (an increase in testosterone coupled with a decrease in cortisol) and risk tolerance behaviors. This effect was purportedly successfully replicated in nine different labs. Several other replication experiments were performed, most notably, by Ranhill et al. (2015), who increased the sample size from 42 in the original study to over 200, and in these experiments, they were unable to successfully replicate Carney, Cuddy & Yap's (2010) original findings. However, after Ranhill et al.'s replication experiments, Cuddy claimed that there were important differences that made the replication of her original experiment

[76] The construct in the case of the replication experiment would concern the causal structure of the effect in the target experiment.

non-diagnostic. For instance, the subjects in Ranehill et al.'s study held two distinct poses, each for three minutes, while Carney, Cuddy & Yap had their subjects hold one pose for two minutes. Therefore, in defending the original experiment against the claim that the 'power pose' effect does not exist, Cuddy argued that this difference between the replication experiment and its target was able to account for the failures. Rephrased under the terminology of the deflationary account, the replication experiments performed by Ranehill et al. were claimed to not be valid; they did not test what they intended to.

Adjudicating Ranehill et al. and Cuddy's clashing claims is a potentially tricky matter. This is in part because uncertainty regarding whether a replication experiment is a valid one can occur for several reasons. One is that there might be a lack of information about the target experiment in the relevant paper. Either the methods or the data in target experiment has not been well documented, or there could be an unintentional lack of transparency regarding the details of various aspect of the experiment. This situation can sometimes be ameliorated through communication with the scientists who carried out the target experiment, but not always.[77] This does not seem to be the problem in Cuddy's power pose experiments. Another reason for this uncertainty, that is perhaps more prominent in the sciences of the mind, is that the precise causal structure of the effect in the target experiment was not well understood by the original authors. In the example above involving the study of power poses, this occurs on multiple levels. In particular, the precise causal structure of power poses, if it exists at all, is in no way clearly understood by Carney, Cuddy & Yap. There is a complete absence of clear criteria that would allow such an effect to be operationalized, including what should or should not count as embodied expansiveness; the phenomenon is not clearly demarcated. This means that there is a severe degree of uncertainty regarding how power poses can be intervened upon and what the boundaries of the effect in Carney, Cuddy & Yap's experiment are.

This translates into there being general uncertainty regarding the assessment of whether a replication experiment is a valid one or not. On the received or standard interpretation, if an

[77] Halina (2021) highlights such a case in which Karin-D'Arcy and Povinelli (2002) had difficulty replicating an experiment by Hare et al. (2000) and were in contact with the original authors to discuss ambiguities regarding the precise methods that were implemented so that a replication experiment could be successfully performed. This represents in many ways a best-case scenario.

experiment is not a replication of the target experiment, it will fail to be diagnostic. Again, this will typically be conceived of in a binary sense. However, uncertainty regarding the precise structure of the effect in the target experiment requires that this be reconsidered. Whether or not a replication experiment is a valid one, as has been argued, can be a matter of degree. This means that the validity of a replication must be established iteratively so that the grain of specification of the claimed effect and its scope can be slowly identified. Importantly, the diagnostic properties of the replication experiment will in most cases inherit the uncertainty that characterizes the validity of the target experiment. Perhaps most strikingly, even without replication experiments ever being attempted, the high degree of uncertainty with regards to the claim in the original experiment should have been front and center. This would have tempered expectations and would have hopefully curbed the backlash that Cuddy faced, which was in many ways devastating and at times unjust (see Dominus 2017).

In addition to this, a direct consequence of the validity of replication experiments being graded is that the success conditions for replications should also at least be capable of being discussed with graded language, as a pragmatic measure reflecting uncertainty.[78] A replication experiment is typically considered successful in so far as it evidences the 'same' effect is found in the target experiment. While this might seem to be somewhat straightforward, the problem is that there is not always consensus regarding what counts as the same effect and there will inevitably be context specific variation regarding what this amounts to. Moreover, if it is not clear that the replication experiment in fact tests the causal structure of the effect tested in the target experiment, the success or failure of such a replication experiment cannot be stated in the typical binary terms, particularly given what this typically implies.

To take another example, in a recent large scale replication project on preclinical cancer biology, Errington et al. (2021, pg. 3) put forward the following criterion; "a replication attempt is successful if the original effect and the replication effect are in the same direction."[79] As the authors admit, this is a remarkably low bar to meet, and their endorsement of this criterion for

[78] Success conditions are formulated as being graded here in part in order to remain somewhat agnostic about the ontological status of the graded success conditions for replication experiments.
[79] This is similar to a criterion for replication put forward by Guttinger (2020) discussed earlier in this part of the dissertation (see Part 1 Section 1).

evaluating whether or not a replication is successful seems to hinge at least partially on the fact that they are evaluating a body of experiments rather than an individual experiment in isolation. The assumption here seems to be that if there was no effect in preclinical cancer biology in general, then one would expect the direction of the effects to be randomly distributed. This is not what Errington et al. (2021) found. However, again there is a sense in which the degree to which a given experiment is a valid replication and the degree to which it is a successful replication can interact in complex ways in this study, and this would remain the case even if the replication experiment's validity had been secured. There are six ways of evaluating replications that Errington et al. (2021) highlight.

(1) Evaluating replications against a null hypothesis
(2) Comparing original effect size with the 95% confidence interval of the replication effect size
(3) Comparing the replication effect size with the 95% confidence interval of the original effect size
(4) Comparing the replication effect size with the 95% prediction interval of the original effect size
(5) Comparing effect sizes in the original experiments and the replications
(6) Combining the original and replication effect sizes

There is no universally agreed upon or acontextual criterion for what makes a replication successful or not, and frequently this is a matter of degree and interpretation. Even in NHST (null hypothesis significance testing), where a successful replication is determined by a statistically significant result being obtained that goes in the same direction as the original effect, clearly demarcated criteria of success that are non-arbitrary do not exist. Such an approach will inevitably struggle to justify clear interpretations of borderline cases and would be better served by adopting an approach that is sensitive to graded success and failure conditions, even if they are only adopted as a pragmatic measure.

Moreover, various statistical approaches, which will be locally decide upon, come with their own limitations. For example, in a recent paper, Fletcher (2021) critiques the following five types of replication measures that have been put forward by the Open Science Collaboration (2015).

(1) a subjective assessment by experts
(2) whether one arrives at the same conclusions in a significance test of the null hypothesis that there is no effect
(3) the results of a significance test of a difference in effect sizes between the original and attempted replication

(4) comparing the original effect sizes with confidence intervals from the attempted replication
(5) statistical meta-analytic methods.

Fletcher (2021, pg. 57)

Each comes with its own limitations and there are potentially conflicts between each way of measuring whether a replication is successful so that no unique outcome in every case can be guaranteed to be arrived at. In addition to this, and more importantly for the current purposes, none of these admit clear, justified, and uncontroversially sharp divisions between what counts as a successful replication and what does not. Fletcher sees this last issue as a clear problem. But it need not necessarily be one if replication measurement is conceived of as coming in degrees. This admittedly makes things more complicated and requires more explicit deliberation. However, assuming this entails more careful consideration and analysis, this would surely be a positive development given the type of panic that has been all too common in response to the replication crisis.

4. Replications Assess Validity

Exhaustive background knowledge of the causal structure of an experimental effect secures the assessment of validity; both of a replication experiment as well as of the experiment it targets. In the same way, limitations on background knowledge, constrains the validity assessment of both. In this way, although replication experiments are constituted by the stipulation of an epistemic agent, validity assessments nevertheless are central to their function. However, the fact that the deflationary account of replication centers validity is, perhaps surprisingly, highly unorthodox in the literature.

It is often claimed both in the philosophy of science and in metascience literature that replications play the role of assessing reliability alone and emphatically *not validity*. For example, Irvine (2020, pg. 11) writes, "Mere direct replication shows nothing though, by itself, about the accuracy or validity of the measurement procedures or the truth of the underlying theory being tested. […] while repeatable procedures clearly do something in a reliable way, it is not clear from the fact of repeatability alone what it is." Feest (2019, pg. 899) makes a similar claim when she writes, "[…]

direct replication can (perhaps) provide evidence for the existence of something, but it cannot say existence of what." Likewise, Zwaan (2013, pg. 1) writes that direct replications, "[…] tell us about the reliability of those findings. They don't tell us much about their validity". And more recently, Errington et al. (2021, pg. 18) writes, "Successfully replicating a finding also does not render a verdict on its credibility. Successful replication increases confidence that the finding is repeatable, but it is mute to its meaning and validity." Given this, coarsely, there are at least three things that it is has been claimed that replications are *not* in the business of evaluating.

(A) The accuracy of the measurement procedure used in the target experiment
(B) The validity of the target experiment
(C) The truth of the theory that is being tested

Under one interpretation of (C), if a replication experiment is successful then there seems to be reason to think that it could count as evidence in support of the theory; at least in certain cases. However, interpreted in another way, one will plausibly never know with certainty whether any theory is true in any sort of deep metaphysical sense.[80] Instead, there will be more or less evidence in support of theory at any given time; empirical inquiry remains necessarily open.[81] Given that (C) is ambiguous in a way that is potentially obfuscating, it will be left to the side.

Instead, the focus here will be on (B), and it will be assumed that (A) is contained in (B). While the view that reliability and validity come cleanly apart is a strong one in the replication literature, detailed analysis of the claim is notably absent from it. Unwittingly, this has served to mask the weakness of the position. Making an argument that this is the case will require performing some initial conceptual work. In doing this, a series of initial idealizations will initially be made.

[80] Yet another interpretation would be that auxiliary assumptions concerning the validity of the target experiment are needed to conclude that a successful replication raises the probability of a theory being true. Nevertheless, flushing out this interpretation requires building in assumptions that are tangential to the argument being made. For this reason, it will be set aside.
[81] Plumbing the details of this would require wading through the realism v. antirealism debate in the philosophy of science and given that there is no clear prospect of settling this debate in general, or within this context in particular, this can be safely set aside for the time being.

RELATIONSHIPS BEWEEEN VALIDITY AND RELIABILITY	
(1) Experiment is valid →	Reliability of the phenomenon **can** be clearly assessed.
(2) Experiment is not valid →	Reliability of the phenomenon **cannot** be clearly assessed.
(3) Target experiment is valid →	A valid replication experiment can be produced. The target experiment's reliability **can** be assessed.
(4) Target experiment is not valid →	A valid replication experiment cannot be produced. The target experiment's reliability **cannot** be assessed.

Table 13. An idealized mapping of the relationship between validity and reliability in replication experiments and their targets.

Assuming a high level of idealization, the relationship between validity and reliability can take on several forms (see Table 13 above). It is assumed here that each are not accidently valid, and going forward it will be assumed that this cannot be the case for the sake of simplicity; that an experiment is valid (including both a replication and a target experiment) entails knowing that it is valid and why in some relevant explanatory sense. It will also be assumed that validity is not achieved due to a variety of epistemic limitations and that invalidity is not intentionally being sought after.

If it is known with certainty that an experiment is valid (1), then enough is also known to be able to begin assessing the reliability or stability of the phenomenon. If the experiment is not valid (2), then reliability or stability of the phenomenon will not be able to be clearly assessed. This is because if the target experiment is not valid, then the causal structure of the effect in it will not be known (assuming the above conditions). If this is the case, then what is or is not being produced will also to a certain degree remain mysterious, which causes the reliability or stability of a given phenomenon to remain mysterious.

For example, in a case presented by Woodward (2000) in a discussion of counterfactual dependence, a method for potassium-argon dating was found to be reliable when used on certain types of rock but not others (Glen 1982). It was eventually shown that by removing atmospheric argon from the rock samples, that the method could be made reliable. However, achieving

reliability in this context came only with a deeper understanding of the causal structure of the potassium–argon dating method in various contexts and an enhancement of its validity (i.e., its ability to measure what it purported to measure). By improving the understanding of how the method works, and what would cause it to not work in certain contexts, the method of measurement became both more valid, which also allowed it to become more reliable.

Once again, the same approach applies to replication experiments. If it is known with certainty that the target experiment is valid (3), then enough information is available to also perform a valid replication experiment of the target and to assess its reliability. This is because if the target experiment is valid, then the causal structure of the effect must also be known to some degree. This creates the foundation for a valid replication experiment in so far as it makes clear what must be targeted. However, uncertainty at one level (the level of the target experiment) flows into uncertainty at another (the level of the replication experiment). If the target experiment is not valid (4), then there is not enough information available to perform a valid replication experiment or to assess its reliability. In this way, the assessment of reliability in the context of replication experiments necessarily starts with an assessment of validity.

TWO TYPES OF RELIABILITY	
(1) 'Results' Reliability	Measurement procedure yields the same results.
(2) 'Causal' Reliability	A common causal structure is reliably produced.

Table 14. Two types of reliability.

Achieving a deeper understanding of this relationship requires further specifying what is meant by reliability, which admits at least two interpretations (see Table 14 below). As Feest (2022) highlights, one view that is found in the psychology literature is sometimes taken to mean that the measurement procedure used in an experiment produces the same results. However, merely 'producing the same results' is a weak interpretation of reliability that is consistent with having an extremely high degree of uncertainty regarding what is being measured or what causes a measurement procedure to yield the 'same results.' On this conception the presence of a wide

variety of distinct phenomena involving distinct causal profiles yielding a common output is consistent with the finding being 'reliable'. This notion is an intuitive one and while it might be a necessary starting point, it is ultimately entirely too weak to be of much use answering the types of questions that arose from the replication crisis regarding the reliability of research. This is primarily because the results centered conception of reliability lacks the tools to determine, with any sort of relevant degree of specificity, what exactly it is that is 'reliably' being measured. If this is the conception of reliability that the replication crisis literature turns upon, extremely little should follow from it.

A second competing and more demanding view of reliability, however, can be found in writers (Franklin 1999; Mayo 2000; Woodward 1989) who, in an ideal epistemic scenario, view it as tightly bound with issues concerning validity. However, as Sullivan (2008) highlights, many of these conceptions target particular aspects of experiments (i.e., data, instruments, techniques, methods etc.) rather than entire experiments conceived holistically. However, on Sullivan's (2008) conception, one that is endorsed by Feest (2021), reliability applies to a whole experiment and requires having a grasp on what brings about the effect that is in question, and what conditions would cause it to be confounded so that it can be used to bear on the evaluation of a hypothesis.[82] On this account of reliability, validity becomes a precondition for reliability rather than something that is extraneous to it.

Still, one might think that this places reliability and validity close together in a way that makes them virtually indistinguishable, which would make the case for retaining the two concepts somewhat implausible. Sullivan (2008), in her discussion of the roles of validity and reliability in neuroscience recognizes this problem and assigns clearly distinct roles to the two concepts. Validity essential comes to play the role of external validity, while reliability is achieved through simplifying measurements. In this way the two concepts pull in opposite direction. Viewed within a broader context, however, this view is somewhat idiosyncratic, and it is unclear how or if it would generalize.[83] An alternate view of the distinction between the two might be to reserve

[82] Feest here writes "claim" rather than "hypothesis" as Sullivan does. In general, Sullivan's language and treatment of reliability is more extensive and subtle.
[83] Sullivan's scope is explicitly constrained to neuroscience alone. This is no way intended as a charge against her account.

reliability assessments for aspects of experiments that are validated. This would potentially make its role distinct from that of assessing the validity of a holistic experiment. Regardless of the precise relationship that ends up being landed on, the reliability of replication experiments cannot plausibly be assessed independently of their validity.

5. Approaching Replication Experiments

One virtue of the deflationary account can be found in its ability to capture the deep diversity that is contained within the practice of replication, while simultaneously holding replication together as a trust-inducing practice. Moreover, it provides a way of approaching both the formulation of replication experiments and the evaluation of them. Across the natural, social, and behavioral sciences, assessing the validity of replication experiments should be approached with respect to the following three 'orienting questions': (1) 'With what degree of certainty is the phenomenon under investigation characterized and individuated?', (2) 'With what degree of certainty is the investigation itself characterized and individuated?', (3) 'What aspect of the target is the replication intended to validate?'. How these questions are answered holds direct implication for the potential evidential import of any given replication experiment. This will in turn determine how any local practice of replication should be structured.

While validity of replication experiments has been argued in previous sections to be a matter of degree, these experiments can also be analyzed across the three 'orienting questions' that are both qualitative and quantitative in character. This means that what counts as a valid replication experiment and whether or not a replication counts as successful will be analog rather than digital in character; that is, it will admit degrees in the ways that was argued for in the previous sections.[84] Again, this means that on the deflationary account there will be no hard lines between types of replications such as conceptual and direct, and for that matter, there are no 'types' of replication on this account at all.

[84] Fletcher (2021) and Fanelli (2017) have alluded to similar positions regarding the evaluation of replications, however, neither has given an account of how this might work.

Approaching replications with these three 'orienting questions' in hand comes with several advantages, two of which will be briefly highlighted here. For one, these 'orienting questions' are tightly interwoven with one another. For example, the goal of the replication (3) will be dependent on the degree of certainty with which the target of replication is characterized (2), in a way that does not allow them to be ultimately considered independently of one another. That is, when approached systematically, dependencies between how these questions are answered can be revealed in so far as they are approached holistically. Moreover, a more general analysis can yield specific hypotheses and predictions that can be studied using meta-analytic techniques. This could, in turn, potentially aid in the broad diagnosis of the status of replication experiments in research programs or entire sub-disciplines. Another perhaps more straightforward advantage of employing these 'orienting questions' when approaching replication experiments is that they clarify the task of evaluating their validity. A more in-depth case for the appeals of adopting this approach will be made throughout the course of this section, primarily by means of the case study presented below.

Given that (un)certainty plays a significant role in the approach, it is worth very briefly addressing it more explicitly. Certainty is a rich concept that has been given ample discussion in both philosophy and decision theory. Here, it is broadly assumed that uncertainty is the absence of certainty. No strong stance will be taken with regards to how to measure it. The kind of (un)certainty that will be focused on here is of the epistemic variety and it is here understood as being a property of beliefs (Reed 2022). Uncertainty with regards to how to characterize or individuate the factual properties of a phenomenon, holds direct implications for the evaluation of a replication experiment's validity.[85] This is particularly the case when the existence of the phenomenon is debated and/or when the empirical methods that are used for investigating the phenomenon are novel and/or poorly understood. This is an epistemic situation that comparative cognitive scientists often find themselves in. Within the contexts of approaching replication experiments with the above 'orienting questions', what is of primary interest is the degree to which the available evidence warrants any determinate conclusion being drawn, and what

[85] Within the context of decision theory this might be represented as a case of option uncertainty (Bradley 2017, pg. 35).

influence this has on actions that shape an approach to replication experiments; both their formulation and evaluation.[86]

To be clear, the appeal of the deflationary account is explicitly not that it provides a singular procedural foolproof guide to 'good' scientific research, or that it cannot be exploited when paired with the wrong conditions or 'bad' actors. Neither the deflationary account of replication nor any of the proposals that come with it are situated to 'solve' the replication crisis or to act as failproof guards against fraud or bias. This is not a reasonable demand to place on any account. Instead, approaching replication experiments with these 'orienting questions' and understanding them within the context of the deflationary account is intended to provide a pragmatic approach to assessing their validity, which will have direct implication for how replications are formulated and evaluated. In this way, the tasks of performing and evaluations replication experiments can move in the direction of becoming less mysterious. This holds the potential to ameliorate the navigation of many issues that have arisen in the wake of the replication crisis.

5.1 A Case Study: Eurasian Jay Pilfering Prevention

A case study will be introduced in this section that will be used to illustrate and fill out the deflationary account and the approach to validity that has been proposed. It involves a recent purportedly failed theory of mind replication study in comparative psychology that was carried out by Amodio et al. (2021).

Previous research has claimed to show that Eurasian jays (Garrulus glandarius) are able to adapt their behavior according to the mental states of a conspecific, and a relatively substantial literature over the past 20 years has been built on these findings (Bugnyar et al., 2016; Dally et al., 2004; Dally et al., 2005; Emery and Clayton, 2001). Legg et al. (2016), for example, claimed that behavior exhibited in the presence of pilfers was intended to prevent cache loss. More specifically, it was claimed that when Eurasian jays know they are being observed, they will prefer to cache their food in a removed location; in this experiment a seedling tray that is in a distant

[86] While the literature on (un)certainty is admittedly a massive one, going deeper into this issue at this point would detract from the larger argument being made.

compartment, rather than one that is easily visibly accessible to a pilfer. This was thought to build on work done by Legg and Clayton (2014), which claimed to show that Eurasian jays prefer to cache their food behind an opaque barrier (a location that was out of the sight of the pilfer) rather than a transparent barrier (a location that was in view of the pilfer). It was also thought to build on Shaw and Clayton (2013), which claimed to show that Eurasian jays attempted to reduce the amount of auditory information that is available to the potential pilferer in order to protect their caches.

While not all of these experiments have explicitly attributed mindreading capacities in any form to Eurasian jays and have instead focused on the correlation between the cues that could or could not be mental states, a related line of research has attempted to target mindreading capacities more directly, specifically their sensitivity to conspecifics' desires. Ostojic et al. (2013) claimed to show that males are sensitive to the food preferences of females and provide them with food accordingly. That is, only when a male jay was able to previously observe what a female jay ate, was he purportedly able to infer her desires and feed her what she wanted. Ostojic et al. (2014) built on this research that introduced a conflicting and neutral condition in which the desires of the male and the female either matched or were clashing and had results that were consistent with the original finding. In a further experiment, Ostojic et al. (2016) controlled for experimenter expectancy bias, because previous results were consistent with an explanation that accounted for the results through an appeal to the unconscious expectation of the experimenter. They then repeated the experiment and 'blinded' the experimenter by removing them whenever possible. When this was not possible, they made the testing conditions opaque to them (e.g., when food was given to a male jay the experimenter did not know whether the female jay had been fed or not). In this case, as well, they came out with the same results.

Amodio et al. (2021) then considered these experiments as a cluster that plausibly indicate that Eurasian jays make use of several strategies to protect their caches from loss by responding to the cues (perspective or desire) of a conspecific that could be a pilfer. In sharpening what they take this evidence to show, they claim that while the existing evidence is not sufficient to prove the presence of a theory of mind, nor to isolate a specific cognitive mechanism, taken as a whole, the evidence suggests that Eurasian jays are capable of flexibly responding to the social cues of

conspecifics. Over a series of five experiments, Amodio et al. (2021) both attempted to extend and subsequently replicate studies in this line of research. In doing so they claimed that they were able to find no evidence that the jays were aligning their behavior in response to the mental states of other agents.[87]

They established their first two experiments by combining two previous experimental protocols from the research program; one from Legg and Clayton (2014) which was used to establish the manipulation of the observer's perspective and the other from Ostojic et al. (2017), which was used to manipulate the observers desire for different foods. They framed these initial experiments as a test of questions that were remaining from previous experiments in the research program. Amodio et al. (2021, pg. 3) write, "Building on previous studies in the caching context, we tested whether jays can integrate information about a conspecific's perspective and current desire to selectively protect those caches that are at most risk of being pilfered." However, neither of these two initial experiments that were performed provided clear support for this hypothesis that scrub jays can integrate information from multiple cues to protect their caches from being pilfered by another scrub jay, and neither was able to demonstrate the presence of a statistically significant effect.

This presents a problem; it is not entirely clear how to characterize these experiments. As highlighted, in experiments 1 and 2 Amodio et al. (2021) state that they are attempting to build upon previous findings, however, they also see this as a test of the 'reliability' of Legg and Clayton (2014) and Ostojic et al. (2017), particularly when they write, "[…] the inconsistencies with previous research could also be due to previously reported effects not being reliable enough to form the basis of follow-up studies" (Amodio et al. 2021; pg. 9). While the authors are careful when making these claims these stipulations surrounding these initial experiments cause them to veer into the territory of being replication experiments.

[87] Interestingly the title of their paper makes claim to "little evidence", but from the content of the paper it is unclear what this "little evidence" refers to. They give no indication of the presence of positive evidence aside from the existence of previous studies.

In addition to this, there is perhaps an interesting methodological lesson here regarding the prerequisites for genuinely build on findings. If the neither the effect nor the methods for producing the effect are well established or understood, any attempt to 'build' on findings should be seen as tentative at best.[88] This would remain the case had Amodio et al.'s (2021) experiments been successful. In this way, framing them more explicitly as replication experiments and more explicitly assessing their validity would serve to clarify their status.

Given that the combined experimental protocol in in experiments 1 and 2 failed to yield positive results, Amodio et al. (2021) backtracked and explicitly performed replication experiments that were targeted at the experiments that the combined experimental protocol was based on. Experiment 3 was a simplified version of Legg and Clayton (2014) that was intended to test whether the scrub jays could use visual perspective taking to protect their cache from being pilfered from a conspecific, and experiment 4 was a purported 'direct' replication of Legg and Clayton (2014). Experiment 5 was intended to determine whether the transparent barrier that was present in experiment two was in fact a confounding variable, which was an effect that was claimed in Ostojic et al. (2017).

In the following, these replication experiments will be analyzed considering the deflationary account and the guiding questions presented above. In doing so, it will be shown how performing such an analysis can aid in the interpretation and approach to formulating replication experiments more broadly.

5.1.1 With What Degree of Certainty is the Phenomenon Under Investigation Characterized and Individuated?

The starting point of scientific inquiry can be broadly characterized by the occupation of a less than ideal epistemic standing with regards to a phenomenon of interest. The purpose of performing scientific inquiry is precisely to increase knowledge regarding this phenomenon and to improve this epistemic standing. However, in many cases, the precise characterization and

[88] In a slightly different way, this is a point that has been previously alluded to by Hüffmeier et al. (2015), however, their hierarchy involves performing direct prior to conceptual replications.

individuation of the target phenomenon is subject to a particularly high degree of uncertainty. In these cases, it can be unclear whether the object of inquiry exists full-stop or if the boundaries of the phenomenon are being accurately captured in a way that allows straight forward inquiry to be carried out. These types of 'fuzzy' phenomena are unfortunately typical in comparative cognitive science and this state of affairs has direct implications for how the science is carried out. This is because the degree of uncertainty regarding the characterization and individuation of the phenomenon is determinate of the type of evidential import can be derived from the results of an experiment.

As has been discussed previously in this part of the dissertation (see Part 1 Section 2.3), in recent years, the role of theory and concepts in the practice of replication, in addition the ability findings to be generalized (Yarkoni 2022), has gained increased attention. One upshot of this literature has been that there is a continuous space between phenomena that are grasped sufficiently well, and ones that are not, and there is no prospect of identifying a sharp line between the two. This implies that a graded account of the individuation and characterization of the phenomenon is needed. However, there are multiple ways for a scientific phenomenon to not be well understood. Given this, the project of understanding the space of fuzzy phenomena, and the particular type of fuzzy phenomenon that is present in any given local inquiry, will require detailed analysis. In addition to this, any sort of evaluation here will clearly be qualitative in character. This is because it is neither possible nor desirable to assign a single number to the, 'understanding of the phenomenon under investigation'.

The general understanding of 'phenomenon' adopted here is the product of multiple experiments and studies, and it is assumed broadly that this is the only way in which phenomena could come onto the scene.[89] As a starting point, a position put forward by Woodward and Bogen (1988) regarding phenomena will be adopted, who argue that they are only made observable through

[89] This account differs from stories that are often told in the cognitive sciences that tend to set empirical investigations as starting from zero. Coarsely, the idea here is that scientists land upon hypothesis and phenomena of investigation intuitively or in some mysterious irrational way. This can be seen in Heinrich (2018) and Andrews (2020), among others. By contrast, it can be assumed that the existing empirical research informs the content of hypotheses and phenomena of investigation, and no investigation can be framed as intelligibly starting from zero. This is not to say that there is no space for intuition or irrational motivations, but that these alone do not provide a sufficient account. This is congruent with a position put forward by Sullivan (2009).

data. In addition to this, the relevant phenomena will be viewed as always already being theoretically loaded and empirically embedded, albeit in a variety of ways (Wimsatt 1981). While multiple studies can approach a single phenomenon from a variety of perspectives, and with a variety of goals that direct inquiry, the more a phenomenon is studied in a variety of ways, the better the understanding of it will be. This has been one of the underlying assumptions of methodological triangulation (Heesen et al. 2019). It need not follow from this that a unified or integrated account of an underlying mechanism can be formed at any given moment (Craver 2007), or that the various lines of inquiry will necessarily be compatible with one another, even in the long run. The adoption of such a position would be too strong and is ultimately unwarranted. Rather, the idea is a more minimal one; more lines of inquiry are thought to hold the potential to yield a better understanding of the phenomenon in a general sense (Kuorikoski & Marchionni, 2016; Basso 2017).

This is central to an approach to replication experiments on the deflationary account in so far as the larger understanding of the phenomenon under investigation, independent of the causal structure in any given experiment, has direct implications for the experimental protocols that can be formulated, in addition to how they are carried out.

For example, using the Amodio et al. (2021) as a case study, the following question can be asked; to what extent is mindreading as a phenomenon understood? In short despite over forty years of explicit empirical investigation, mindreading remains extremely fuzzy. In part, this is because every experiment that has been carried out is underdetermined in a significant way causing serious theoretical disputes that remain open, and there is no overarching background theory in discipline that can be drawn upon to direct or provide significant constraints on the study of the phenomenon more generally.[90]

In addition to this, particularly in cases in which the understanding of the phenomenon that is being studied are poor or 'fuzzy', as is the case in the mindreading research program conceived more broadly, there will plausibly be multiple levels at which the phenomenon is approached and there will often not be a unique answer to which aspects or features of the phenomenon are

[90] These issues will be given more explicit treatment in the subsequent part of this dissertation (see Part 2 Section 2).

relevant. Indeed, this problem is in many ways to be expected. In the case of Amodio et al. (2021) in particular, this will potentially include the larger cognitive profile of scrub jays as a species, the individual cognitive profiles of the animals that are being tested, as well as experiment in the human theory of mind research program that inform the shape of the phenomenon more broadly. Again, hard boundaries will not be able to be set from epistemic perspectives that are characterized by severe uncertainty. However, by first making clear the background understanding of the phenomenon under investigation, expectations regarding the replication experiment can be clearly and transparently set.

5.1.2 With What Degree of Certainty is the Replication Experiment Characterized and Individuated?

While closely connected to the characterization and individuation of the phenomenon under investigation just discussed, the characterization and individuation of the target of replication should importantly be held apart from it. Evaluating it broadly involves determining the extent to which the target experiment is understood as an object of inquiry; that is, understanding the causal structure of the effect in the target experiment and the way in which it is sensitive to confounds and interventions. This will vary, at least in theory, independently of the characterization and individuation of the phenomenon under investigation. For example, when a new experimental method is being used to study an aspect of a phenomenon that is otherwise well understood.

This can be a difficult task regardless of the target of the replication experiment, in part due to problems arising from underdetermination and a necessary reliance on auxiliary hypotheses that comes as a part of it. However, the fact of underdetermination clearly does not render empirical inquiry inert. This is because underdetermination comes in degrees and auxiliary hypotheses can be more or less confirmed in any given instance. Explicitly building an assessment of the causal structure of the effect tested in the target experiment into the formation and evaluation of replications can be more or less transparently evaluated.

However, even if what caused the effect of interest in the target experiment seems to be well understood, it is often unclear what changes it will be sensitive to. This has been a substantial topic of interest in the literature on replication. For example, Collins (1985) argued within the context of assessing replication experiments that it is not always clear what is necessary for producing the results of an experiment.[91] In doing so, he placed a focus on the skill of the experimentalist, which cannot be easily captured in the methods section of a paper, despite claims to the contrary (e.g., Machery 2020, among many others). Particularly regarding novel findings, Collins highlights the way in which consultation with the authors of the target experiment is almost always a necessary feature of performing replication experiments. A similar line of argumentation can be found in Soler (2011), who has termed the inevitable role that tacit resources, skills, and presuppositions play in experimental practice the 'opacity of experimental'.[92] While this is at least in part a debate about 'contingentism' versus 'inevitabilism' in characterizing the historical progression in science, appeals to this feature of replication experiments has also been used to make sense of replication failures, and to emphasize the difficulty of interpreting the results of attempted replication experiments.

In a more general sense, both authors highlight the 'human' role in experimentation and the fact that this role cannot be easily articulated. In part, this is meant to capture the fact that experimentalists make choices about every aspect of an experiment, and these choices can be unconscious, implicit, or even irreducibly implicit (Soler 2010). This leaves open the possibility that essential aspects of a target experiment will inevitably not be captured by the replication experiment, which clearly poses a significant challenge for the project of producing valid replication experiments. However, there should be more sensitivity to the fact that understanding experiments can be difficult, the fact that the practice of experimentation includes elements of tacit knowledge, regardless of the form, need not necessarily be detrimental to replication experiments.

[91] The tactic dimension of experimental practice has been a running thread in social studies of sciences. For classic works on this theme see Kuhn (1969), Polanyi (1958, 1967), and Fleck (1935).
[92] In doing so she highlights that this is in tension with the principle that experimenters be substitutable with one another. This is in tension with Machery's (2020) resampling account and its broad appeal to random sampling.

As is emphasized in the deflationary account, only understanding the experimental conditions in the 'right way' is necessary for a valid replication experiment to be carried out. Here, everything hangs on what the 'right way' amounts to; that is, the causal structure of the effect tested in the target experiment needs to be captured. In this way, the account evades the problem of implicit knowledge. If implicit aspects of an experiment are in principle not able to be captured, then the causal structure of the effect tested in the target experiment is not plausibly understood. Appeals to expertise or implicit knowledge as they have been repeatedly done in the replication literature only to highlight the extent to which the causal structure of the effect tested in the target experiment has either not been sufficiently articulated or not been sufficiently understood. If this is the case, more work needs to be done and the challenge needs to be faced head on.

Given this, the following questions can be asked of Amodio et al.'s (2021) replication experiment; to what extent is the causal profile of the target experiments understood and what implications does this have for both how this replication was formulated and how it should be interpreted? In answering this, focus will be placed on experiment 4 discussed above in particular, which was stipulated to be a direct replication of Legg and Clayton (2014). This is done here because experiment 4 lends itself to the clearest analysis. However, other experiments will be touched on as well. Ultimately, it will be argued that the causal structure of the experiment that produced the effect of interest in the target experiment is not understood in a way that would allow for the replication experiment to be claimed to be a convincingly valid one.

In the experiment being targeted in replication experiment 4, Legg and Clayton (2014) tested whether Eurasian jays (Garrulus glandarius), when being observed by a conspecific would choose to cache food behind a transparent or an opaque barrier. Their hypothesis was that the jays would prefer to cache behind the opaque barrier while being observed by a conspecific in order to prevent pilfering from taking place. When not being observed at all (i.e., while in private), the jays selected between the opaque and the transparent cache sites at chance. However, when a conspecific was observing them, they displayed a statistical preference for the out-of-view location behind the opaque barrier.[93]

[93] This is resonant with a set of experiments by Hare et al. (2000, 2001) discussed previously in this part of the dissertation.

As previously emphasized, crucial to understanding the effect exhibited in this experiment, as well as understanding the way in which it should be characterized, is determining which aspects of the experiment are involved in the relevant casual structure that produced the effect, which are superfluous to it, and what interventions will cause the effect to be interrupted. The degree of certainty regarding this will determine the extent to which any replication experiment can be determined to be a valid one. One place to start making sense of this, is to look more closely at the methods and the conditions that were tested. I take it that the general brevity of the paper and the relative simplicity of the methods used make it a good case study in illustrating the larger point of carefully considering this many features that were involved. In doing so, how the target experiment is characterized in the paper with a high degree of uncertainty will be highlighted, and the implications that this will have for any replication experiment will be brought out. The features of the target experiment will be listed numerically so they can be more easily referenced. While Legg and Clayton list the 'procedure' separately, it will be treated here as a feature that cuts across multiple aspects of the experiment.

(1) **Environment/apparatus:** The jays that participated in this experiment were housed in a large outdoor aviary and were tested in a small indoor space that was made up of two compartments. The precise measurements of these compartments have been provided by Legg and Clayton in the methods section, as have the materials that have been used to construct the barrier. The observing jay was in one compartment, while the caching jay was in the second. This compartment contained a T-shaped barrier that was placed 1 meter off the ground. One of the 'arms' of the T was constructed of opaque plastic, while the other was constructed of transparent plastic. Behind each of these arms was a seedling tray that was filled with sand. Sand was used so that the observed jay was unable to hear the food being cached.

(2) **Food:** The jays were fed a maintenance diet of soaked dog biscuits, cheese, seeds, nuts and fruit, and were provided with unobstructed access to water. In this experiment, the jays were provided with a bowl of 30 peanut halves that they were able to cache. Over a 15-minute period, the caching jay was given the opportunity to cache in one

of the two seedling tray that were filled with sand. Two and a half hours later, the caching jay was then allowed access to the caching trays in order to retrieve the cached peanut.

(3) Participants: There were eight participant jays in the experiment. Three served as subordinate observers and another three served as dominate observers. Most jays, apart from one, were placed in the caching condition prior to being placed in the observing condition. This was intended to prevent caching behavior from being influenced by being in the role of the observer.[94] The remaining two jays were excluded from the study because they did not habituate to the compartments containing the T shaped barriers.[95]

Legg and Clayton implemented control conditions in order to constrain the space of possible explanations for the behavior exhibited, and in doing so defend their thesis in the discussion section that, "[…] the jay's choice of caching location appears to be dependent on what conspecifics can or cannot see" (Legg and Clayton 2014, pg. 1226). There were two conditions in this experiment: one with an observer and one without. The main control that was implemented was intended to rule out the effect of dominance on the caching behavior. Legg and Clayton implemented this by running each participant through three trials; one while being observed by a dominant, one while being observed by a subordinate, and one while not being observed. Aside from the structure of the experimental protocol itself, no other conditions were implemented.

The choice of what to include in the methods section of a paper is presumably dictated by what the authors believe are the plausibly important elements of the experiment that led to the effect exhibited. Interpreted more strongly these can be framed as a claim regarding the causal structure

[94] It is not entirely clear why the inverse condition was not controlled for. It seems just as relevant to understand what influence the being in the caching role previously had on behavior that is exhibited in the observing role. While it seems plausible that the low sample size might have placed a limitation on testing this condition, Legg and Clayton do not comment on this choice.

[95] While the elements that Legg and Clayton have chosen to make explicit have been independently considered, the enumerated features here cannot plausibly be considered independently of one another in so far as they holistically constitute the causal structure of the effect that was exhibited in the experiment.

of the effect. The content of this methods section is of upmost importance in so far as it dictates what gets included and what ends up being excluded from any attempted replication experiment. For example, Legg and Clayton provide the descriptions of the physical space, including both the compartments and the larger enclosures, presumably that they take to be an important aspect of the effect in the target experiment. This is thought to be supported by the contrastive behaviors that the jays exhibit in response to the T shaped barriers. That is, they chose to cache behind the opaque rather than the transparent barrier, which might suggest that these features of the barrier played a role in this behavior. The setup alone is enough to produce some effect but absent further testing, it is not enough to understand it in any sort of detailed sense that will allow which aspects of the experiment are important to producing the effect to be clearly stated, which are not, and how will the effect be sensitive to interventions and confounds. For example, Legg and Clayton specify the type of plastic that is used in the transparent behavior; Perspex ©. Is this important or can another type of plastic be used? The lighting conditions in the compartment have not been specified. Does this play a crucial role in the effect? Moreover, the housing history of these jays might plausibly be relevant here, which is something that Amodio et al. (2021) allude to in their discussion. In addition to this, the material that the compartments have been made of has not been specified. There are numerous auxiliary hypotheses regarding the environment that are a part of this experiment and only some of these have been explicitly addressed by Legg and Clayton. This seems to be because there is insufficient knowledge regarding these auxiliary hypotheses, as well as the ones that were meant to be more directly tested, to state with precision, which are responsible for the effect.

The same goes for the other features of the experiment that have been addressed by Legg and Clayton. For example, the food. The type of diet that the jays have been fed and the peanuts that have been used for caching has been specified to a reasonable degree. But aspects like feeding times and quantity have not been. The time of day at which the experiments have been conducted could plausibly play a crucial role in this effect. The same goes for the history of the jays that were participants in the experiment, as Amodio et al. (2021) again allude to. Is it important that they have been subject to cognitive tests for most of their lives and that they have been raised in a large enclosure? Exactly which aspects of this are important, given that it is safe to assume it features prominently in the causal structure of the effect? The point here is simply to

highlight how open the system is in which these experiments are being conducted, and how this 'openness' poses serious challenges for accounting for the experiment in question, absent further constraints. Given these and other limitations, the understanding of this experiment is severely limited, which makes the prospect of carrying out a clearly valid replication experiment of it virtually non-existent.

This is even more so the case given the content of the main claim of the paper, and consequently the content of the effect; that jays are sensitive to what conspecifics see and cannot see while caching. This type of hypothesis has been notoriously different to nail down in the animal cognition literature in so far as it imports many highly troubled features of the non-human animal mindreading debate.[96] Typically seeing is thought to imply a mental state, while something like line-of-gaze is not, although it is often entirely unclear what this distinction amounts to. If Legg and Clayton mean to characterize the behavior of the jay in terms of theory-of-mind, this will have different sorts of implication for understanding the effect exhibited than if it was not.

5.1.3 What Aspect of the Target is the Replication Intended to Validate?

On the deflationary account, a replication is an experiment that is stipulated by an epistemic agent to be a test of a hypothesis regarding the causal structure of the effect tested in a target experiment. However, the validity of a replication experiment, which is central to its utility, is something that can, and potentially must, be assessed in a variety of ways. This is particularly the case given that replication experiments can take on degrees validity as has been argued in the previous section of this part of the dissertation (see Part 1 Section 3.2). This feature of replication experiments has not always been adequately acknowledged in the replication crisis literature; however, it has been given attention by some.[97]

For example, Schmidt (2009) has argued that replications have five functions that they are unable to fulfill simultaneously: (1) addressing sampling error (chance result), (2) controlling for artifacts (lack of internal validity), (3) addressing researcher fraud, (4) testing generalizations to different

[96] This will be discussed in more detailed in Part 2 of the dissertation.
[97] For some, a single replication experiment will fulfill multiple functions at once.

populations, and (5) testing the same hypothesis of a previous study using a different procedure. While (4) and (5) are typically understood to be a unique to what has been called conceptual replication, (1)-(3) are thought to be the domain of what has been called direct replication.[98] In addition to this, the many other functions of replication experiments are touched upon in the various typologies that have been discussed earlier in this part of the dissertation (Part 1 Section 1.5). The possible functions of replication are of course not limited to these. The larger point, however, is that while each of these broadly function as validity assessments, it is possible for them to come cleanly apart from one another, and in many cases, there might be no other way of assessing the validity of the replication. For example, researcher fraud can be plausibly tested without directly addressing sampling error. By making this more explicit, a better account of the various functions that replications take on can be given within the context of the deflationary account.

However, an initial distinction should be made here between the 'goal' and the 'function' of a replication experiment. In the following, a 'goal' will denote a function that the epistemic agent intended the replication experiment to fulfill, while the 'function' will denote the function that the replication experiment was in fact able to fulfill. In many replication experiments, these two come explicitly apart.

This can be illustrated by applying Schmidt's (2009) scheme to the Amodio et al. (2021) case study that has been analyzed along each dimension thus far. First, the goal of this replication experiment is seemingly not to (1) address sampling error in so far as the same population was used in the replication experiment as was used in the target. While Amodio et al. (2021) raised concerns about the population's history in attempting to account for the jay's failure on the task, they did not take up this aspect of the target experiment in their replication experiments. In making use of the goal versus function distinction just presented, it was neither the goal nor the function of this replication to address sampling error.

However, one of the goals of the replication experiments that was carried out does seems to be captured by (2); controlling for experimental artifacts. While Amodio et al. explicitly attempted to

[98] Once again, the distinction between direct and conceptual replication breaks down on the deflationary account.

do this, even on their own lights it is not clear that they were successful in doing so. From the failure of replication experiments 1 and 2, Amodio et al. concluded that the experiments show an inconsistency with the effects that had been claimed in the literature thus far, which is then used as a basis for challenging the findings of both Legg and Clayton (2014) as well as Ostojic et al. (2017). In experiments 3-5 no statistically significant effect was identified, which served as the basis for similar conclusions. For example, Amodio et al. (2021) highlight the low power of their studies and the possibility that their sample was in some way non-representative because they used the same birds that were tested in the original study five years earlier in Clayton's lab. The alternative explanation for these failures that is given, is that their behavior was somehow altered after years of being tested upon. This would in essence amount to a shift in motivation. However, they highlight that this would be a specific shift that had to do with cache protection, rather than caching in general, given that caching behaviors remained stable.

In general, the authors prevaricate between hedging their conclusion and making stronger claims about the reality of the effect, as is evidenced in the title of the paper, and they are, to their credit, more caution than many in comparative psychology when framing their claims. Nevertheless, they conclude that the failure of all five 'replication' experiments provides reason for thinking that the original effect size was in fact overstated, and that this is consistent with the low power of the target studies when paired with publication bias. They even go on to claim that the failure of their 'replication' experiments, "[…] suggests that previous effect sizes are likely overestimated, or at the very least, that the effects cannot be consistently elicited in the same or similar samples of birds" (Amodio et al. 2021, pg. 16). As hedged as this conclusion is, this is too strong given the evidence that is available to them. This means that while the stated goal of the replication experiment was to control for experimental artifacts, the function of the replication experiment was only successful in doing so with an extremely high degree of uncertainty. While this replication experiment can serve as the basis for further replication experiments, a clearly valid replication experiment in this sense remains well out of reach for Amodio et al (2021).

In proceeding through the list of functions that were presented by Schmidt, there is no reason to think that Amodio et al. (2021) held the goal of (3) addressing researcher fraud or (4) testing generalizations to different populations. Initially, the first set of experiments were meant to test an

of extension of previous experiments in this line of research. This seems to fall under Schmidt's function (5) testing the same hypothesis of a previous study using a different procedure. As previously highlighted, the first two experiments failed, which led to the attempted replication of the original experiments by Legg and Clayton (2014). However, while the goal of the replication was presumably to test the same hypothesis under the same procedure, it does not seem that it in fact functioned in doing so. Again, it is not clear that the same hypothesis was being tested in any way other than one that binds it to the intentions of the experimenter. Put simply, there is just too much uncertainty, and too many untested auxiliary hypotheses present in these experiments to provide an assessment of this.

5.2 Takeaways from the Approach

An approach to replications experiments within the context of the deflationary account has been introduced. As a part of this, Amodio et al.'s (2021) study has been analyzed through a series of 'guiding questions.' This allowed for several conclusions to be drawn with regards to the status of Amodio et al.'s replication experiment that will now be summarized.

First, the precise characterization and individuation of the target phenomenon (i.e., mindreading) is subject to a particularly high degree of uncertainty and is 'fuzzy' in character. This provides reason to think that any replication experiment of a target experiment that is intended to study this phenomenon will be particularly unconstrained and implies that the ability of Amodio et al.'s replication experiment to achieve a high degree of validity should broadly be approached with skepticism.

Second, the causal structure of the effects in the target experiments that were replicated is poorly understood. Each aspect of Legg and Clayton's (2014) target experiment was characterized and individuated in a way that left room for significant doubt regarding the causal structure of the relevant effect. This should count as another initial block for any replication experiment of this target experiment being able to plausibly achieve a high degree of validity.

Third, Amodio et al.'s replication experiment was constrained to the following functions: controlling for experimental artifacts and testing the same hypothesis of a previous study using a different procedure. Defining the validity assessing function of the replication experiment will serve to define the bounds within which such an experiment can be approached and interpreted. However, given Amodio et al.'s epistemic positioning with regards to the target experiments, their ability to control for experimental artifacts is extremely limited, and their prospects of carrying out a valid replication experiment are bleak.

All of this adds up to a situation in which the conclusions that are drawn from Amodio et al.'s replication experiments should be approached with significant caution and skepticism. In short, these are in no ways clearly valid replication experiments. Importantly, this also means that even had the experiments been successful, any sort of evidential import taken from this should be characterized with a high degree of uncertainty.

The takeaways from the approach, however, are not constrained to an analysis of Amodio et al.'s replication experiments. Beyond the analysis of this experiment, the approach that is underwritten by the deflationary account serves several purposes. One is to evade the all too simple replication versus non-replication distinction that has been the driver behind much of the replication crisis discourse. By accounting for replication experiments with these 'guiding questions', an account has been given that captures both the diversity and the unification of the practice of replication, while simultaneously acknowledging the varying epistemic standpoints from which replication experiments are conducted and how this changes what a replication experiment amounts to. In addition to this, the approach broadly makes room for wider analysis of patterns among replication experiments within research programs or broader disciplines. Questions such as how some 'guiding questions' function as proxies for others, or how sets of replication experiments can be captured at a higher-level characterization, can be posed and systematically studied. In this way a finer grained and more progress generating account of replication experiments can emerge.

6. Advancing Replication

Replication experiments are rare. They are difficult to publish and there is currently little incentive to carry them out. Historically, when they have been performed, it has been very difficult to establish consensus around how to interpret them. Given this, deflationary account is aimed at removing some of the mystery surrounding performing and evaluating replications. In adopting this account, any debates arising around a replication experiment in any given context can become more tractable, and expectations can be managed regarding what sort of evidential import replication experiments are able to hold.

Rather than providing necessary and sufficient conditions for replication experiment or attempting to divide them up into a clean typology, the deflationary account that has been defended provides a framework that can be employed in particular research contexts where these questions can best be evaluated according to local standards, while simultaneously allowing them to appeal to overarching guides. Nevertheless, this account is not equipped to adjudicate in the abstract what should or should not count as a valid replication experiment beyond stating that it must be a valid test of a hypothesis about the causal structure of effect tested in the the target experiment. As has been argued, this can take place in several ways. Again, these are debates that will be most productively developed by shifting away from a 'rubber stamp' procedural 'just do it again', 'naïve' type of approach. Given this, there might be a concern that the practice of replication is left too open to manipulation. However, the specters of manipulation and fraud will pose a challenge to every account of replication and there is no automatic or surefire way to guarantee that this does not happen. Demanding that an account of replication provide such a constraint is too demanding and unrealistic.

Another consequence of deflationary account is that it is at the moment unclear if there is a general replication crisis that spans the sciences, given that replication experiments have been evaluated in numerous ways, and they are approached from a diversity of highly contested epistemic standpoints. This is even more so the case given that the validity of replication experiments should be seen as coming in degrees as has been argued. This means that any sort of coarse-grained binary verdict regarding the presence or the absence of a replication crisis will likely be unjustified. Still, the discourse around the replication crisis writ large has given rise to several foundational debates about research quality and trustworthiness, as well as a much-needed

analysis regarding what a replication is and how it should function, as has been demonstrated both through the introduction of the deflationary account and the analysis of the other currently existing accounts of replication. This at least partially implies that when it is asked what types of policy measures should be taken, sweeping a-contextual changes are currently not warranted.

Instead, careful analysis of the epistemic situation underlying a research program should take priority. In addition to this, it worth holding in mind that replication is not the only tool in the box, and although it is one of the primary practices that we use to determine reliability, it is not a cure all, and there are several issues pertaining to reliable and valid scientific practice that it will not be able to immediately illuminate. This is particularly the case in immature sciences as has been highlighted, and in cases involving ghost literatures (Machery 2021).

In concluding this first part of the dissertation, the primary initial positions and the key proposals underlying the deflationary account will be briefly recounted. It was first argued that naïve accounts of replication relied heavily upon an underspecified and unconstrained notion of identity or relevant similarity in determining the utility of replications. This made them inadequate for evaluating many of the issues that emerged in the literature of the replication crisis. Then, two recent novel accounts of replication were analyzed; Machery's (2020) resampling account and Nosek and Errington's (2020) diagnostic account. The resampling account failed in so far as it includes cases that should not count as replications and excludes cases that should count as replications. The diagnostic account failed because it overburdened the practice of replication and introduced a definition upon which potentially anything counts as a replication. The discussion then shifted to causal accounts of replication. Four accounts were introduced and analyzed: Norton (2016) Anjum and Mumford (2018) Feest (2021), and Irvine (2021). Coarsely, while these accounts held several appeals, their primary and shared limitation was found in the fact that they were overdemanding in a way that either explicitly or implicitly claimed that the practice of replication should potentially be abandoned. At this point, a novel demarcation criterion was introduced on which an experiment or line of research should be subject to the replicability criterion *iff* there is an adequate understanding of the causal structure that is responsible for the effect of interest in the target phenomenon. Then, three typological accounts of replication were analyzed; LeBel et al. (2017), Hüffmeier et al. (2015), and Leonelli (2018). These accounts were

shown to generally fail in so far as they did not test the causal structure of the effect in the target experiment and instead relied generally on a similarity relation between potentially irrelevant aspects of the target and the replication experiments. In the final 'critical' section of the first part of the dissertation, three interpretations of the replication crisis were analyzed. There it was argued that there can either be a replication crisis or a theory crisis, but that both cannot exist simultaneously.

At this point the deflationary account of replication was introduced, and the positive account was more generally developed. On the deflationary account, a replication is an experiment that is stipulated by an epistemic agent to be a test of a hypothesis regarding the causal structure of the effect tested in a target experiment. A replication experiment is valid if it is in fact a test of hypothesis regarding the causal structure of the effect tested in the target experiment. Whether or not a replication experiment is valid therefore becomes the central question. However, if experimental validity comes in degrees, then the validity of replication experiments as well as its success must also come in degrees. This means that more detailed and fine-grained analyses of replication experiments must be performed. In this way, it was argued that background knowledge of the causal structure of an effect secures the validity assessment, both in the replication experiment as well as its target. This means that replications cannot assess reliability without simultaneously assessing validity, contrary to the orthodox view. Given this the validity of replications experiments can be assessed via the following three 'guiding questions': (1) With what degree of certainty is the phenomenon under investigation characterized and individuated? (2) With what degree of certainty is the investigation itself characterized and individuated? (3) What aspect of the target is the replication intended to validate? In the final section of this first part of the dissertation this analysis was put to work to analyze a replication study performed by Amodio et al. (2021).

The next part of the dissertation will build on much of the analysis done in the first part. However, the focus will shift more directly onto comparative cognitive science and the empirical and theoretical challenges that it has faced.

2

ESCAPING THE VALIDITY CYCLE

0. Introduction

As argued in part one of the dissertation, the ability to evaluate the evidential status of an experiment is highly contingent upon the type of knowledge that is held of the causal structure of the effect in that experiment. Part two of the dissertation builds on this work and shifts the focus away from the higher order project of evaluating experiments, and onto the question of why, on many in the discipline's own lights, first-order empirical projects in comparative cognitive science have repeatedly struggle to produce the type of knowledge that would make them conducive to clearly diagnostic replication experiments that might come to target them. The broad aim here is to distill and address a pattern of problems that have affected a significant portion of research programs in the discipline of comparative cognitive science.[99] This will include an analysis of existing diagnoses for the emergence of this pattern of problems as well as an analysis of various proposals that have been made to overcome them. While the picture on offer of might seem initially bleak or overly pessimistic, realistically assessing the difficult epistemic situation that much of the discipline is faced with, is a precondition for achieving the type of knowledge that is it generally directed at producing. In this way, the analysis offered should be recieved as a cautiously optimistic one. However, whole-heartedly adopting this optimism will require abundant

[99] To be clear from the outset, comparative cognitive science contains, and will contain, a huge diversity of research programs and the claim is in no way that the problems and challenges introduced in this part of the dissertation are present in equivalent or uniform ways across each of these contexts. The problem is presented here in a somewhat 'generic' sense in order to capture what will be identified as a significant trend that needs to be addressed.

modification to 'business as usual' in much of the discipline. This is to say that the goals set out by those working in comparative cognitive science can plausibly be achieved, but that doing will be hard won over a substantial period of time.

This second part of the dissertation unfolds over four main sections.

Section (1) diagnoses the problem that a significant portion of research programs in comparative cognition face in terms of what will be termed the 'validity cycle.' Instantiations of the 'validity cycle' are illustrated with two types of paradigmatic research program in the discipline: one established (mindreading, episodic memory, and flexible planning) and one emerging (the concept of death). The diagnosis of the established research programs will create the foundations for testable predictions to be made regarding the development of an emerging one.

In section (2), the question of why the 'validity cycle' emerges is addressed. The section employs a number of challenges facing empirical (human) psychology that have been made in the meta-science literature, in order to make sense of what is happening in the 'validity cycle' in comparative cognitive science. The structuring hypothesis is that weak theory, hypothesis under-specification, and experimental underdetermination interact to repeatedly give rise to instances of the 'validity cycle' in comparative cognitive science. However, each of these problems is analyzed separately, and the limitations of the existing diagnoses on offer are critically discussed within the broader context of the sciences of the mind.

In section (3), a series of recent proposals for addressing issues stemming from the 'validity cycle' are addressed: going bottom-up, using dimensions, formalizing theory, and changing testing. It is argued that these proposals encounter difficulties that leaves them open to falling back into the 'validity cycle'. Refinements are proposed to each of them and brief proposals for how they can be combined are put forward.

Section (4) concludes.

1. The Validity Cycle

Behavioral experiments in a significant portion of research programs in comparative cognition are generally unable to yield decisive evidence from their experimental data. This is at least in part due to the following three interconnected problems; (1) weak theory is tested with (2) underspecified hypotheses, using (3) methods that necessarily result in data that is not clearly interpretable. The higher-level issue that these three interconnected problems result in will be referred to as the 'validity cycle'. The 'validity cycle' (see Figure 4 below) is a model of a cluster of problems that a significant portion of research programs in comparative cognitive science face as they attempt to produce valid experimental work. It repeatedly yields a situation in which it is extremely difficult for comparative cognitive scientists, to obtain clear evidence from the types of experiments they typically perform.

Importantly, it is very much a model. This is because it is intended to represent only an aspect of a target system in order to facilitate a particular type of representation and explanation (Godfrey-Smith 2007). There is no claim here to the 'validity cycle' being able to capture the target system in its totality, or more importantly to being able to capture the entirety of challenges that the discipline of comparative cognitive science faces. That is explicitly not the project.

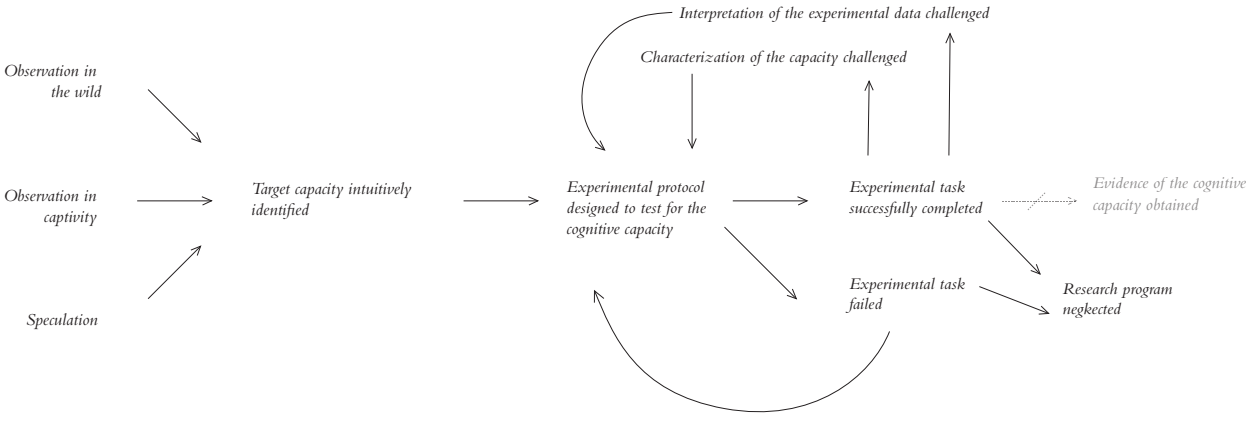


Figure 4. The 'Validity Cycle'

On this model, research programs targeting novel capacities typically begin in one of three ways: through ethological observations, through observations in captivity, or through speculation on the part of the experimentalist, which are sometimes motivated by philosophical analysis. The target

capacity is then intuitively identified and then operationalized as part of an experimental protocol. However, this is typically where things get more complicated. As soon as it is claimed that an experiment provides evidence in support of the attribution of a capacity that the hypothesis claims to test for, both the characterization of the capacity itself as well as the task-capacity relationship are open to contestation.

Colloquially, these claims take one of the following forms, or both: (1) 'The cognitive capacity has been mischaracterized in this experiment, therefore any attempt to operationalize it will fail to yield evidence' or (2) 'The task-capacity relationship is unclear in a way that blocks the possibility of the experiment providing evidence of the presence of a cognitive capacity'. Put simply, contested theory and experimental underdetermination interact to repeatedly render the evidential import of any given experiment relatively inert. Regardless of which line of contestation is taken, the result is the same; a new experimental protocol must be generated. In a variety of truly ingenious ways, comparative cognitive scientists attempt to overcome this problem, primarily but not exclusively, by further refining the behavioral experiments with the intention of clearly isolating the target variable. This takes place alongside projects directed at theory development, although these have been comparatively rarer in the discipline itself. In doing so they inevitably end up confronted with the same type of problem that they started with. From here, the cycle spins and spins until the research program is eventually spit out and is no longer the object of sustained focus.

That the 'validity cycle' captures something significant and concerning about a significant number of research programs in the discipline, or that it is in fact a vicious cycle, might at first seem somewhat surprising. This at least in part because, it seems to have the potential to represent a virtuous iterative process, whereby increasingly strong theory is developed and empirical evidence regarding the status of a cognitive capacity is converged upon, much as it is in research programs in other sciences (again, see Chang 2004). However, time and time again, the virtuous outcome is not achieved, and the iterative process that has been delivered in other scientific contexts, appears to be repeatedly coming up short. This phenomenon is widespread enough (e.g., mindreading, causal cognition, imitation, play, flexible planning, among many other), to pose a general and substantial threat to the discipline's self-defined project of producing cumulative knowledge and

achieving converging scientific consensus around that knowledge. On some accounts, this also potentially poses a threat to its scientific integrity more broadly conceived (Farrar et al. 2019; Penn 2011).[100]

One way to think about this would be the following: If comparative cognitive science was a scientific discipline that was self-consciously and explicitly exploratory in nature, meaning it did not engage in explicit hypothesis testing and did not routinely make strong claims on the basis of its experimental evidence, this might not be cause for concern. This is because all research programs will start from an epistemic standpoint that comparatively impoverished and plausibly exploratory research will represent a promising route towards epistemic enhancement. The problem here is that this is in no way reflects the discipline's conception of itself, how it operates, or the types of claim it generally makes.[101]

In this way, the integrity of the science and its ability to progress can at least be legitimately drawn into question.[102] This is in part because claims in the discipline are frequently presented as possessing far more certainty than the available evidence warrants, and the discipline simultaneously displays no serious sign of revising this situation. In this way, it is very plausible that something needs to change assuming the integrity of the science needs to be restored. However, there is no consensus either in the science of comparative cognition, or in the philosophy that targets the discipline, regarding the precise direction this change needs to move in.

The diagnosis of the three problems above, when considered respectively, are not necessarily novel in character. However, when their interactions are considered holistically, the novel features of the diagnosis emerge in so far as it highlights the multifaceted nature of the problems that a significant portion of research programs in comparative cognition face, without claiming to capture them in their totality. To be slightly more explicit, the 'validity cycle' is distinct from

[100] Integrity in this context does not imply intentional misconduct or fraud. There is currently no evidence that these are systematic issues at the moment in the discipline. Instead, the integrity of a science as it is invoked in this context can be questioned if there are systematic practices that inhibit the advancement of knowledge in the discipline.
[101] If there is any doubt about this, skim the abstracts of the top comparative cognition and psychology journals to see how many of the publications of the discipline are engaged in acknowledged and explicit exploratory research.
[102] To be clear, drawing something into question does not settle in advance how that question will be answered.

diagnoses like the so-called 'logical problem', which are focused on a given research program's ability to experimentally isolate the cognitive capacity of interest, or the problem of underdetermination writ large. Instead, the 'validity cycle' highlights the repeated inability to get a grip on exactly what is being tested for at the level of theory, hypothesis, and experiment, and how these issues interact to produce a multitude of problems for evaluating evidence.

1.1 Instances of the 'Validity Cycle'

Comparative cognition is a large and highly diverse discipline with a rich and complex history. This means that any general statements about it, including diagnoses of its disorders, will invariably admit exceptions. Attempting to deal with these exceptions can quickly turn into a fool's errand or a whack-a-mole style endeavor. Despite this, there are clear analytic and strategic advantages to addressing problems that are plausibly systematic within comparative cognitive science at the level of the discipline, even if every single research program is not afflicted with every single problem.[103] To appropriate a slogan from critical theory; *systemic problems require systemic solutions*. But they also require systemic analysis to get off the ground. It is at exactly this level that the 'validity cycle' as a model of the challenges facing the discipline is situated.

Given this, a series of research programs in the discipline that are clear contenders for being cases of the 'validity cycle' will be discussed. The claim is not that exactly the same thing is happening in each of these cases or that they have encountered and dealt with the same problems in exactly the same ways. In some cases, these will be considered diachronically, in so far as how exactly they develop over time is significant. In others, however, something closer to a synchronic snapshot will be given. The 'validity cycle' as a model will be shown to capture something significant about a research program if the following conditions are met:

1. Attempts to generate experimental evidence of the presence or absence of a cognitive capacity are repeatedly challenged as either (a) not targeting the correct characterization of

[103] Another way, perhaps overly simplistic way to put this might be, if this diagnosis does not capture what is going on in a particular research program, then it should be considered to be out of scope.

the capacity (the theory prong) or (b) not providing a valid test of the capacity, assuming it is properly characterized (the experimental prong).

2. The repetition of the cycle does not clearly constitute an iterative process of convergent and cumulative knowledge regarding the capacity in animals. That is, the broad goals defined by the practitioners of the discipline are not being met.

In the following, a series of research programs that are 'established' to varying degrees will be shown to demonstrate symptoms of being caught in a 'validity cycle': (1) mindreading, (2) episodic memory, and (3) flexible planning. The first two will be given a more detailed analysis than the last, only because they have been the subject of far and away more philosophical and empirical work. Building on this, an instance of what will be labeled an 'emerging' research program will be discussed, which focuses on the (4) non-human animal concept of death research program. The strategy is to show how the development, or lack thereof, in the non-human animal mindreading research program, and the problems that it has repeatedly faced, can predict the development of an emerging program in discipline and the problems that it will face, assuming significant discipline-wide changes are not made (see Table 15. below).

Research Program	Theory Prong?	Experimental Prong?	Validity Cycle?
1. *Mindreading*	Contingent upon problems around representation, among others…	Behavior reading, associative learning? Testing for the task not the capacity.	✓
2. *Episodic Memory*	Proliferation of theories, no clear prospect for convergence.	Just semantic memory? Testing for the task not the capacity.	✓
3. *Future Planning*	Not probed (small research program)	Associative learning? Testing for the task not the capacity.	✓(?) (potentially forming)
4. *Concept of Death*	A minimal definition; small and developing research program.	Data interpretation and encoding unclear. Testing for the task not the capacity.	? (predicted)

Table 15. Instances of the 'Validity Cycle'

Before diving into the case studies in more detail, it is worth making plain what the 'validity cycle' models as not being achieved and asking whether it is reasonable to set this out as a goal. Clear evidence is primarily being presumed to be able arise from a valid experiment, even if this validity comes in degrees as was argued in the previous part of the dissertation (see Part 1 Section 3.2). On a certain interpretation, this might seem to be excessively demanding, however, this need not be the case. A diachronic analysis of research programs achieving this goal will exhibit evidence of this being achieved as a part of an iterative process. Admittedly, this can be difficult to assess because history is long and there are no objective criteria that can determine when an assessment should start.

In addition to this, it has recently been repeatedly claimed that individual experiments in the discipline have been overburdened and that evidential pools included collections of experiments and background knowledge should form an inferential basis.[104] While there is something to this, the strategy can only plausibly gain traction assuming individual experiments are able to provide some degree of evidential clarity. Otherwise, it remains mysterious how a convincing case can be squeezed out of a body of extremely unclear highly underdetermined experiments. In this way, obtaining clear evidence from valid experiments is a reasonable goal that should be upheld, even if individual experiments are ultimately unburdened.

1.1.1 An 'Established' Research Program 1: Theory of Mind

In the broadest sense, mindreading, or the use of a ToM (Theory of Mind), is the ability to represent the mental states of others. Animals do not have a ToM in the same way that adult humans do.[105] There is broad consensus on this point (Shettleworth 2010). Beyond that, however, consensus has failed to establish itself. Even after over 40 years of research, almost every aspect of this research program remains unsettled. The prospects for this situation shifting are currently not particularly promising. If anything, funding cuts, labs closing, ethics regulations regarding the use of great apes in research, and a general environment of wanning interest in the research program, indicate a turn for the worse. While chimpanzees have exhibited seemingly remarkable behavior

[104] This will be discussed in more detail later in this part of the dissertation (See Part 1 Section 3.2).
[105] Mindreading and theory of mind will be used interchangeably.

over a wide variety of experimental conditions that were intended to test whether they possess a ToM, there has never been an experiment or study in animals that was able to yield anything approaching uncontraversial evidence in support of a ToM attribution.[106] Without exception, the research program has produced ambiguous results. From Premack and Woodruff's (1978) initial conative state attribution experiments, to Hare et al.'s (2000, 2001) food competition studies, to Lurz et al. (2018), and Krupeyne et al.'s (2016) implicit false belief tasks, the research program has repeatedly faced the same type of problem: it has failed to experimentally isolate mindreading as variable and to achieve a clearly valid experiment. It will be argued that this represents a prototypical instance of one prong of the 'validity cycle' being repeatedly followed.

That this research program has struggled and faces substantial challenges is not a new insight; far from it. There are two closely related ways in which the severe problem it is facing has been framed. The first holds that the research is sometimes characterized as being perpetually afflicted by what has been termed the 'logical problem' (Lurz 2011); the claim that it is in principle impossible to test for mindreading in animals with the behavioral methods that are typically implemented in the discipline. Lurz's characterization of the logical problem is largely constrained to the perceptual state attribution paradigm. Because of this, he frames the problem as an inability to split the proximate stimulus in experiments that make use of visual perception, in order to infer the distal cognitive state (Lurz 2011). The proximal stimulus in many animal mindreading studies involves the eyes (see, e.g., Hare et al. 2000, Hare et al. 2001, Hare et al. 2003; Flombaum & Santos 2005; Hare et al. 2006; Melis et al. 2006; Kaminski et al. 2006), hands (Kaminski et al. 2004; Tomasello et al. 1994, 1997, 1999), and entire bodies (Povinelli and Eddy, 1996). These proximal stimuli, however, invariably underdetermine any inference that might be made regarding a possible distal representation made by an animal, be it mindreading involving or otherwise.

Vonk and Povinelli's rendition of the problem, by contrast, has a more general scope, and is captured by the idea that for every mindreading hypothesis, there is a complementary behavior–

[106] A definition of theory of mind will not be given here because this is exactly what has been contested and resisted consensus. Despite this, claims to converging consensus program are sometimes made (Halina 2015, Krupenye and Call 2019). However, these are either unwarranted or are evaluating a particular subset of the scientific community. This raises concerns regarding how consensus is evaluated which will not be addressed here.

reading hypothesis (Vonk and Povinelli 2006).[107] In addition to this, not everyone sees this lack of progress as a noteworthy problem. For example, it has sometimes been claimed in the literature that this is a feature of all experimental data, and that this problem is in no way unique to the non-human animal mindreading research program and that it represents no special problem at all (Halina 2015).

However, none of these characterizations have properly captured the problem at hand or proposed a sufficient solution. If the problem facing the research program was just one of experimental underdetermination, that would be one thing. If it was just a problem of hypothesis or theory under-specification that would be another. But the fact that it is both of these, in addition to other issues that are plausibly discipline-wide, creates a set of interacting problems that are distinct and have not been adequately diagnosed in the discussion of the research program thus far. The diagnosis of this problem in terms of the 'validity cycle' is in part intended to be a response to this.

As touched upon above, achieving a valid mindreading experiment using black-box behavioral methods that can yield clear evidence has proven to be extremely difficult for comparative cognitive scientists. This has resulted in an arms race of increasingly complicated and esoteric tasks and experimental protocols. While there has been a consistent struggle to both determine what is being tested for, as well as how to test for it, focus in both the philosophical and the scientific literature has been historically placed on the later. This has been the case since Gil Harman (1978), Daniel Dennett (1978), and Jonathan Bennett's (1978) respective BBS response articles to the Premack and Woodruff's (1978) inaugural study.

Recently, however, philosophical work has seen a reversal of this trend, and with good reason; determining what might count as good evidence of ToM, is dependent on what we take mindreading as a cognitive capacity to be, or how it is characterized. While there has been a shift towards explaining the complexity and diversity of the ways in which we make mental state attributions beyond straightforward beliefs in the adult human program (e.g. Westra 2017,

[107] In an all-too-common instance of sexist injustice, this rendition of the problem is sometimes called 'Povinelli's problem' despite the fact that Jennifer Vonk is the lead author on the paper.

Spaulding 2018), those working on the non-human animal mindreading research program have gone the other way and attempted to specify the details of the capacity with the hopes that doing so will allow it to be more productively operationalized.[108] Both programs have attempted to move towards further specification, albeit in distinct ways.

As alluded to above, perhaps the most neutral way of defining ToM is as the representation of the mental states of others; that is, as a mental representation that targets another's mental representation (a type of meta-representation).[109] For much of the past 40 years, this type of definition of the capacity was thought to allow comparative psychologists to remain agnostic with regard to the details of the capacity, while continuing empirical inquiry, and slowly but surely isolating mindreading as an experimental variable. However, as the research program failed to adequately progress, more specified definitions of the capacity slowly emerged, however, their implications for the experimentalists remained largely unclear.

One slightly more specified and influential account can be found in Whiten (1996), who characterizes, or modeled, the mental states of others as intervening variables. Here, the essence of ToM is a variable that mediates the connection between a variety of situations that an animal observes and the variety of behaviors that it exhibits. This is placed in contrast to 'smart behavior reading', which would make a series of disparate connections between the input (the observed situation) and the output (the behavior that is exhibited). Here, according to Whiten, there will be nothing that the distinct predictions have in common.[110] Accounts stemming directly from this view have been interpreted in at least two ways. The first holds that the intervening variable needs to have 'mentalistic' content to genuinely count as an instance of ToM, while the second holds that any intervening variable that correctly classifies perceptually disparate behaviors into an equivalence class will suffice. This second view has faced difficulties because grouping behaviors that include the representation of cognitive states into equivalence classes is not obviously

[108] Notably this is not a shift that is representational of research programs in the discipline. Most research programs in comparative cognitive science are small in a way that prevents a sustained focus on theoretical development from emerging. Despite any shortcomings of such attempts in the ToM research program, the broad neglect of theoretical projects in other research programs has placed them at an epistemic disadvantage.
[109] I take it that representing a mental state is a minimal form of using it. In this way, the representation of a mental state entails its use.
[110] There is no principled reason to think that this would be the case absent very significant controls and extensive trials.

indicative of the possession of a ToM. This is particularly the case given that the class is not explicitly established on the basis of cognitive representations. Instead, it is established on the basis of a proxy that is reducible to mere behavioral representations absent an intervening variable, or some other unobservable, but importantly not a cognitive property.

Lurz (2011) has argued that the ability to classify perceptually disparate behaviors into equivalence classes according to the cognitive state that they have in common, notably must have mentalistic content. Within the dominant perceptual state attribution paradigm, this would include the ability to group behaviors that are encountered in a variety of situations into a single class according to the cognitive states that they share.[111] And importantly for Lurz, they must be represented as such. This has been a site of somewhat heated contention in the literature.

For example, at points, Povinelli and colleagues have prevaricated on this issue, sometimes even within a single paper (Penn and Povinelli 2007), but they have more recently settled on the claim that the intervening variable must have mentalistic content in order to count as a genuine case of ToM (Penn & Povinelli 2013). Andrews (2016) has refused the distinction and argued that it should be abandoned because knowledge of the mental just is knowledge of the behavioral in so far as the unobservable mental qualities can only be accessed through observations of behavior. However, there is good reason to think that the distinction cannot be so easily discarded.

If each of these conceptions were clearly defined, that is if it was known what it might specifically mean for a mental state variable to have content, let alone mental content, this might merely be a verbal dispute about what to call mindreading. However, given that this is not the case, further specifying the relevant mental content as, for example, a type of perceptual content as is done in level-2 perspective taking tasks, would be a promising avenue to gaining traction in a research program that has repeatedly struggled to do so over an extended period. This is particularly salient because all parties involved in the debate are attempting to give a directly theorized realist account

[111] It is a bit unclear why this could not just be done according to cognition coarsely construed.

of the mental lives of non-human animals, in contrast to a mere a project that is modelling focused.[112]

Definitions of mindreading are typically not interpreted as 'models' that are useful for making predictions, even if models at certain points feature in that inferential process. It seems to be relatively clear that both scientists and philosophers want to know what is really going on in the heads of these animals. Do they have a ToM or not?

The difficulties of specifying ToM as a scientific theory have become even more apparent considering what has been called the semantic problem(s) for animal mindreading research (Buckner 2014). Buckner's starting point is that there is a consensus in the ToM literature around construing both the empirical challenge and the underlying theory in terms of mental representation. However, what these mental representations amount to has never been adequately probed. Given this, Buckner claims that the animal ToM research program faces a particular instance of the problem of distal content that is rooted in philosophical debates around representational mental content, rather than solely a vexing empirical challenge that is rooted in experimental underdetermination and the inability to separate out and test a mindreading variable.

While naturalized theories of content generally make use of covarying features in an environment to obtain distal content, such as *tree*, that is over and beyond the mere proximate content, such as *green leaves and brown bark*, these theories achieve this in distinct ways that are consequential for the experimental data in the ToM literature. Within the context of the animal mindreading research program, this means being able to distinguish the proximate content of observed behavior from the potential distal content of a mental state variable. Despite the fact that representation is at the core of almost all characterizations of the cognitive capacity in question (e.g., ToM), and at the core of most cognitive capacities in animals for that matter, the debate has generally avoided specifying a theory of representational content, which results in a broadly underspecified theory of mindreading even in cases where there are attempts to nail it down. This

[112] While models are typically thought of as indirectly exemplifying features of a target in a way that often abstracts and simplifies it, direct theorizing cannot sustain the introduction of falsehoods or deviations from accuracy. For more on this see Frigg and Hartmann (2018).

is a problem because interpretations of the experimental data vary with the theory of representational content that is adopted. In this way, experimental underdetermination remains a problem on Buckner's account, but the lack of consensus regarding a theory of representational content feeds and exacerbates that problem.

Considering this, Buckner, advocates the adoption of Dretske's informational theory of representational content to solve this problem (Dretske 1997). On this account, the co-variation of a cognitive state with the cognitive state of the mindreading target is sufficient to count as ToM. However, this move is largely stipulative and it does not address the underlying problem that it introduced. In this way, it is not clear why this account should be adopted, particularly if what counts as ToM comes apart under various theories. Committing to a particular account too quickly or absent sufficient evidence will potentially lead the research program further astray. In this way, the problem is displaced onto giving an account of mental representation more broadly, which poses particular difficulties because the problems facing representational content are not limited to the distal-content problem (e.g., the problem of error, and the multiple problems of indeterminacy for representational content). In this way, converging upon a theory of mental representation would be a significant theoretical step forward in this research program. However, such a convergence does not appear to be within grasp and there is emerging consensus that the only thing that can be agreed upon is that no existing account of mental representation is currently adequate (Godfrey-Smith 2006).[113] There are substantial and deeply entrenched philosophical debates that stand in the way of this being easily obtainable. Of course, this is not to say that settling on a theory of mental representation is a full-stop prerequisite for the empirical study of ToM in animals, but rather that until such a theory is settled upon, significant and deep theoretical problems regarding the interpretation of experimental data are likely to continue to linger.

The turn in the ToM debate away from the minimal account of theory of mind, towards more fine-grained accounts seems like it might have been promising step in the direction of theory development. However, rather than leading to a more detailed account that would make the problem more empirically tractable and allow experimentalists to work pointedly towards

[113] Similar issues arise in the shift from representations to concepts given that literature is similarly contested.

achieving more clearly valid experiments, this turn has resulted in a proliferation of more general problems. Again, the 'validity cycle' keeps spinning and the criteria set out at the start of this section are plausibly met. Both the experimental prong and the theory prong of the 'validity cycle' have been repeatedly taken in this research program, and this has not obviously resulted in an iterative convergent process. Instead, there is a credible case to be made that the research program has instead repeatedly struggled to progress. And again, this situation is exacerbated and solidified by several other factors such as funding cuts, ethics regulations, small sample sizes etc.

These issues are in a certain way up-stream from the experimental prong, and hypothesis-based challenges that the research program has more broadly faced. Given this, any claim to evidence of the presence or absence of ToM in animals is ultimately premature because it would be based on several assumptions about the theoretical commitments implied in a ToM as well as the task-capacity relationship, both of which are yet to be settled. This places the research program in a difficult position in which it is largely unclear, even to those who are involved in the highly theoretical aspects of the debate, exactly what the phenomenon is that is being tested for. In other scientific disciplines, this might be a standard moment of calibration that is, iteratively arrived at. However, neither the flood of empirical work nor additional theoretical work has obviously contributed to the type of convergence that has been repeatedly sought after. It is unclear if it eventually will.

1.1.2 An 'Established' Research Program 2: Episodic Memory

Subjectively, memory might seem like a monolithic entity. However, on most accounts, it is composed of a variety of enmeshed systems, all of which are theorized to have distinct functions and neural architectures. Mainstream taxonomies of long-term memory cut it into two kinds: non-declarative and declarative (Michaelian 2017).[114] Non-declarative memory is thought to involve a type of informational encoding that allows a subject to use representations sub-

[114] Short-term memory and working memory bridge the declarative or non-declarative divide, which highlights the complicated and intersecting structures that the various memory systems are thought to have. These will be left out of the current account.

personally.[115] This includes procedural memory that underlies skills and habits, non-associative learning, and perceptual learning, among others (Ness and Calabrese 2016). By contrast, declarative memory is thought to generally involve a type of informational encoding that allows a subject to explicitly use representations on the personal level.

Both semantic and episodic memory are thought to be types of declarative memory. Attempts to conceptually demarcate these two kinds of memory, and to experimentally isolate them as variables, have occupied the bulk of episodic memory research program in humans over the past 50 years. Originally, the distinction was theorized in terms of the different sources of the information that each kind of memory was thought to involve (Tulving 2002). While semantic memory was thought to encode and store information pertaining to general objective facts, episodic memory was originally theorized to encode and store information pertaining to the 'what, when, and where' of subjective experience (Tulving 1972).

As experimental work that was inspired by folk and intuitive definitions began to emerge, it became increasingly plausible to some that episodic and semantic memory were discrete memory systems with independent functions and neuroanatomies. This conception of episodic memory allowed for specified hypotheses to be formed that were more sensitive to specific aspects of 'recollective experience' (Tulving 1983). For example, the difference between noetic awareness, or the phenomenal experience that accompanies thinking about the world, and auto-noetic awareness, or the phenomenal experience that accompanies remembering or recollecting, was ignored in early experimental work, as were the 'where' and 'when' aspects of episodic memory (Tulving 1983). Early neuroscientific methods sought to alleviate some of the vagueness inherent in interpreting behavioral data, but the dearth of empirical evidence and the persisting lack of conceptual clarity, remained obstacles for establishing episodic memory as a major neurocognitive memory system (Tulving 1986). These issues have led many to question the utility and feasibility of conceiving of episodic memory as a separate neurocognitive system at all (Glenberg 1997).

[115] The personal/sub-personal distinction refers to two different types of explanation. It can be traced back to Dennett (1969). Personal level explanations takes place at the level of whole individuals. These often invoke terminology in the space of belief-desire psychology. Sub-personal explanations do not target whole individuals and can invoke neurophysiological features. According to Dennett, both types of explanation are psychological. Although the personal/sub-personal distinction has been highly influential the boundaries of each type of explanation have been, and remain, highly contested.

More recently, Tulving has attempted to give a more precise definition of episodic memory and has argued that the core of the capacity can be found at the intersection of the concepts of self, autonoetic awareness, and subjectively sensed time (Tulving 2002). It is not entirely clear, however, if this definition is in fact more precise or progress inducing than the 'what, when, and where' definition, in so far as the same types of debates that ensared the previous definition have re-emerged.

In addition to this, PET and FMRI studies have resulted in the **h**emispheric **e**ncoding/**r**etrieval **a**symmetry model, or the HERA model, on which the left prefrontal cortex is predominantly correlated with the encoding information into what is thought to be episodic memory while the right prefrontal cortex is predominantly correlated with retrieving information from what is thought to be episodic memory. There have also been clinical observations that have been claimed to support Tulving's more recent conception, but their evidential status is largely anecdotal and weak (Tulving et al. 1988).

Despite the seemingly substantial developments that have been made over the years, the episodic memory research program in humans is still faced with many of the same set of problems that it encountered at its outset; it remains unclear what is being tested for or how to go about testing it. The most fundamental and persisting problem is that there is no consensus regarding what episodic memory is, and the introduction of a dizzying number of conceptions of the capacity has placed restrictions on any empirical undertaking. Part of the problem lies in the fact that the clashing and increasingly specified definitions of the capacity on the table prime the research program, if it can still be referred to as a unified research program, for a variety of verbal disputes, which in turn makes behaviorally testing for the capacity a convoluted project at best. These clashing categorizations are in part a consequence of the fundamental fact that there are competing accounts of the function of episodic memory.

Past oriented accounts typically appeal to the fitness that is incurred by an organism that can exploit information about subjectively experienced past events, rather than being confined to the

facts about the objective world that are characteristic of semantic memory.[116] Among other things, this is thought to allow for the reinterpretation of past information considering subsequently acquired information in a way that would not be thought to be possible while remaining within the bounds of semantic memory. This view has been criticized by future oriented accounts, which hold that there is no obvious adaptive benefit to recalling past experiences that will never occur again, and that this should constrain our conception of episodic memory. These critics hold that the function of episodic memory is ultimately to imagine future events. Episodic recollection and future-directed imagining on this account are thought to be a part of one unified system (Clayton & Wilkins 2018). Moreover, metacognitive accounts hold that autonoesis allows for the emergence of a sense of subjective certainty (Klein 2014) or the ability to distinguish between remembering and imagining (Mahr & Csibra 2018). If episodic memory is indeed to be at least in part functionally defined, it remains unclear how that should be done, particularly as definitions of the capacity continue to proliferate. This theoretical challenge poses a serious threat to the research program.

In addition to these disputes around the function of episodicity, assuming that episodic memory involves something like mental time travel, whatever that turns out to mean, there are still several crucial details that need to be worked out for the empirical program to plausibly gain traction. For example, the commonsense conception holds that recalling the past is in fact a process of simulating or imagining it. On this view, episodic memory is a creative process that is centered on the right type of informational preservation (Michaelian 2016). What this amounts to in detail is spelled out in several conditions, such as the content-matching condition and the factivity condition, among others, all of which are up for dispute under this approach and are ultimately meant to allow for the separation of remembering from merely imagining. Against this, the causal conception holds that there must be a causal connection between the originally experienced representation and the recalled one for there to be a genuine act of remembering (Martin and Deutscher 1966). Given that content generation is compatible with memory on this account it

[116] The term objective here is meant to create a constrast with the subjective aspects of episodic memory. For example, John might have first person subjective memories of cycling through the Scottish Highlands. They might also remember that Eddy Mercx won 34 individual stages at the Tour de France. However, he did this before John was born. The latter is a fact about the objective world that John knows. The former is something that John experienced subjectively.

remains a question how much content generation can be present if there is to be an act of genuine remembering and how memory can be seen as reliable if it permits a level of construction (Michaelian 2016). Finally, the simulation conception abandons the causal condition all together and holds that merely imagining is sufficient for episodic recollection (Michaelian 2016). Obviously, there are a lot of choices to be made here, all of which hold strong consequences for any comparative cognitive science-based project seeking to identify episodic memory in animals.

The animal research program inherits these problems while simultaneously introducing many of its own. This has made testing for episodic memory in animals, regardless of which account is settled on, very difficult. Unlike some other cognitive capacities such as mindreading, which is a research program that was originally tested in animals, episodic memory was long thought to be unique to humans, and the empirical work for the first thirty years of the research program reflected this assumption (Suddendorf and Busby 2003). This was at least partially because experimentally demonstrating conscious recollection was thought to require language, which limited testing to humans (Tulving 1983; Griffiths et al., 1999). However, a series of studies done by Nicky Clayton's lab in the early 2000's on episodic memory in scrub jays jump started the animal program (Clayton and Dickerson 1998). Since then, there have been attempts to test for episodic memory in scrub jays, honeybees, hummingbirds, pigeons, rats, gorillas, and monkeys, among other animals.

Again, what it thought to set episodic memory apart from semantic memory on some accounts is the ability to phenomenally 're-experience' a past event that is lush with conscious subjective detail. Because this phenomenal aspect of the episodic memory features so centrally and given the fact that comparative cognitive scientists have historically refrained from testing for conscious states for a variety of reasons (Andrews 2020), it was sometimes claimed that the phenomenal aspect of episodic memory could not be tested for. Given this, Clayton claimed that in fact, it was only episodic-like memory was being tested for in her experiments, meaning that they were only interested in the content of the information available to the animal, because without having access to an animal's subjectivity, which in humans is done primarily through verbal reports, nothing could be said about its subjective experiences. This strategic agnosticism, however, puts it at odds with several definitions of the capacity that have been implemented in the human program.

Clayton's food caching paradigm in scrub jays has been highly influential.[117] In the original experiment, scrub jays were given the option of caching either worms or nuts at various caching trays. When given the opportunity to return to these caches, they generally return to the cache with worms. This is because scrub jays prefer worms to nuts. However, worms quickly decay and become inedible in a way that nuts do not, and scrub jays were able to learn to be responsive to this fact. If the worm was cached long enough for it to decay, they would return to the nut cache, thereby reversing their original preference. From this, Clayton and Dickerson (1998) claimed that scrub jays had episodic-like memory according to Tulving's 'what-where-when' criteria. That is, they were sensitive to the type of food (what), its location in the tray (where), and the temporal duration of the caching (when).

However, it is not clear from this definition that episodic-like memory is in fact distinct from semantic memory, given that semantic memory can contain information pertaining to what, where, and when something took place (Klein 2013). If this is the case, it is also unclear in what sense the memory is even 'episodic-like' or why these results should not just be seen these as an impressive or rich instances of semantic memory.

Interestingly, many comparative cognitive scientists working with various species have embraced the episodic-like definition. However, the question remains whether the scrub jays are merely encoding information in some sort of objective sense, which on some interpretations would place it in squarely in the domain of semantic memory, or whether they are informationally encoding an experienced event that is subsequently recalled in an episodic form. Moreover, it is important to emphasize that it is separate question whether scrub jays, or any of the animals that are being tested, are conscious, which is notoriously difficult to test. It seems very plausible that animals could have subjective experiences while simultaneously not encoding the information episodically, whatever that turns out to mean. In addition to this, even if an event is episodically encoded it could at least conceivably not be done consciously. However, if episodic memory requires something like autonoesis, a first step towards testing for this more demanding criterion might be to show that an animal has some sort of consciousness experience simpliciter. There

[117] The reader will remember a brief discussion of this in Part 1 Section 1 of this dissertation.

have been recent proposals to test for this more robust conception of episodic memory that includes phenomenology that was originally advocated by Tulving (Templer and Hampton 2013, Boyle 2020).

Despite its various developments, the episodic memory research program forms a plausible instance of the 'validity cycle' according to the criteria presented above. This is, however, not to claim that the non-human animal episodic memory research program is epistemically situated in a way that is identical to the mindreading research program. Clearly these are distinct and have importantly distinct challenges facing them. However, the non-human episodic memory research program has been clearly and repeatedly become very familiar with the theory and the experimental prongs of the 'validity cycle'. Moreover, the theoretical challenges facing the research program seem particularly salient for the project of assessing whether the cycle is vicious or virtuous. The non-human animal episodic memory research program is very much a live one, so any declarations will necessarily be contingent upon that fact. However, there is good reason to think that the cycle, in this instance, does not clearly constitute an iterative process of convergent and cumulative knowledge regarding the capacity. This is particularly given the case that it remains highly unclear what exactly is being tested for, and in the case of the animal program, what separates the capacity from semantic memory. This is not to say that there have been no important advancements in the research program over the past fifty years, which would be a clearly untenable position, but rather that, for the moment at least, the path toward convergent and cumulative knowledge remains one that is significantly removed.

1.1.3 An 'Established/Emerging' Research Program 3: Flexible Planning

A multitude of human activities require planning for events that are removed from our current sensory information. Performing these activities has been thought to require a distinct cognitive capacity that has been coarsely referred to as 'flexible planning'. To take a fictional human example, Odysseus knew that at some future point he would hear the sirens' song and feel compelled to jump overboard to be with them, which would inevitably lead to his death. For this reason, he lashed himself to the mast before his ship approached the sirens, even though at that very moment he was not feeling the pull of the sirens.

While human cases seem to be intuitive and easy, particularly when they are fictional, by contrast, identifying genuine instances of 'flexible planning' in animals has been difficult and highly contentious. Some have claimed that 'flexible planning' must involve a novel behavior. This criterion was introduced as an attempt to control for certain confounding alternative explanations. These explanations have typically involved the reduction of the behavior that has been claimed to be an instance of 'flexible planning' to reinforcement learning alone or ones that claim that the behavior is based on a motivational state that is distinct from the one the animal is currently in (Suddendorf and Busby 2005). While this conceptual modification initially promised to bring clarity to some of the experimental work, novelty remains a highly disputed and underspecified concept in comparative cognitive science and identifying genuine uncontroversial instances of it have proven difficult.

Other attempts to further specify the concept have included claims that 'flexible planning' should be 'flexible' across domains or be 'domain general'. This has been thought to include situations in which two or more domains, such as a technical domain or a social domain, are able to be planned across (Kabadayi and Osvath 2017). While the precise definition of a domain is left unspecified, the tasks are usually coarsely distinguished in a way that makes hard boundaries between domains unnecessary.

Another complication for the research program comes from the fact that planning is not necessary for all anticipatory behavior, particularly those behaviors that are reliant upon innate and rigid cognitive mechanisms. If the animal in question is thought to have ecological or behavioral predispositions that limit it to domain specific planning, flexible planning is typically not theorized as being a driver behind such behavior, and it is remains unclear in what respect genuinely flexible planning is involved. In many ways, these theoretical decisions reflect limitations in the empirical methods.

For a long time, it was thought that non-human animals were not capable of flexible planning (Atance and O'Neill 2001; Gulz 1991; Noble and Davidson 1996; Corballis and Suddendorf 1997). However, at least three shifts in adjacent research programs in the late 1990's contributed

to this view being substantially questioned.[118] **First**, the introduction of the Bischof-Kohler hypothesis, which claimed that animals are trapped in the present in so far as they are unable to recall a past or imagine a future state that they are not currently in (Bischof-Kohler 1985), started to garner increased attention. **Second**, there was an increase in work on tool use in non-human animals, which was thought to potentially require the use of 'flexible planning' (Whiten et al. 1999). And **finally**, research on episodic memory in both humans and non-human animals advanced and started to centralize mental time travel and episodic foresight, which were theorized as being connected to 'flexible planning' (Clayton and Dickinson 1998). These, and other factors, laid the groundwork for the first flexible planning experiments in animals (Shettleworth 2010).

For example, Naqshbandi and Roberts (2006) performed an experiment on rats and squirrel monkeys, in which they were given a choice between two varying amounts of food. As the food choice was presented to the animals, their water bottle was removed from the cage. The thought was that both animals would prefer the larger amounts of food, but that this would also make them thirstier than they would have been had they chosen the smaller amount. When the animal choses the smaller amount of food, the water bottle is returned more quickly than it is when the larger amount is chosen. This was thought to test for the animal's ability to foresee and plan for their future state of thirst. While the rats showed no effect and barely preferred the larger amount of food to the smaller amount on all trials, the squirrel monkeys reversed their baseline preference for the larger amount of food after 6 trials. Naqshbandi and Roberts (2006) claimed that this provided evidence against the Bischof-Kohler hypothesis and evidence for the existence of flexible planning in animals. However, it was subsequently argued that delayed reinforcement or punishment was responsible for the preference shifts because of their gradual character and that flexible planning had nothing to do with this behavior at all (Shettleworth 2010). The Bischof-Kohler hypothesis was still alive and well.

More recently, the debate has come to a head of sorts, in a purportedly breakthrough study done by Kabadayi and Osvath (2017), in which they claimed that ravens were able to flexibly plan in

[118] There are at least two ways of interpreting how these research programs have fed into each other. On one interpretation evidence of closely related cognitive capacities has been combined to generate progress across research programs. Another is that evidence from experiments in a variety of context is so vague and underdetermined that it can be used to support a wide variety of hypotheses.

domains that were not natural to them. This was particularly compelling because the food caching domain, around which many of the studies in corvids had previously been oriented, was claimed to be an illegitimate test of 'flexible planning' because they were thought to be reflective of a specific caching adaptation. In their experiment, the ravens were tasked with selecting a tool or token that would allow them to obtain a future reward. In the majority of cases, the corvids were able to make the correct selection even in the presence of a distractor reward (i.e., a reward that was less appealing than the one they would obtain when selecting the tool or token). From this, Kabadayi and Osvath (2017) concluded that ravens are capable of 'flexible planning.'

However, an alternative explanation of their successful performance on the task was waiting just around the corner. Redshaw et al. (2017) argued that the corvid's success on the task could be accounted for with associative learning. Given that the ravens had to be trained to use the tools and tokens in the task over the course of 35 trials, Redshaw et al. argued that there was good reason to prefer the 'lean' hypothesis that associative learning and not 'flexible planning' better accounts for the ravens' successful completions of the task.

Associative learning represents one of the most plausible success stories of comparative cognitive science, with the research program being a seemingly cumulative one. While the ideas behind associative learning are very old, the past 150 years have seen an explosion of models of various types of associative learning. Its comparatively outsized 'success' has frequently come to flood research programs that have aimed to study other cognitive capacities, and when it comes to accounting for much of animal behavior, its status as an explanatory model is still very much up for debate (Lind 2018). There might also be reason to think that its power as a model is at least in part due to its flexibility (Dacey 2023). Given this, Osvath and Kabadayi (2018) contested this associative explanation of the data and argued that their experiment sufficiently controlled for the possibilities that the alternative explanation provided.

In this way, the 'flexible behavior' research program can be usefully modelled with the 'validity cycle.' New experiments are performed with the hopes of being able to plausible evade associative learning explanations (e.g., Boeckle et al. 2020), however, this has not seemed to convince the critics. In this way, the cycle continues, however, in a manner that is again distinct from the

mindreading and the episodic memory research programs. The challenges that the 'flexible behavior' research program faces are distinctly empirical at the moment, however, this does not mean that the theory prong of the 'validity cycle' is necessarily off the table. Assuming that the debate continues, which is in no way guaranteed, there is good reason to think that research will be challenged on this front as well.

1.1.4 An 'Emerging' Research Program 4: the Concept of Death

It has been assumed by many in the animal cognition literature that non-human animals cannot possibly have a concept of death because it is simply a concept that is too rich and demanding (Regan 2004; DeGrazia 1996; Harman 2011; Bradley 2015). This intuition is a strong one but generally it has remained an intuition.[119] Monsó's (2019) project is to lay the conceptual foundations that would allow for this question to be empirically tested and to help move the research program beyond mere speculation. Given that there is substantial variation across cultures and history with regards to how death is conceptualized, she starts by presenting seven necessary conditions for the possession of a 'full-blown' concept of death. The idea here is that while there is substantial variation in concepts of death, these conditions are meant to represent the 'core' of the concept.

> (1) *Non-functionality*: death implies the cessation of all bodily and mental functions.
> (2) *Irreversibility*: dead individuals cannot come back to life.
> (3) *Universality*: all living things, and only living things, die.
> (4) *Personal mortality*: death will also apply to oneself.
> (5) *Inevitability*: eventually, all living things must die.
> (6) *Causality*: death occurs due to a breakdown in the bodily functions.
> (7) *Unpredictability*: it is impossible to know in advance the exact timing of death. (Monsó 2019, pg. 7)

[119] Speculatively and intuitively, this might have something to do with the way in which death occupies a simultaneously central and peripheral role in our lives. We will never experience our own deaths and it is often difficult to come to grip with the death of others in our lives. Why this is the case is ultimately an empirical question, and it will likely admit cultural and historical variation.

From this full-blown conception, Monsó argues that (1) non-functionality and (2) irreversibility, highlighted in blue above, are the crucial components of death. This serves as the basis for a minimal concept of death that can be used to test for the possession of the concept in animals. Monsó (2019, pg. 9) writes…

> "A creature can be credited with a minimal concept of death once she classifies some dead individuals as dead with some reliability, where 'dead' is understood as a property that pertains to beings who:
>
> (a) are expected to have the cluster of functions characteristic of living beings, but
> (b) lack the cluster of functions characteristic of living beings, and
> (c) cannot recover the cluster of functions characteristic of living beings."

Monsó's strategy here relies on two steps. The first is to identify the essential features of a full-blown concept of death given that there is so much cultural and historical variation in the understanding the concept. The next is to pare this down to its two essential features. Monsó's case is even more interesting because she explicitly addresses the meta-philosophical features of her account. Rather than treating death as an open ended cluster concept, Monsó argues that adopting a set of necessary and sufficient minimal conditions is favorable because it will purportedly allow for a clear attribution of a minimal concept of death to animals than something like a cluster concept, which would permit attributions if fewer than all of the criteria are instantiated.[120] However, this is of course contingent upon the acceptance of the criteria, and as has been seen in the repeated instances of the 'validity cycle', this is one of the primary sites of contestation. Moreover, if the full-blown concept is not first accepted, it is very possible that the minimal concept will not be as well. They are tightly dependent on one another.

While there is some research on whether animals possess a concept of death in comparative cognitive science, the research program is largely an emerging one, meaning it has not been afforded much attention when compared to established ones such as mindreading and episodic

[120] For an example of a cluster concept that has been used in a psychological context, see Bermudez (2005).

memory, both of which have been subjected to substantial theoretical dispute. Given this, and the features that it shares with the mindreading research program, as well as those that are directed at studying similar sophisticated cognitive capacities that have been highlighted so far in Part 2 of the dissertation, there is the possibility of drawing attention to the problems that will be encountered before a large number of resources and energy is poured into experimental work; that is, before heated disputes get off the ground, and before inflated claims can be made, contested, and captured by the 'validity cycle'. The focus here will primarily be on the theory prong of the validity cycle, but the ways in which experimental and observational work is complicated by the proposed theory will also be highlighted.

The overarching problem here can be located in the fact that the theory proposed by Monsó is underspecified.[121] This is a problem because underspecified theories have direct consequences for the interpretation of observational data, and the formation of hypotheses that are tested in experiments. The three steps of the account will be worked through before it will be considered in more detail.

The (a) part of the definition can be immediately probed, and it can be asked what it means to 'expect' that something will possess a cluster of functions typical of a living being. Several features of this aspect of the definition are highly ambiguous. Moreover, why should it be thought that expectation should be built into the definition at all. If a corpse is recognized immediately as a dead one, in many contexts at least, no expectation of functions that are typical of a living being should be present. In certain circumstances, such as if one were to come upon a member of a social group that has died this might be part of experiencing them as being dead, but this need not be built into the concept of death at all. Of course, understanding what it means to be dead might plausibly entail some understanding of what it means to be alive, but this does not imply that understanding death should have anything to do with the expectation of life.

This leads directly into problems with (b). What does it mean to expect functions as being typical of a living being as opposed to functions being typical of that individual at a specific period, or a member of a group, or a species? Why think that there is any need for such a high-level concept?

[121] This issue will be filled out in more detail in the subsequent section.

While Monsó is sensitive to species variation, highlighting the differences between the cessation of function between animals that are cooperative versus those that are competitive, her account leaves open the possibility that the loss of function is typical of an individual and not life itself. And more fundamentally for this aspect of the definition, it is unclear what it means to recognize a function.

And finally (c), what it means to not be able to recognize that an organism will not be able to recover a function is multifarious. There are certain cases that Monsó highlights like sleeping, in which function purportedly ceases but is then recovered. However, it is unclear if this should count as a cessation of function at all, but rather engaging in just a different type of function. While sleeping individuals are not typically highly mobile, they are generally responsive to external stimuli, in the way that dead individuals are not. Functions arise, pause, and cease constantly for a variety of reasons. What this brings out is that it is not clear why an understanding of permanence should be required or what exactly it entails. It seems like the cessation of functions that are typical of living being should be enough to cover what this point is intended to capture. It is clear, however, that the temporary cessation of function should be ruled out.

All of this combines into an overarching problem that is familiar from many of the research programs that have been subject to the 'validity cycle'. Namely, the theory is problematic in a way that creates downstream consequences for hypothesis formation, experimentation, and the interpretation of evidence. And what type of evidence does Monsó think should count towards an attribution of the concept of death? She lists the following sources of potential evidence: varied behavior towards corpses, unhygienic/maladaptive behavior towards corpses, different treatment of corpses vs. asleep individuals, investigative behavior towards corpses, aggressive behavior towards corpses, caring behavior towards beings with limited functionality, mourning behavior towards corpses, eventual ignoring or abandoning of a corpse, age or experience-relative difference in behavior towards corpses. This is an extensive list that virtually covers the range of possible reactions that one could have in response to a corpse in an environment. This means that in so far as one can respond to significant changes in individuals in an environment, there will be some reason on this account to think that the animal has a concept of death. If these sources of evidence were able to be clearly operationalized, that would be one thing. For example, if

mourning behavior could be reliably identified in animals as such, this would surely provide compelling and persuasive evidence for the possession of a concept of death. The problem is, however, that behavior is wide open to interpretation, and any behavioral coding practice that might be implemented in these sorts of observational studies will be highly underdetermined. This type of observational evidence is not able to distinguish the hypothesis that an animal has a concept of death from the multitude of competing alternative hypotheses that are able to equally well account for the observed behaviors. Monsó is rightly sensitive to the anecdotal nature of this type of evidence and argues that observational studies must be complemented by experiment lab-based work.[122] She even has a proposal for how to do this.

The study of the concept in animals obviously has the potential to veer quickly into ethically dubious territory. For this reason, Monsó's proposal is artefact based, and would involve an animal first understanding the function of the artefact. Then the artefact would cease to function, and the animal would have to recognize non-functionality and irreversibility. According to Monsó this would provide indirect evidence of the "necessary cognitive requirement" for a concept of death and would demonstrate the ability to "process the crucial sub-components" of such a concept (Monsó 2019; pg. 16). However, there are deep theoretical assumptions here regarding cognitive capacities, the relationship between them, and generalizability. For example, assume that non-functionality and irreversibility were sufficiently operationalized in a particular experimental context, and it is in no way clear that this could be the case. It needs to also be assumed that non-functionality and irreversibility are conceptualized at this high level of abstraction and not in a way that is highly context bound. These are exactly the issues that have plagued a multitude of research programs in comparative cognitive science such as the causal cognition research program.[123] Distinguishing between the possession of a context bound behavior and a richer understanding of a higher order concept that would permit an inference from behavior involving an artefact to behavior involving a living or dead organism will undoubtedly prove to be extremely challenging.

[122] To be clear, the anecdotal nature of this type of work would pose substantial challenges, however, as highlighted this is not the core theoretical challenge for the research program.
[123] This research program will be extensively discussed later in this part of the dissertation.

Given these issues, a prediction can be made that the research program will follow the path laid out in the 'validity cycle.' It is very plausible that both prongs will be repeatedly taken. This will include a series of experiments in which both the concept of death and the experimental results are repeatedly contested in a way that superficially resembles a process of calibration that can be observed in sciences that are more devleoped. However, the progress that would be entailed by calibration will not be clearly won and the research program will be, for a variety of reasons, eventually neglected. Again, this is a prediction, and it may come to pass for reasons that are distinct from those that are found in the research programs that were previously discussed, or not at all. Only time will tell.

2. Weak Theory, Hypothesis Under-specification, and Experimental Underdetermination

Why do the issues that the 'validity cycle' captures arise repeatedly in comparative cognitive science? The question is a difficult one that will not yield a monolithic answer. However, in building on the examples laid out above, three tightly connected issues will be introduced that serve as plausible systematic contributors to the repeated emergence of the 'validity cycle'; weak theory, hypothesis under-specification, and experimental underdetermination.[124] To be clear from the outset, there is no direct line or necessary connection between these three problems. However, there is reason to be concerned about potential contingent dependencies that have arisen between them. In a certain way, weak theory holds the potential to systematically lead to underspecified hypotheses, which holds the potential to systematically lead to experimental underdetermination. Despite these holistic concerns, each problem will be given individual treatment in so far as doing so provides a unique entry point to confronting possible contributions to the repeated emergence of the 'validity cycle' in comparative cognitive science.[125]

[124] Because the difference between hypotheses and theories may seem initially opaque, a number of intitial distinctions can be made. Hypotheses are directly tested in experiments and produce experimental data. Among other thing, they dictate what types of control conditions need to be implemented in that particular experiment. Theories, by contrast explain or predict phenomena (Bogen and Woodward 1988). Theories are not directly tested in experiments. This will be particularly clear in the discussion of overarching theories below.

[125] This analysis will also provide an outline of what it means to be an immature science as it is referred to in the title of this dissertation. This will be further discussed in Section 2.4 of this part of the dissertation.

2.1 Weak Theory

The past couple of years have seen the emergence of what has come to be called the 'theory crisis' (e.g., see Oberauer & Lewandowsky 2019), which has been repeatedly framed as a neglected source of the so-called 'replication crisis.' In the most general sense, the idea behind the 'theory crisis' is the state of the theories that are being probed in disciplines such as psychology, cognitive science, and others that have been hit by the so-called 'replication crisis', is at least in part responsible for the emergence of the crisis itself, and that many of the replication failures should in fact be seen as unsurprising.

This concern about the general state and quality of theory in the broader psychological and cognitive sciences is not an especially new one. Interestingly, however, the issues that have been raised have received almost no attention in comparative cognitive science. Meehl (1978), for example, started sounding the alarm over 40 years ago with the observation that many theories in psychology emerge and never develop or reach an advanced stage of maturation. Instead, their development takes the form of fads, with a siloed focus on low hanging fruit, only to then disappear as quickly as they surfaced. Directly building on this, almost 15 years later, Lykken (1991) argued that theories in psychology in general are neither affirmed, refuted, nor substantially developed. Instead, they fade from the attention of the core of the discipline only to be replaced quickly with the next 'hot' topic. In a similar vein, Wachtel (1980) highlights the way in which psychology has failed to progress at the rate of the hard sciences, primarily due to an overemphasis on experimental work in the field and the general neglect of theoretical work.[126] On Wachtel's account, this is a consequence of long-standing incentives and pressures in the discipline. In this way, pessimism about the state of theory and its advancement in the psychological and cognitive sciences, should be seen as nothing new.

What is new is that these problems are now fed directly into the larger sense of panic caused by the onset of the replication crisis, and the many proposals for reform and revolution that have

[126] The disctintion between the 'hard' and 'soft' sciences is invoked often and almost never made precise. The implication typically being that the harder a science is, the more rigorous it is. The natural sciences, including physics and chemistry for example, are typically taken to be hard, while sociology and psychology are typically taken to be soft. For more on the history of the disctintion between the 'hard' and 'soft' sciences see Shapin (2022).

come with it. This has subtly shifted the diagnosis of the problem from being one that primarily concerns the limitations of the development on the discipline that characterized pre-2000's criticism, to one in which the status of theory has resulted in evidence in these disciplines being viewed as potentially unreliable and wildly misguided in a deep and systematic sense. This has placed a renewed emphasis on the role of theory in the sciences of the mind, in a way that has given the task of remedying it a sharp sense of urgency. For some, the alarm has reached a fever pitch (e.g., Barrett 2010, Shettleworth 2010, Hemelrijk and Bolhuis 2011, Vonk and Shackelford 2013, Mikhalevich et al. 2017, among many others). However, this new-found attention has also brought forth the fact that the terms of the diagnosis and the debate that it gave rise to remain unclear, unsettled, and deeply contested. This is at least in part because the previous diagnoses were not well worked out. For example, although Meehl (1978) identified several problems with the function of theory, as well as the limitations of null hypothesis statistical testing, he did little in the way of showing how theory should be constructed and cultivated in a way that would bypass the problems that had been identified.

This is an even more pressing problem in the contemporary literature where there is no clear or agreed upon definition of what counts as 'weak' or 'latent' theory or more crucially how to go about developing it into theory that is 'strong'. In illustrating and alluding to the problem in human psychology, Muthukrishna and Heinrich (2019) use the example of a well-known series of experiments from Bargh et al. (1996) in social psychology; a field that has been strongly affected by the replication crisis. This set of experiments tested whether American undergraduate students walk slower when they are reminded of the elderly through unconscious representations of 'slowness' and is meant to function as a more general test of an automatic behavior priming effect. In challenging this work, they essentially pose two questions for it: (1) what could this experiment contribute to a general theory of human behavior and, more implicitly and (2), how should such research be interpreted?

Theory in disciplines that fall under the psychological and cognitive sciences has sometimes been thought to be weak because these disciplines lack an overarching theory, background theory, or general theoretical framework that could serve to constrain more local theorizing.[127] The general

[127] These terms should be largely seen as synonymous and interchangeable.

idea is that an overarching theory functions to guide instances of theorizing in a principled way. An example of an overarching theory can be found in Muthukrishna and Heinrich (2019) who propose dual inheritance theory as a potential theoretical framework for psychology. Dual inheritance theory holds that genetic and cultural evolution interact to shape biology and behavior.[128] Through the use of formal models that make specific predictions regarding both genetic and cultural evolutionary processes in line with dual inheritance theory, the thought is that certain cognitive phenomenon such as learning can be studied in a more directed and constrained way.

While the presence of an overarching theory provides constraints on research, it is important not to exaggerate the role that such a theory is able to occupy. Constraints play an essential role in disciplines like physics, but they are often couched in several assumptions and in no way result in a simple, uncontroversial, and straightforward output as some philosophers of science in the replication literature, or many in the metascience literature, have implied. It is reasonable to assume that stronger constraints will support stronger inferences, however, no discipline is fully constrained and many of the problems facing the science of the mind will be persistent in the presence of an overarching theory, albeit in a potentially less severe form.

Given this, (1) is a somewhat curious question for Muthukrishna and Heinrich (2019) to pose. This is because the work done by Bargh et al. (1996) which claimed to show behavior being automatically triggered by features of the environment, was directly mobilized in support of dual process theory; a plausibly general theory of human cognition and behavior (Kahneman 2017).[129] This means that the problem cannot be that the research is not directed at an overarching theory, but rather that Muthukrishna and Heinrich (2019) question the ability of this 'type' of experiment to contribute to such a theory or be structured by it in any sort of plausible or productive manner. Interpreted in this way, their challenge is more intelligible and perhaps even a plausible one. In its current form, however, Muthukrishna and Heinrich take this fact to be obvious, despite its highly controversial character. Moreover, it is still unclear what exactly makes this theory 'weak'. Given

[128] They also allude to Homo economicus as a potential general theoretical framework that exists in economics Another example of an overarching theory might be the standard model in physics.
[129] Bargh et al. (1996) situate this experiment as part of a body of experiments that are thought to demonstrate unconscious priming effects.

that they refrain from critiquing this work in detail, the reader is left to fill in the details on their own. This neither makes their somewhat ungenerous critique particularly convincing, nor does it specify what exactly is defective in the theory or what would make it better. The precise character of the problem is left unclear. Despite this, a more full-blooded critique can be made and further specified. It can be boiled down to the following three points:

1. The relationship between representations of 'slowness' and the elderly is left intuitive and underspecified.
2. It is unclear that a representation of 'slowness' is in fact what is triggered in the experiment and that is responsible for the effect.
3. The relationship between slowness and the 'elderly stereotype representation' is left intuitive and vague.

Given these problems, it should come as unsurprising that the research has been subjected to several replication experiments that could be interpreted as successes and/or failures, and that this has repeatedly been the site of dispute (Aarts et al. 2002; Cesario et al. 2006; Doyen et al. 2012). In part, this is because the causal structure of the effect that is intended to be tested in these experiments is opaque, vague, and underspecified. This has also meant that the authors' claims regarding the results of these experiments requires several inferential leaps that are neither clearly specified nor justified. In this way, it is reasonable to doubt the ability of these experiments to provide evidence in support of (i) a particular theory of automatic behavioral triggering that is rooted in an elderly stereotype or (ii) an overarching discipline structuring theory such as 'dual processing theory'.

Importantly, this is not an experimental problem that merely gaining more data alone will be able to resolve. Instead, a more specified theory is required. In addition to this, it also does not seem to be a problem that an overarching background theory would obviously be able to solve either. Perhaps the thought is that such underspecified theories would not arise in a discipline with a highly constraining background theory, but that of course cannot be taken for granted. This becomes particularly clear, considering Muthukrishna and Heinrich's proposal that dual-inheritance theory should be adopted as the overarching theoretical framework for the sciences of

the mind and behavior. Given their argument, this would not be particularly attractive. It is one thing to argue that evolutionary biology, including dual-inheritance theory, should function as a constraint on the sciences of the mind and behavior; this is a largely uncontroversial and innocuous claim. It is an entirely different point to argue that dual-inheritance theory can simply be transplanted into the sciences of the mind and behavior and thereby resolve the problems with theory that have emerged in these disciplines. The problem here is that there is no reason to think, particularly given the case that Muthukrishna and Heinrich make, that it would be able to provide sufficient constrains on the research to do anything like the work that they are asking it to do; namely evaluating theories and experiments in psychology.[130] While dual-inheritance theory has provided limited direction in some research contexts such as those that have been focused on tool use or social learning, in the vast majority of cases, particularly in comparative comparative cognition, it will leave inquiry significantly unconstrained.

Another take on the problem of theory in the sciences of the mind, comes from Borsboom and Markus (2013), who argue that weak theory and the absence of theory in the psychological and cognitive sciences has meant that there is a relative overreliance on statistical methods in these disciplines. They illustrate this and the consequences of a 'dearth of theory' with a thought experiment in which psychologists are tasked with building a bridge in a way that they currently perform inquiry in their discipline. According to Borsboom and Markus, this would involve innumerable iterations of bridges, undirected regressions targeting a variety of features of them, and complex factor analyses. On this approach, it would plausibly take thousands of years to empirically establish Newton's second law (F = ma). The thrust of the argument is that working without theory, or working with extremely 'weak' theory, makes the sciences of the mind slower and less productive than their more counterparts that are working with 'strong' theory. In a similar way, Collins' (1985) claim that weak foundations leave more room for doubt and uncertainty, hits directly on the role of theory in these disciplines.

To take up another example, without having to actually build an airplane one can be modelled in various ways so that it can be predicted how it would fly in a variety of conditions without ever

[130] In this way it should also be seen as unsurprising that most theories will come out as underspecified and/or nonsensical from the perspective of dual inheritance theory simply because it cannot bear on them.

leaving the armchair. Robust theories, such as those of ram drag or mass flow rate, and the background theories that they exist against, are in a large part responsible for the (1) precise predictions, (2) explanations, and (3) manipulations that make aerospace engineering so powerful. This is part of the reason that one can wake up at 7:00 AM in London and fall asleep in a hotel room in downtown Tokyo 16 hours later. Of course, this theory did not emerge from thin air, and it is inert outside of a long an iterative experimental process that has driven these theories forward. The argument is, however, that once strong theoretical development is established, which occurs necessarily within a broader experimental context, it can be mobilized to yield more expedient results than approaches that are less theoretically driven.

The core issue with this type of analysis is that this does not offer a clear way forward. Fundamental physics is different from the sciences of the mind in a variety of ways, so highlighting the distinct manners in which they have managed to proceed, or the differences between the two in doing so, is not always going to be an obviously productive strategy. In this way, the problem is similar to the one presented by Muthukrishna and Heinrich's proposal; simply importing a framework from another seemingly more mature discipline is not obviously going to do the trick (see Table 16 below). Practitioners in the sciences of the mind can be envious of the epistemic situation of other disciplines all they want, but this will not obviously lead to better theory. And understandably, this might be a hard point to come to grips with given that more established or mature sciences often function as guides for emerging or immature sciences in a wide variety of ways, particularly in the philosophy of science. Even if this analysis does not offer a clear way forward, it might offer a beginning of a diagnosis of one of the sources of difficulty that the sciences of the mind have faced.

Author	**Science used as model?**	**Appeal**	**Difficulty**
Muthukrishna and Heinrich (2019)	Evolutionary biology- dual inheritance theory	Clarifies the appeal of an overarching theory.	Not action guiding or sufficiently constraining; evolutionary biology / population genetics does not solve the problem

Borsboom and Markus (2013)	Fundamental physics	Takes a step towards clarifying how an established/strong theory works	Not action guiding or sufficiently constraining; psychology cannot simply mimic the trajectory of physics

Table 16. A summary of the positions reviewed in this section of this part of the dissertation.

Building on this, another way in which theory in the sciences of the mind has been thought to be weak is located in the fact that they are formed in a largely intuitive manner. At first glance, this is not obviously a problem. Traditionally in the philosophy of science, the context of discovery and the context of justification have been thought to be importantly distinct (Reichenbach 1938). As Popper (1959) extensively argued, theories are arrived at in a wide variety of ways, and rightly so. For example, Kekule is reported to have first seen the benzene ring in his dreams.[131] In this lineage in the philosophy of science, how a theory is generated has virtually nothing to do with its status, strength, and/or explanatory power. Instead, what is important is how a theory is tested and verified. This view certainly has a strong and intuitive appeal, but there is also something more complicated going on here.

If high quality or 'strong' theories were consistently being dreamed up perhaps there would not be reason for concern. After all, when things are going right, there does not seem to be much reason to look more closely at the process that underlies the generation of such theories. However, when there appears to be a systematic issue with the theories that are being tested in a discipline, there is plausibly good reason to look more closely at how such theories are being generated even if these two 'contexts' are ultimately distinct from one another. For example, if a child were responsible for generating all of the theories that were to be tested in discipline of theoretical physics, this would have no bearing on the verification of these theories or the ability to test them. However, this would severely constrain the space of theories that were up for being tested or verified and would clearly have severe consequences for the discipline and its ability to progress by its own lights.

[131] There is apparently substantial controversy concerning whether Kekule in fact dreamed the structure. However, the larger point is not contingent upon settling this debate. For more on this see Strunz (1993).

An example of a theory that has been claimed to have been intuitively arrived at can be found in the paradox of choice literature (Scheibehenne et al. 2010); the theory that humans prefer less choice and that having less choice will increase well-being. As Muthukrishna and Heinrich (2019) claim, but do not provide evidence for, this theory is rooted in the researchers' own experiences while shopping. They write (Muthukrishna and Heinrich 2019; pg. 224), "This approach has many disadvantages: from WEIRD psychologists using their WEIRD intuitions to generate the alternative hypotheses, contexts, and bounds of the phenomena to the WEIRD participants used to test hypotheses to the specificity of the question itself." The claim is in part that theories that are arrived at intuitively tend to remain familiar and have the potential to seem obviously true to researchers; they are plausibly more subject to biases such as confirmation bias. This fact is thought to make them potentially more immune to critical inquiry. In addition to this, theories that are intuitively arrived at are likely to remain highly limited in the type of content they contain. In comparative cognitive science, concerns around how theories are generated have arisen because cognitive capacities are claimed to be conceived in terms that are thought to be derived primarily from an experimenter's intuitions around what they would do were they in the animal's position in an experiment (Penn 2011).

Another way in which a theory might be considered weak has to do with the level of explanation at which it is situated and how this level determines the types of constraints that can be placed upon it, and in turn, how such a theory is able to be specified. In comparative cognitive science, black-box style behavioral experiments have created significant barriers to further theoretical specification (Sober 2015). This has meant that the target of investigation is frequently severed from potential sources of evidence at other levels of explanation that might be able to place principled constraints on hypotheses. While comparative cognitive science is saturated with background theory from other disciplines such as evolutionary biology and psychology, the level of explanation on which many cognitive capacities are conceived, has historically made it difficult for this background and adjacent theory to constrain or shape hypotheses in a way that would allow them to be clearly specified. To be clear, there is nothing necessary about this situation. That is, nothing automatically blocks comparative cognitive scientists from developing stronger theories by drawing on work in adjacent disciplines and using them to generate progress.

However, the challenges posed by connecting other sources of evidence up with hypotheses as they are typically stated in the discipline is a very significant one, particularly when this evidence is being derived from a discipline that is in the early stages of maturation. For example, connecting lower-level evidence from the neurosciences up with higher level folk psychological concepts has proven extremely difficult, and often only the most general and vague of constraints from the research is able to play this type of structuring role.

Given this, a theory might be thought of as 'weak' or 'latent' is the degree to which it (i) cannot yield hypotheses that are clearly able to be confirmed or disconfirmed to a significant degree or (ii) it is unable to lead to its own further development towards being able to explain, predict, and/or manipulate facts about observables.

In concluding, there are at least five senses in which a theory might be considered to be 'weak.' Its contribution to the 'validity cycle' has not afforded a singular diagnosis. For a summary of these senses, see Table 17 below. Most importantly this analysis brings out the fact that what constitutes so-called 'weak' theory remains highly contested in the literature and needs further fundamental work.

Diagnosis	**Problem**
1. *Lack of an overarching theory*	Inquiry remains necessarily unconstrained
2. *Distorted generative process*	The wrong set of theories might end up placed up for confirmation
3. *Wrong level of explanation*	Limitations are placed on possible constraints and specification
4. *Inability to yield confirmatory hypotheses*	Theory cannot play its intended epistemic role
5. *Inability to yield further development*	Theory cannot play its intended role in knowledge development

Table 17. A summary of the variety of senses in which a theory has been thought to be weak in the theory-crisis literature.

2.2 Hypothesis Under-Specification

A related, albeit distinct, problem emerges at the level of hypothesis formation. Theories and hypotheses are sometimes treated synonymously. However, there are important distinctions to be made between the two. Hypotheses are tested directly through the experimental process, but the same cannot be said of theories.[132] The shape of these hypotheses directely dictate particular experimental protocols and control conditions. The results of these experiments produce data which can be used to confirm hypotheses. In turn, these are used to confirm or disconfirm theories. A consequence of this is that theory places restrictions on the space of hypotheses that can be generated. In this way, the status of theory and the status of hypotheses are deeply dependent on one another. However, there is a distinct problem with hypothesis under-specification that is independent of the problem of 'weak' theory discussed above.

One way to begin thinking about hypothesis under-specification in a more general sense is to look at the use of the term in linguistics. There, a predicate is thought to be underspecified if it has more than one distinct meaning (Muskens 2000).[133] Imagine planning to meet a friend at the café on campus at noon. It is relatively clear what is meant by café, but if there are 20 different cafés on campus, you will know where not to go, but you still do not know where the meeting at noon will be. In a similar way, a hypothesis is underspecified if there are at least two partially exclusive interpretations of its content. In practice, in the science of comparative cognition, this can play out in several ways. In some cases, there is an agreement regarding the definition of a capacity that is implicitly underspecified. In others, researchers operate with multiple explicitly conflicting underspecified hypotheses.[134] Regardless, in all cases, when an underspecified hypothesis is being tested it is appropriate to ask; exactly which hypothesis?

Hypothesis under-specification becomes a problem when it blocks experimental evidence from being obtained. It can directly result in the underdetermination of both hypothesis and theory

[132] Another way to bring this out is to look at the way in which there cannot be something like an 'overarching hypothesis.' This would seemingly require divorcing hypotheses from the experimental context in which they are tested.
[133] Under-specification is importantly distinct from similar phenomena such as vagueness, which is characterized by permitting borderline cases.
[134] It is not that the debate is unsubstantive. Imagine Lab A is using definition X while lab B is using definition Y, and they are engaged in a dispute regarding how to define the thing they are testing for. Instead, in this case, Labs A and B are using a shared definition, and they are not sure if it picks out X or Y.

selection. In practice, this problem is given attention at varying moments in the experimental process, but in comparative cognition, this often occurs after the experiment has been carried out, the data has been obtained, and even as claims to evidence in favor of the presence of a capacity, or in support of a theory, have already been made.

Imagine, for example, that researchers are testing a hypothesis H. This hypothesis affords at least two distinct interpretations: H_1 and H_2.[135] An experiment is designed and carried out that is meant to be a test of H. Critics then questioned this experiment's ability to test H in a variety of ways. Among other points they argue either that H_2 is what was being tested for all along, that H_2 is what should have been tested for all along, or simply that H_1 was not being tested for in the experiment. Regardless of the particular criticism, the experiment will likely struggle to provide uncontroversial evidence of the either interpretation of the hypothesis. This is because it is unclear which interpretation of the hypothesis should be tested for in the experiment. Importantly this will determine how things like control conditions in the experiment should be run. This type of under-specification can leave researchers particularly exposed if the cognitive capacity being investigated turns out not to be captured by the higher level of characterization at all.

Note, that the problems that arise here are at least conceptually independent of the theory that the hypothesis is designed to test, and that even if the theory was 'strong' and well-formed in each of the senses described above, this phenomenon of a hypothesis being underspecified can still theoretically arise. It will be less likely, but it is surely still possible. Additionally, this does not mean that hypotheses must, or even can, be maximally specified, or that this would even be intelligible as a goal in every case. Further specification may be irrelevant, or it simply may not be possible to achieve given the current state of the evidence. In addition to this, if increasingly fine-grained interpretations of hypotheses are nested in the right way, underspecifying a hypothesis can be a constructive strategy for breaking into a nascent research program that is in the process of development. This, however, does come with tradeoffs in so far as underspecifying a hypothesis limits the evidence that can be obtained, and assuming it can be tested with a substantial degree of validity the possibility of controlling for confounds in these situations becomes very challenging.

[135] It is assumed here that H_1 and H_2 are not both nested under H, or that they cannot be lumped together under H.

In sum, the problem of hypothesis under-specification arises at least in part, when the hypothesis cannot play the role in the empirical process that it is designated to. That is, underspecified hypotheses are a problem when they are not able to generate experiments that are able to yield data that contribute to their confirmation or disconfirmation.

2.3 Experimental Underdetermination

Even if the problems of weak theory and hypothesis under-specification were able to be resolved or bypassed in comparative cognitive science, the problem of experimental underdetermination would remain a threatening one. However, given the current status of the psychological and cognitive sciences, these problems plausibly feed directly into the problem of experimental underdetermination. In the context of comparative cognition this typically involves the inability of a behavioral task to provide clear evidence of cognitive capacity and the inability to sufficiently control for extraneous variables.

Underdetermination poses a general problem for the philosophy of science, where the task of theory selection is thought to always be underdetermined by the evidence (Bonk 2008). This version of the problem is not one that comparative cognitive science should be expected to solve, and it should in no way be expected to be a benchmark for the science broadly construed. While attempting to infer a cognitive mechanism from behavior alone might be thought to inevitability require underdetermination, underdetermination should be framed as coming in degrees, and part of advancing the science of comparative cognition will necessarily involve reducing underdetermination as much as possible. However, when framed in this graded way, most instances of it in the discipline at the moment are seemingly very high.

At this point, an important distinction needs to be brought out. The relationship between data and theory is clearly distinct from the relationship between data and hypothesis. This has typically not been acknowledged or at least made explicit in the literature on underdetermination in the philosophy of science. While the task of theory selection is one that is underdetermined by experimental data, particularly in the use of null hypothesis statistical testing, it is the task of hypothesis selection that is first underdetermined by the data. In this way, the underdetermination

of the task of hypothesis selection is primary and should be given greater and more explicit emphasis, particularly in the literature on the science of comparative cognition.

One way to cast the problem of underdetermination can be found in Deborah Mayo's severity principle. This idea that has its origins in the work of Karl Popper and his concepts on falsification and corroboration. On Popper's account, experimental data is said to have been able to provide evidence in support of a theory or hypothesis only in so far as it is able to survive serious criticism. Mayo (2018) extends this line of reasoning and argues broadly that error probabilities are required to assess how well probed experimental data is, which she calls severe testing. On this line of reasoning, a test is severe in so far as it is (1) internally valid and (2) statistically significant. An internally valid experiment (1) is set up to properly to test the relationship between the dependent and independent variables and can eliminate extraneous variables.[137] This entails having the ability to produce falsifying evidence. Additionally, a test is statistically significant (2) if the p-value is below the significance or threshold level. Here then, the relevant type of underdetermination is present to the degree to which the test can be considered severe; primarily due to a failure to meet the internal validity criterion.

In some ways, this approach is bootstrapped off the more traditional conception of the underdetermination in the philosophy of science, which is commonly traced back to the Duhem, who writes the following within the context of a discussion of physics:

> *"In sum, the physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses; when the experiment is in disagreement with his predictions, what he learns is that at least one of the hypotheses constituting this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed."* (Duhem 1954 [1906], pg. 187)

In its most general construal, this is what is typically referred to as the Duhem-Quine Thesis. What follows from this is that it remains unclear whether any given set of experimental data in fact confirms or disconfirms a given theory. However, this construal glosses over the fact that the

[137] For a more substantial discussion of internal validity refer to part one of this dissertation.

underdetermination of the task of hypothesis selection underlies this problem. This is because for any hypothesis that is being tested the cause of the success or failure of the experiment could lay either in the target hypothesis or the innumerable auxiliary hypotheses that are being simultaneously tested. Jumping straight to theory in this way muddles the problem.

Type of underdetermination	***Definition***	***Theory or hypothesis***
1. Constrastive	Predictive success underdetermines the task of the theory selection	Primarily theories
2. Holistic	A given hypothesis can never be tested in isolation	Primarily hypotheses

Table 18. Two types of underdetermination and their targets

As Stanford (2017) highlights, what falls out of the Duhem-Quine Thesis are in fact two distinct types of underdetermination: contrastive and holistic (See Table 18 above).

Contrastive underdetermination emphasizes the way in which predictive success underdetermines the task of the theory selection. This is because there are innumerable many other theories that could have also predicted the results. This could be due to the ways in which theories or hypotheses can be slightly altered so that they closely resemble each other, or theories that are remarkably distinct from one another but predict the same results. In the most general sense these are theories that are live in so far as they are typically thought to be empirically equivalent to one another. This type of underdetermination is called contrastive because when theories are compared to one another, they are parked on a level evaluative playing field.

In a similar but not exactly equivalent way, contrastive underdetermination can also apply to the task of hypothesis selection when considering the experimental data that has been produced. This occurs in cases in which the data that is produced in an experiment is inadequate for deciding between the null and the alternative hypotheses in cases involving null hypothesis statistical testing. This occurred repeatedly in the non-human animal mindreading research program which was discussed above. The difference between theory and hypothesis targeted contrastive underdetermination seems to lie in its pervasiveness. While theory directed contrastive

underdetermination might be an ever-present feature of empirical inquiry, hypothesis directed contrastive underdetermination does not seem to be, particularly in null hypothesis statistical testing.

Another type of underdetermination can be found in **holistic underdetermination**, which emphasizes the way in which a given hypothesis can never be tested in isolation. When a theory is used to generate hypotheses that are tested in experiments, we need to simultaneously test a battery of other auxiliary hypotheses. This includes things like methods, instruments, sample size, representativeness, as well as the hypothesized role that various features of the environment have in testing the relevant effect. This type of underdetermination becomes clearly relevant when the success or failure of any given experiment is interpreted. If the effect was shown to be significant, it cannot be stated with certainty that this was due to the mechanism that was hypothesized to be present or some other assumption that is captured in one of the innumerable auxiliary hypotheses that are also being tested. That is, a 'positive' result might have been obtained because the instrument malfunctioned, or the sample was not representative, etc… The exact same thing applies to 'negative' results, where this type of underdetermination is more often invoked within the context of replication. Importantly, holistic underdetermination is not obviously resolved by repeating the target experiment and confirming the auxiliary hypotheses one by one. This is because each experiment requires the introduction of additional, and ever proliferating auxiliary hypotheses, to say nothing of the problems posed by contrastive underdetermination.

These two types of underdetermination, of course, do not exist in isolation. That is, they interact with one another and co-exist within any given empirical context. The word holistic here should be taken seriously. This means that *any* assumption that could potentially feature in the testing of the target effect can be characterized as an auxiliary hypothesis.[138]

While contrastive underdetermination has traditionally focused more clearly on the role of theory, holistic underdetermination placed more focus on the role of auxiliary hypotheses in producing experimental data. However, these were still targeted at the task of theory and not hypothesis

[138] This might introduce an additional type of underdetermination regarding how we characterize and individuate auxiliary hypotheses, assuming that this can be done in innumerable and potentially unconstrained ways.

selection. Despite this, the relationship between data and hypothesis is a unique one that should not be collapsed into the task of theory selection. By glossing over this inferential step and the subsequent step of inferring theory selection based on evidence obtained through hypothesis selection, a potentially important site of difficulty is lost from view. Holding these two steps apart is crucial for understanding an empirical process that is shaped by underdetermination and the possibility of navigating it.

This variety of underdetermination all adds up to a particularly sticky epistemic situation for sciences like comparative cognition where the characterizations of the relevant phenomena are highly contested. Moreover, this is a discipline where claims are routinely confronted with alternative explanations for the successful results of black-box behavioral experiments, and this is compounded in cases where underspecified hypotheses are on the table. In this way, the problem of underdetermination, in the broadest sense, highlights just how difficult clarity in this context is to come by.

2.4 Summary

In the most general sense, the core problem lies in the fact that every aspect of the empirical process is highly unconstrained; this spans the sciences of the mind (see Table 19 below for a summary of this section). The point of spelling things out in this way has been to clarify how crucial each step of the empirical process is, and how loose these connections can be in the science of comparative cognition. This should provide an initial perspective of the status of the research conceived more holistically as in a certain sense constituting an 'immature' science. That is, if the general lack of development leads to weak theory, underspecified hypotheses, and experimental underdetermination in a way that does not allow these experimental moments to fulfill their epistemic function, then there is reason to think that the science is immature to some degree. Moreover, while these explanations might seem highly abstract there is a fair amount of conceptual confusion in the comparative cognition literature that uses the terms like theory and hypothesis interchangeably, which have distinct roles in the empirical process.

In philosophy, there is seemingly always another distinction to be made, independent of its apparent utility. However, the point in introducing these distinctions is that doing so allows details of the problem to be more precisely addressed, and in doing so the factors that repeatedly contribute to the emergence of the validity cycle will be able to be more adequately intervened upon. Although these might at first seem like trivial distinctions, they have clear consequences for diagnosing the problem at hand. Subtle confusions have been shown to be highly consequential, such as in the collapsing of statistical and substantial hypotheses (e.g., Dacey 2023, Bausman and Halina 2018). The more this situation can be addressed at a high level of abstraction so that it can trace down into particular contexts, the better.

Contributor to the validity cycle	***Contribution***
1. Weak theory	• Lack of an overarching theory • Distorted generative process • Wrong level of explanation • Cannot yield confirmatory hypotheses • Inability to develop
2. Underspecified hypotheses	• Struggle to produce experiments that evidentially distinguish the hypotheses
3. Experimental underdetermination	• Contrastive and holistic underdetermination limit both theory and hypothesis selection

Table 19. Summary of the contributors to the validity cycle.

3. Ameliorating Proposals

With some of the systematic problems facing the science of comparative cognition on the table, focus can now be turned towards attempts that have been made to address and overcome these problems. Given the diagnosis in the previous section it will be assumed going forward that some significant aspect of the discipline needs alteration; that things are not all right. In this section, four distinct proposals that have been put forward are analyzed. Each proposal attempts, in its own way, to advance the science of comparative cognition by moving the research beyond the issues discussed previously in this part of the dissertation (Part 2 Section 1).

The **first** involves taking a bottom-up approach to the study of animal minds. The **second** involves breaking the capacity that are studied into dimensions. The **third** involves adopting formalization techniques in the discipline. The **fourth** involves shifting the way experimentation is done towards attempting to identify signatures of cognitive capacities.

It will be argued that each of these approaches has significant deficiencies. After identifying these deficiencies, it will be shown how modified versions of each of these proposals can be placed in a complementary relationship with one another in order to forge a promising path forward.

3.1 Go 'Bottom-Up'

On some diagnoses, the problems that the discipline of comparative cognition is facing are a result of what is called the top-down approach, which is conceived as loosely aligning with the dominant methodology in comparative cognition that attempts to derive cognitive mechanisms almost exclusively from behavioral research.[139] Against this, a bottom-up approach has been advocated by a number of authors as a promising way forward.

However, both the problem that is being addressed and the proposed solution are unclear in a way that undercuts their respective pragmatic value. In this way, not only are there a variety of positions that fall under the labels top-down and bottom-up, but these various positions themselves have remained highly underspecified. This means that two challenges remain unmet in order for this ameliorative proposal to gain traction: First, what distinguishes a top-down from a

[139] So-called top-down and bottom-up approaches are abundant in the sciences and engineering. They can be found in disciplines such as chemistry, computer science, and ecology, among many others. In the creation of nanostructures in chemistry for example, the top-down approach starts with a macroscopic structure that it reduces to a nanoscale through a process of subtraction such as photolithography in the manufacturing of computer chips (Iqbal et al. 2012). The bottom-up approach, by contrast, starts with atoms or molecules and builds them up into a nanostructure through a process such as molecular self-assembly (Dahman 2017). Another instance of top-down and bottom-up approaches can be found in artificial intelligence. In this context, the top-down approach understands intelligence in terms of the manipulation of representation or symbols, independent of neural structures. This is also known as GOFAI ("Good Old-Fashioned Artificial Intelligence"), or symbolic AI (Haugeland 1989). The bottom-up approach attempts to understand intelligence by creating artificial neural networks that are modelled after structures found in the human brain. This is also known as connectionism (Hawthorne 1989). In many of these contexts, top-down and bottom-up approaches are in a winner-take-all style competition with one another, with the dominant method of the respective discipline or research program at stake.

bottom-up approach? Second, what are the features of a bottom-up approach that make it a potentially progress-generating remedy?

In the following, a novel account of five distinct senses in which the top-down and bottom-up approaches are implicitly and explicitly construed in literature is given: (I) anthro- v. zoocentric (starting point), (II) anthro- v. zoocentric (goal), (III) complexity v. simplicity, (IV) special v. conserved, and (V) whole-system v. sub-system. The variety underlying the terms 'top-down' and 'bottom-up' has been substantially overlooked in the debate and has resulted in the potential of the proposal being blocked or simply not taken up. In this section, the focus will be initially placed on describing the various senses in which the terms have been used rather than prescribing a singular way in which the terms top-down and bottom-up should be interpreted. This means that throughout this section a strong pluralism is adopted with regards to the normative question of what types of content should fall under the terms top-down and bottom-up, both within individual disciplines as well as across them.[140] This pluralism is rooted in a broader terminological diversity that is common in interdisciplinary and multi-disciplinary scientific contexts.[141] However, this pluralism should not simply be left undifferentiated. By distinguishing these multiple senses from one another genuine progess with regards to these proposals can be achieve.

While the hierarchy implied in top-down and bottom-up is not explicitly stated in any of the five senses, they can be most straightforwardly distinguished from one another by determining what the labels 'top' or 'bottom' bind to. The differences between them are subtle at points and the five senses sometimes overlap, and co-exist simultaneously, especially in their respective cases for adoption. However, clearly demarcating them for one another is a prerequisite for meeting the second challenge presented above. While some have the potential to become error inducing, particularly when adopted in isolation, each sense of top-down and bottom-up also holds the potential to avoid misleading commitments that have introduced substantial problems to the

[140] For an example of how unclear the terms top-down and bottom-up are in the predictive processing research program, a context where clarity is often assumed to exist, see Rauss and Pourtois (2013).
[141] In the philosophy of science, the terms top-down and bottom-up and sometimes strongly associated with certain interpretations of the mechanistic tradition (see e.g. Craver 2007). In this section, there will be no attempt to constrain usages of the top-down and bottom-up terminology to these conventions.

discipline. This means that the lessons that can be derived from the respective approaches are, in principle, reconcilable with one another.

3.1.1 Anthro- v. Zoocentric (Starting Point)

As Eaton et al. (2018, pg. 1) highlight, there are many, "[…] top-down approaches that begin and end with a focus on humans." These will be labeled anthropocentric top-down approaches and they come in two versions: starting point and goal. The anthropocentric top-down (starting point) approach is centered on the human and is implemented when the characterization of cognitive capacities that are the target of inquiry in comparative cognitive science are derived in some substantial sense from cognitive capacities that are thought to be present in humans.[142] These are research programs that have attempted to advance by "applying top-down strategies and interpreting animal behavior through a human-centric lens (Eaton et al. 2018; pg. 1)." Moreover, they argue that these, "top–down approaches begin with a concept of a capacity clearly defined and identified in a species, such as humans, and subsequently direct researchers to seek evidence for the same capacity in otherspecies (Eaton et al. 2018; pg. 3)."

This approach is routinely adopted because cognition has often been probed in humans in ways in which it has not been in other non-human animals. For example, as mentioned above, episodic memory was studied for close to 30 years in humans before it was studied in animals. Historically, the capacity was assumed to be demanding in a way that non-human neural architecture simply could not support (Tulving 2002).[143] This meant that, upon its inception, the animal research program had a wealth of evidence about the capacity from the human program to draw upon. This has made adopting it as a starting point particularly appealing. De Waal and Ferrari (2010, pg. 203) make a similar point in their discussion of the imitation research program and show the way in which a "top-down definition" was adopted which resulted in a situation in which only humans were thought to posess true imitation.

[142] The actual cognitive capacities present in humans are often exaggerated. Instances of this exaggeration have been labelled anthrofabulation. For more on this please

[143] This does not mean that these cognitive capacities are always well-understood or even well-defined in the human research programs. Nevertheless, on this approach, it is thought that the evidence upon which any inquiry into could be based clusters in our species. For more on this see Buckner (2013).

Just like the anthropocentric top-down approach, the zoocentric bottom-up approach comes in two versions: starting point and goal. On the first version (starting point) the characterization of cognitive capacities that are the target of inquiry in comparative cognition should be primarily oriented around the species under investigation. As Eaton et al. (2018, pg. 6) aruge, "A bottom-up approach would be to assess natural behaviors in the given species and design experiments that further probe these natural behaviors or abilities." They go on to argue that they would be able to do this without being anchored in a human cenetered capacity.

This approach is in part rooted in specific difficulties that the anthropocentric top-down (starting point) approach has faced. As argued by Vonk & Shackelford (2013), many top-down approaches overly emphasis continuity and in doing so lack sufficient evidence as to why certain patterns that would be thought to be indicative of the presence of a cognitive capacity would be present in the relevant target species at all. For instance, when tasks that are thought to test for a cognitive capacity in one species are transplanted onto another, without taking the characteristics of the species under investigation into account, a misleading cognitive profile of that species can emerge, particularly if researchers are not sensitive to the difficulties of negative existential claims. In addition to this, reliance on exaggerated conceptions of human cognitive capacities can distort cross-species comparisons (Buckner 2013). Particularly egregious cases of this plausibly occur when experimentalists rely too heavily on intuitions regarding what cognitive capacities would be involved in successfully completing a task (Penn 2011, Barker and Povinelli 2019), as was touched upon briefly in the previous section.

3.1.2 Anthro- v. Zoocentric (Goal)

The second sense of the anthropocentric top-down approach is 'goal oriented' and holds that the capacities that are studied in animals should be used to provide the minimal criteria for characteristically human capacities that are typically framed as being 'full-blown'. This version of the anthropocentric top-down approach is 'goal oriented' in so far as the goal is to understand the evolution of human cognition, even if animals are being simultaneously studied. This approach is highlighted by Eaton et al. (2018, pg. 4) who write, "Psychology is also hampered by its

insistence on striving to explain human behavior, with humans positioned atop a pinnacle. Other species are of interest only as a source of comparison or a means to shed light on necessary criteria for uniquely human traits."

This version of the approach is routinely implemented because the study of animal minds is seen as being able to make substantial contributions to our understanding both of the evolution of the human mind and the extent to which animals have 'human-like' cognition, or cognition that is 'continuous' with that of humans (Shettleworth 2010). An example of this can be found in the animal communication research program, where there are accounts of intentional communication that are conceived of in a way that makes their realization less demanding and more intuitively amenable to evolutionary emergence (Moore 2018). Such an approach is often motivated by a commitment to a strong version of evolutionary continuity (Greenwood 2016).

While these two versions of the anthropocentric top-down approach (i.e. starting point and goal) can feed into one another they do not always come as a package. To be clear, the first is about *how* cognition in animals should be studied, while the second is about *why* a certain line of research is adopted. On the first (starting point), humans are used to obtain knowledge about animal minds, while on the second (goal), knowledge about animal minds is obtained because it tells us something about human minds. Both center the human, albeit in distinct ways.

On the goal version of the zoocentric bottom-up approach, cognition in animals should be studied primarily for the purposes of obtaining knowledge of the species or individual under investigation. This approach is highlighted by Eaton et al. (2018, pg. 4) who write, "Behavioral ecologists, ethologists, and biologists […] study species for their own sake, and situate each species in a larger ecological context, often examining the interactions between species." Eaton et al. (2018, pg. 11) continue and argue that, "[…] comparative psychology can extend its reach by focusing not only on how animals can be used as models to understand human cognition, but on how humans can be used to understand animal cognition. Utilizing a bottom-up, rather than a top-down approach, will facilitate the shift to understanding animals for their own sake rather than using them as tools to investigate human centered issues or traits." This is also eluded to by de Waal and Ferrari (2010, pg. 205) who write, "Rather than focus on the pinnacles of cognition,

the field of comparative cognition seems to be moving towards a bottom-up perspective focused on the nuts and bolts of cognition, including underlying neural mechanisms."[144] On de Waal and Ferrari's account this includes focusing on versions of cognitive capacities that are likely to be found in non-human animals rather than focusing soley on the versions that are found in humans.

This version of the bottom-up approach is routinely adopted due to the difficulties that the top-down approach has faced in focusing on establishing precise claims with regard to evolutionary continuity. Commitments to continuity on this approach, however, go awry as the available evidence is overstepped and overly simplified versions of evolutionary continuity are invoked (Brown 2019).[145] By removing the goal of demonstrating either continuity or discontinuity, the species or individual in question is thought to be able to be more productively studied. This is often meant to counterbalance the way in which the field, "[…] still suffers from an overemphasis on top-down approaches that begin and end with a focus on humans (Eaton et al. 2018; pg. 1)."

Again, while the claims of the zoocentric bottom-up approach plausibly feed into one another, they do not always come as a package. They, moreover, probe distinct questions about *how* and *why* a certain approach is adopted. On the first, animal minds are studied on their own terms, while on the second, knowledge about animal minds is pursued as an end in itself. Both center the animal under question, albeit in importantly distinct ways.

3.1.3 Complex v. Simple Capacities

The third sense of the top-down and bottom-up approaches casts the division in terms of complexity and simplicity. The top-down approach in this sense is based on the claim that inquiry in comparative cognitive science should be oriented around 'complex cognition'; typically complex cognitive capacities. This is held in part because a variety of non-human species are thought to be able to serve as informative starting points for the study of cognition. Rather than orienting inquiry around humans or the target species alone, research on any species that the

[144] It might sound like de Waal and Ferrari are merely advocating a mechanistic bottom-up approach here, however, this overlooks the details of their proposal.
[145] This has resulted in the broad formation of camps; one committed to demonstrating continuity and the other to downplaying continuity and purportedly prioritizing 'accuracy.'

cognitive capacity in question can plausibly be attributed to should be used to inform inquiry. For example, this is seen in the mirror self-recognition research program, where experimental protocols that chimpanzees were able to complete were successfully implemented in experiments involving other great apes and monkeys (Gallup and Anderson 2018), and then extended to a variety of other species such as whales, dolphins, insects, and snakes, among many others.

By contrast, the simple bottom-up approach is based on the claim that more 'simple' cognitive processes in the target species should first be understood before the cognitive capacity is able to be fruitfully studied. For example, De Waal & Ferrari (2010, pg. 201) write, "But what if we were to replace our obsession with complex cognition with an exploration of basic processes?"[146] In many ways, this is a reaction to the way in which the complex top-down approach has not been able to deliver clear evidence of the capacities they have tested for. By breaking the processes that constitute complex cognition into a set of 'simple' processes, reliance on intuition and the ambiguities that accompany it are thought to be able to be bypassed along with definitional disputes that are often verbal in nature.

The complexity/simplicity dichotomy typically involves there being more of less of something. While in an ideal scenario this could involve counting something specific, such counting typically remains intuitive as it does in other invocations of simplicity. Complex cognition in some cases might involves more cognitive processes than a simple cognitive process. This is not to claim that processes are the only feature of cognition that can be legitimately counted, but rather that for the complexity/simplicity dichotomy to be intelligible something must be counted either explicitly or implicitly.[147] In almost all cases, however, doing so will be highly unconstrained.

3.1.4 Special v. Widespread

The fourth sense cuts the distinction between top-down and bottom-up in terms of evolutionary distribution. Here, the top-down approach is oriented around evolutionary unique or 'special' cognition. These are often capacities that are thought to present unique puzzles for their

[146] De Waal and Ferrari here also refer to our "obsession" with what they call the pinnacles of cognition.
[147] For more on the variety of simplicity/parsimony claims in psychology, see Dacey (2016).

evolutionary emergence. One reason this approach is adopted is because these cognitive capacities are typically thought to be uniquely human. If they can be shown to be present in other species, claims to human uniqueness can thereby be refuted.

By contrast, the bottom-up approach that is focused on widespread cognition is based on the claim that cognitive capacities that are distributed across taxa often have an evidential basis in a variety of species that allows them to be more easily studied. As de Waal and Ferrari (2010, pg. 204) argue, "Every species, including our own, comes with an enormous set of evolutionarily ancient components of cognition that we need to better understand before we can reasonably focus on what makes the cognition of each species special." They also aruge that (de Waal and Ferrari (2010, pg. 201), "This approach […] will move the field of comparative cognition towards an understanding of capacities in terms of underlying mechanisms and the degree to which these mechanisms are either widespread or special adaptations." Not only is it thought that this will be able to yield a better understanding of the evolution of the capacity, but it is thought that by first identifying the common denominators of the capacity in a variety of species, will allow it to be more clearly characterized.

3.1.5 Whole- v. Subsystem

The final sense of top-down and bottom-up is cast in terms of whole- v. subsystem. The whole system top-down approach is based on the claim that because the cognitive capacities in question are targeted with personal level whole-system explanations, and that this is the level on which they will be most productively studied (see Figure 5 below).

An example of a phenomenon that is targeted with a personal level whole-system explanation can be found in the non-human animal mindreading debate discussed above which is focused on the beliefs, or on activities of whole agents. These are typically phenomena that are studied exclusively with behavior studies. On the whole-system top-down approach, neither a nuts and bolts understanding of the capacity that is being probed is required, nor a fully worked out account of the relationship between the task and the capacity in the experiment is required. This allows comparative cognitive scientists who are primarily concerned with behavioral studies to

remain neutral regarding the mechanisms underlying the capacities they are investigating.[148] By starting merely with intuitive or folk psychological ascriptions and repeatedly iterating them as behavioral experiments yield evidence, the expectation is that progress will be made and knowledge of the status of cognitive capacities in non-human animals will be converged on (Andrews 2020).

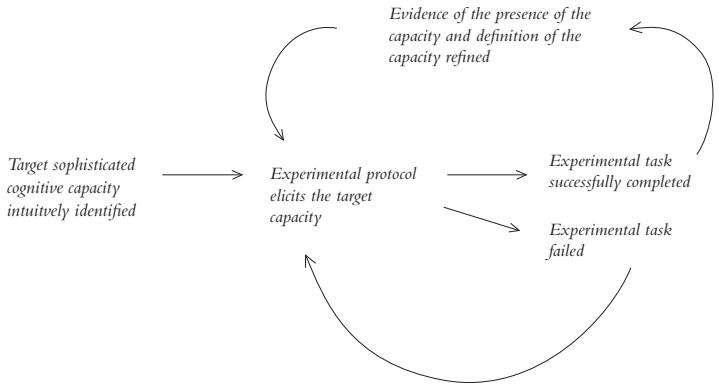


Figure 5. The whole-system top-down approach

By contrast, the subsystem bottom-up approach is based on the claim that the history of comparative cognitive science has shown that the whole-system top-down approach alone has repeatedly struggled to generate probative evidence of the cognitive capacities that are the target of inquiry. As Chitke et al. (2012, pg. 2677) write, "Comparative behavioural research needs to be complemented with a bottom-up approach in which neurobiological and molecular-genetic analyses allow pinpointing of underlying neural and genetic bases that constrain cognitive variation." According to Dennett, sub-personal level explanations targeted things like brains and events in the nervous system.[149] An example of a phenomenon that is targeted with a sub-personal level sub-system explanation would be the neural mechanism of social learning (Olsson et al. 2020). Rather than starting from zero, and relying solely on iterative behavioral studies, the bottom-up approach makes use of existing knowledge to identity the mechanisms that constitute the capacities in question. More generally, the subsystem bottom-up approach is supported by the

[148] This aligns most closely with Eaton et al.'s (2018) discussion of categorization projects.
[149] One noted source of controversy in this distinction was that Dennett considered both both personal and sub-personal explalnations to be properly 'psychological' explanations.

history of the study of comparative psychology, which has repeatedly struggled to obtain clear evidence of cognitive mechanisms from behavioral evidence alone.

3.1.6 Summary of the Disambiguation

Version	***Top-down orientation***	***Bottom-up orientation***
(I) Anthro- v. Zoocentric (starting point)	Human cognition (open)	Animal cognition (open)
(II) Anthro- v. Zoocentric (goal)	The enhancement of the understanding of human cognition	The enhancement of the understanding of animal cognition or fundamental cognitive/biological processes
(III) Complexity v. Simplicity	Complex cognition	Simple cognition
(IV) Special v. Widespread	Evolutionary unique	Evolutionary widespread
(V) Whole- v. Subsystem	Whole-system	Sub-system mechanisms

Table 20. Five senses of top-down versus bottom-up

Given that so-called 'bottom-up' and 'top-down' proposals that have been put forward are far from clear or univocal (see Table 20 above for a summary of the disambiguation), two seemingly successful research programs will be introduced that seem to invoke aspects of both top-down and bottom-up approaches in their various guises. This will be done with an eye to accounting for the variation in the implementation of these approaches. It will also be argued that various difficulties and unclarities arise as a result of specific epistemic limitations that are encountered in particular research programs. This means that a blanket strategy cannot be plausibly prescribed or adopted for the science of comparative cognition. Both the disambiguation of 'bottom-up' and 'top-down' proposals and its use for the analysis of existing proposals is meant to bring clarity to the research more broadly. This likewise applies to proposals such as those given by Logen et al. (2018), who have recently advocated the adoption of a "synergistic" top-down and bottom-up approach. Absent an analysis of the various meanings of 'bottom-up' and 'top-down' and why particular interpretations of these should be adopted, these proposals will remain ambiguous and limited in their potential to shape future research.

This brings out another important point. Namely, that all approaches on the analysis that will be offered are in an important sense 'complementary', albeit in a variety of ways. In some cases, that will mean that multiple 'bottom-up' approaches will be explicitly adopted simultaneously. In other cases, a mixture of 'bottom-up' and 'top-down' approaches will be adopted simultaneously. However, in no case will a singular approach be able to be plausibly adopted.

Under optimal circumstances, these complementary approaches combine the best versions of the top-down and bottom-up approaches that are available within the particular research contexts, while simultaneously providing the methodological resources to overcome the difficulties that were encountered by each individual approach. The starting point of these complementary approaches is that while top-down approaches can completely depart from the subsystem mechanisms that constitute a given cognitive capacity, or simply ignore them, there is no such thing as a purely bottom-up approach to comparative cognition, regardless of the version, that makes no reference to whole system behavior or entirely departs from all top-down approaches, while simultaneously remaining within the realm of comparative cognitive science. While not always explicit, a necessary part of providing explanations of the subsystem mechanisms is referencing the whole system capacities that they constitute. This simultaneous invocation of the top-down approaches by the bottom-up approaches can be carried out in as many ways as there are top-down and/or bottom-up approaches within a given research context. In the most minimal sense, the whole system cognitive capacity can serve as a bare orientation point for directing inquiry into the underlying causal or mechanistic structure.[150] This is a prerequisite for all inquiry on bottom-up approaches.

3.1.7 Caenorhabditis elegans (C. elegans) and Sensorimotor Integration

A robust understanding of the neural basis of vertebrate behavior is beyond the reach of our current science. This is not the case, however, for creatures such as C. elegans, which have proven to be an ideal model for studying both fundamental neural mechanisms and whole animal behavior. C. elegans, also known as nematodes, are one millimeter long, have a fully mapped

[150] While not highlighted explicitly, other top-down approaches are able to serve the epistemic function of the whole-system top-down approach.

connectome, and a fully mapped and invariant cell lineage (Cohen and Denham 2019).[151] This has made it possible for scientists the ability to perform powerful, multilevel, whole-animal analyses on them.

Nematodes navigate extremely complex environments while foraging, escaping predators, and mating (Pirri et a. 2009, Kiontke and Fitch 2013). While they lack the neural functioning and behavioral complexity seen in some mammals, such as visual and auditory perception, among others, nematodes share basic behaviors and a basic neural substrate that is common to almost all animals. This makes them an ideal model organism. Because the biological structure of the nematode can be mapped to its behavior, partially as a result of its size and the state of the technology, work on nematodes can potentially tell us about basic cognitive mechanisms that are widely distributed.

An example of this can be found in the C. elegans sensorimotor research program. Most animals can simultaneously integrate perception and action with one another (Brooks and Cullen 2019). This is because action determines perception and perception is also able to determine action. If animals were not able to distinguish one from the other, or if they were to confuse inputs from one with inputs from another, any movement or perception would turn into a potential blunder, thereby posing substantial challenges for survival. Simultaneously representing the environment (the sensory system) while also acting seamlessly in response to it (the motor system), is captured by a capacity called sensory-motor integration. However, how and at what point information from these two systems is integrated, have both remained live and challenging questions. A detailed account has been difficult to come by, at least in part due to the extreme complexity of most neural systems.

Understanding the mechanisms that constitute sensory-motor integration requires showing how the nervous system separates reafferent information, or information that is self-generated or

[151] At this point one might object that the goal posts are being moved and that in some relevant sense cognition proper is no longer being studied. This is a difficult objection in so far as there is no universally recognized definition of cognition. The classical definition has required computation over representations, but this is highly disputed. If one is sympathetic to that definition there is reason to think that the examples being discussed are relevant, however, a particular definition of cognition will not be defended.

originating in the body, out from perceptual information, or information that originates in the world (Roy and Cullen 2003). Doing this requires being able to form expectations regarding an action before it is taken (Brooks et al. 2015). It is sometimes theorized that inputs from the motor system, are responsible for generating these expectations (Crapse and Sommer 2008). While the broad neural regions that are responsible for this are thought to have been identified in humans, our understanding of the capacity at any sort of detailed level is extremely limited (Tosoni et. al 2008).

However, because nematodes have a completely mapped nervous system, this presents an opportunity for sensory-motor integration to be analyzed down to the cellular and even the subcellular level (Ouellette et al. 2018). This means that the physical and neurological substrate that composes these capacities can be identified with a high degree of precision. In particular, by making use of a variety of methods such as whole-brain Ca^{2+} imaging, which is a microscopy technique that measures intracellular calcium levels in individual neurons throughout the nematode brain, the Ca^{2+} imaging of neuronal activity has allowed for a research program to develop that possesses an extremely high degree of resolution; one that is incomparable to something that would be had in comparative cognition that is performed on vertebrates (Nguyen et al 2016, Ogawa and Miller 2013).

For example, Hendricks et al. (2012) showed that the RIA, which is made up of a pair of unipolar interneurons, are responsible for encoding head orientations as well as sensory input, which they are theorized to do via two types of calcium signal. This makes them operationalizable with the methods discussed above. Building upon this work, Hendricks and Zhang (2013) put forward a model of steering, or destination directed movement, that was based on output from the RIA to motor neurons in the head. This mapped the mechanisms that were responsible blocking the inappropriate activation of the RIA when the head is not bent, thereby allowing for appropriate sensory motor integration. They called this the gate-and-switch model.

Novel developments in the cognitive sciences are making new approaches to the mind possible, and the study of nematodes are playing a central role in these advancements. In addition to the latest technologies that have allowed for the analysis and measurement of brain activities in ways

that were previously inconceivable, several labs such as *Openworm* and *Si elegans* have been able to create computer simulations of the nematode nervous system at the cellular level. The variety of sources of data, and the ability to cross reference the resulting data against each other, as well as the resulting simulations, has proven to be clearly productive and generative.

Part of what underlies the success of the C. elegans sensorimotor research program are its complementary approaches. First, the research program is both 'bottom-up' and 'top-down' in a variety of senses. It uses animal cognition as a starting point in so far as it acknowledges the specific demands that C. elegans face with sensory motor integration, which makes it 'bottom-up' (zoo-centric 'starting point') in orientation. But it also makes use of research on sensory motor integration in humans to constrain the target of entry, which makes it 'top-down' (anthropocentric 'starting point') in orientation. The research program is markedly 'bottom-up' in the sense that it focuses on simple cognition that exists in an organism that has been extensively mapped in a number of ways and can be relatively easily controlled and manipulated.[152] Moreover, it is 'bottom-up' in the sense that it focuses on a cognitive capacity that can clearly be characterized as being evolutionarily wide-spread; there is good reason to think that sensory motor integration is common to most animals. And finally, the research program is 'bottom-up' in the sense that it targets subsystem mechanisms, but it is also 'top-down' in so far it requires an understanding of whole-organism behavior in order have any sort of orientation regarding what the sub-system is subordinate to.

The precise nature of the complementary approach adopted here is constrained by the epistemic status of the research program more broadly. For example, having access to a fully mapped connectome, and a fully mapped and invariant cell lineage opens particular approaches that are obviously not available to every research in comparative cognition. This means that the variety of methods and approaches that can be drawn on in this context are indeed specific to that context. The utility of the various senses of 'bottom-up' and 'top-down' that I have highlighted lies in the possibility of making these potential sources of evidence explicit, and in doing so, a path forward can be forged.

[152] Simple here could easily just quantify over the number of neurons in the system. This is at least in part what makes calcium imaging and simulations of this system possible.

3.1.8 Path Integration and Bio-Robotics

Almost all animals are required to navigate space in order to either acquire resources, migrate, or reproduce. This ability falls under the domain of the spatial cognition research program, which is focused on determining the ways in which animals acquire and make use of spatial information (Shettleworth 2009). One of the most widely distributed types of spatial cognition is called path integration. This is also sometimes referred to as inertial navigation or dead reckoning. In this widely distributed process, animals use idiothetic, or self-generated cues to calculate their egocentric spatial location in relation to a point of origin (Wallace et al. 2002). In the most general sense, it is thought to recruit working memory to take direction and distance travelled as an input to output a vector referring to the point of origin (De Nigris et al 2013). Typically, this is considered as a part of directly navigating 'home' after locating a food source, or subsequently returning to this discovered food source after having returned 'home.' It is conceptually distinct from wayfinding, which is the process of navigating according to perceived objects that are placed within a larger internal representation of the environment, such as a cognitive map (Gallistel 1990). Path integration sets itself apart from these in so far as external cues are not thought to play a primary or dominant role in determining the animal's location within an environment or the path of navigation that it takes.[153]

Path integration has been hypothesized and studied in humans since Darwin (1873). While there have been studies of the capacity in mammals such as gerbils since the 1980's (Mittelstaedt and Mittelstaedt 1980), the vast majority of research has been performed on invertebrates; particularly on diurnal foraging insects such as bees and ants. Most studies start with behavioral observation of what appears to be path integration in the wild. For example, when ants and bees leave their homes for food, they often take circuitous paths as a part of their search that prevent them from having a direct line of sight of their point of origin. Despite this, upon finding a food source, they can directly return home on a unique path (see Figure 6 below).

[153] Particularly within the human research program, these three types of navigation are thought to be co-existing and mutually reinforcing. This makes it difficult to pry them apart.

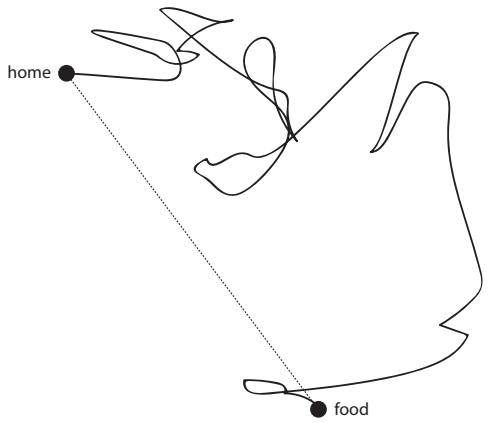


Figure 6. A model of path integration. The solid line represents the path taken while searching for food. Home is taken to be the point of departure and food is taken to be the destination that is not arrived upon ahead of time. The dotted line represents an idealized return journey.

With the phenomenon in seemingly plain view, the research program is then tasked with explaining the mechanism underlying this behavior that makes its realization possible. A series of experiments on desert ants (*Cataglyphis albicans*) adopted the strategy of displacing them after they reached their destination (Wehner and Srinivasan 1981). Having been moved to a new location, the ant then navigated to where its point of origin would have been, had it not been displaced (Wohlgemuth et al. 2001). This was meant to exclude the possibility of the animal making use of landmarks (distal cues) or beacons (proximal cues) (Grah et al. 2005).

While the behavior appears similar across a breadth of species, the underlying mechanisms are various. The desert ant (*Cataglyphis fortis*), for example, makes use of a sort of internal odometer that counts the number of steps it has taken (Shettleworth 2009). Because there are several strategies any given creature can adopt when confronted with a navigation task, one challenge for the research program has been to separate path integration from other methods such as routes, environmental geometry and cognitive maps, among others. For this reason, it is an appealing strategy to attempt to isolate the neural mechanisms underlying path integration in so far as there is a potential many-to-one mapping between behavior and cognitive mechanism.[154]

[154] Of course, this is not the only route available to comparative cognitive scientists. In theory a purely behavioral approach could do some of the work of empirically distinguishing these strategies from one another.

Distinguishing models of path integration systems from behavioral data alone has proven to be extremely difficult (Vickerstaff and Cheung 2009).

Faced with this, the field of biorobotics has adopted a novel methodological approach to this problem. A salient example of this can be seen in Stone et al. (2017), who have a fully anatomically constrained model of the neural mechanisms underlying every step of path integration in honeybees (*Apis mellifera*). Their model is centered on the CX (central complex), which takes sensory information as an input and gives motor information as an output. Unlike the ant mentioned above which counted steps in order to perform spatial navigation, the honeybee is thought to rely primarily on vision during flight-based navigation. In doing so, it makes use of both a (1) **visual odometer** as well as a (2) **visual compass**; both of which are allocentric or geocentric cues.

The (1) **visual odometer** relies on optic-flow.[155] In an experiment by Srinivasan et al. (1996), honeybees were trained to locate sugar water at a specific location in a tunnel painted with stripes that ran perpendicular to the direction of the tunnel. When these stripes were replaced with ones that ran parallel to the direction of the tunnel, the honeybees searched at random and were seemingly disoriented. Likewise, when the tunnel was more wide or narrow than the tunnel on which they were trained, they searched in a way that was predicted by the optic-flow hypothesis.[156] Since these early studies, it has been shown that landmarks are used to reset the odometer as a way to overcome errors, that obstacles are integrated and counted as part of the total outbound journey, and that the optic-flow based visual odometer is sensitive to three dimensional space (Srinivasan 2014). However, these behavioral studies alone have not been sufficient to determine how exactly the odometer is working, and a significant number of live models remain on the table.

Studies in other insects such as fruit flies and cockroaches, however, have identified specific neurons in the CX that are sensitive to large-field motion (Heinze and Homberg 2007, Seelig and

[155] This is distinct from earlier theories that posited that distance was measured in terms of energy expenditure, which was inferred on the basis on the waggle dance (von Frisch 1993).
[156] Recent studies have shown that bees have an odometer that functions in three dimensions. See Srinivasan (2014) for an overview.

Jayaraman 2015). While monitoring the CX neurons with electrophysiology methods, sweat bees were placed in a 360-degree LED arena that simulated large-field motion. Noduli tangential neurons or TN neurons, which are two types of CX neurons, were activated in these experiments. TN1 cells were activated linearly with the increased simulated speed of forward flight, while TN2 cells were activated linearly with the increased simulated speed of backward flight. Neither type of cell responded to rotational flow, thereby providing further support for the hypothesis that the respective cells have a preferred direction of motion. An identical morphology was found in the distant bumblebee (*Bombus terrestris*). This suggested that TN neurons are phylogenetically ancient and widely distributed in *Anthophila* (*Apoidea*).

The (2) **visual compass**, also sometimes called the sun compass, relies on celestial cues (Evangelista et al. 2014). Because the sun is in continual motion, using it as a compass requires an ephemeris function, which tracks the movement of the sun at a particular location and season, as well as an internal circadian clock (Shettleworth 2009). As sunlight passes through the atmosphere it is polarized to varying degrees at different points in the sky, contingent upon the angle of the sun. Many insects utilize this feature in navigation. This has been demonstrated in honeybees in clock-shift experiments that artificially change the direction of the polarized light in a controlled setting.[157] This allowed for the accurate prediction of the direction that the bee would travel in (Dyer and Dickinson 1996). While behavioral data seems to suggest that honeybees possess a visual compass that is sensitive to polarized light, this is not enough to determine exactly how this visual compass is working, nor has it allowed for the underlying neural mechanisms to be identified.

However, polarized-light-based compass neurons in the CX that encode head direction have been identified in the locust, butterfly, and fruit fly in several recent studies. Activity in the CX has been used to predict direction of movement in cockroaches, which has implicated it in the process of steering (Martin et al. 2015). Given this, the CX was hypothesized to form the foundation of the visual compass in bees as well (Stone et al 2017). In several experiments on

[157] As noted above, it is likely that many navigational systems are in play at once, even in honeybees. There has been recent work on cognitive maps in bees, and the ability of clock shift experiments to demonstrate the presence of this capacity, however, this work is still highly disputed. For more on this see Cheeseman et. al (2014).

sweat bees (*Megalopta genalis*), Stone et al. (2017) were able to identify the individual neurons that were activated when the creature's eyes were exposed to polarized light. These polarized-light-based compass neurons were shown to encode head direction that directly corresponded to the location of the sun. From this, it was inferred that the CX was the basis of the visual compass in sweat bees as well.

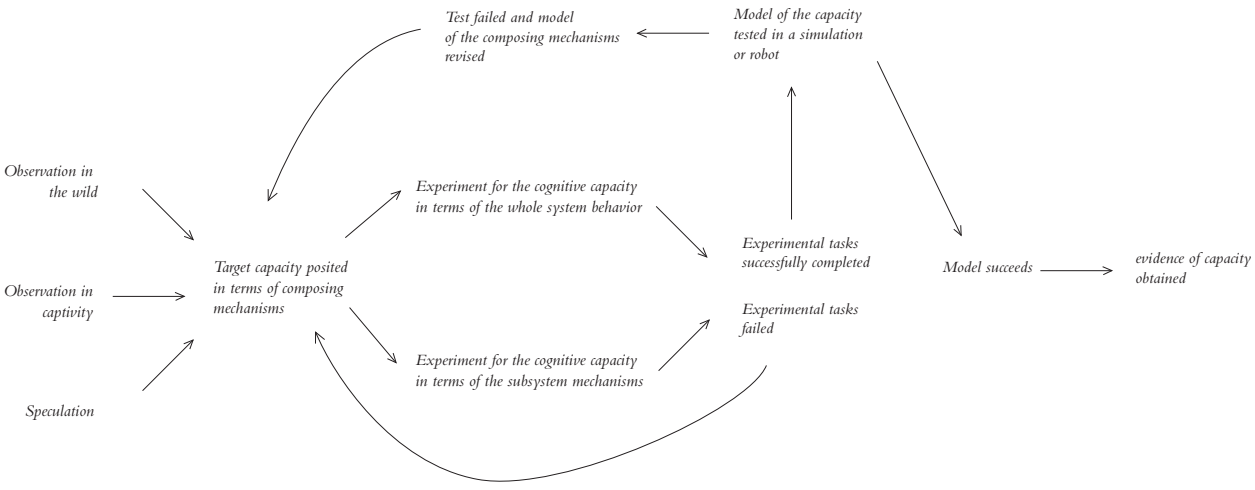
At this point, Stone et al. (2017) were able to map out the neural substrate that is responsible for integrating the visual compass and the odometer, using block-face electron microscopy. This included tracing both the inputs of the relevant neurons of the CX as well as the outputs that converge in CPU1 and descend to the thoracic motor centers that are involved in steering. The CPU4 takes an input from the noduli containing the TN neurons as well as the PB (proto-cerebral bridge), which is where the compass neurons are located in the CX. This has three arborization regions. Significantly, this means that the CX is thought to hold everything that would be required for a path integration neural circuit, which takes inputs from both the (1) visual odometer and (2) the visual compass and outputs information to the motor control system.

Given that the anatomical constraints are in place, Stone et al. (2017) were then able to give a detailed and constrained account of the way in which path integration is calculated in the brain of the bee. As in the nematode sensory integration case, the model for the underlying mechanism has been broadly analyzed down to a cellular level. This mechanism computes the difference between the desired and actual direction of travel, and outputs a steering direction that can close this gap. It is a plausible hypothesis that this mechanism, or some similar variation thereof, is present in all insects (Stone et al. 2017).

One of the most interesting aspects of this study is that the robustness of the model was tested by implementing it in a robot that is placed within the real world. When Stone et al. (2017) implemented their model in a robot, they found that it was able to successfully perform path integration, and it was even able to do so when noise was introduced into the system. This is a prototypical example of the general methodology of biorobotics as practiced by Webb and her colleagues, which generally attempts to test the understanding of the mechanism by situating it within a body that is embedded within an 'ecologically valid' environment (Webb 2001). The

environment in which real world animals are embedded is extremely complex. Given this, testing in this environment provides an advantage over simulation alone because it truly puts the model to the test. Real live environments throw unexpected factors at the robot that are hard to generate in simulations. In addition to this, and perhaps most importantly, this methodology is iterative in so far as it allows for modifications to be made based upon performance. For example, if there was some feature of the environment or the capacity that was not accounted for in the model that made it unable to perform path integration, this could be corrected and then be retested in a robot within the relevant environment. While this was not done in the case of Stone et al. (2017), it is an approach that has proven extremely useful in the discipline more broadly.[158]

Moreover, Webb has systematized this methodology (Webb 2001). The first step is to identify the whole system behavior. The second is to hypothesize the low-level neural mechanisms that are responsible for actualizing that behavior. Webb's analysis often takes place on the cellular level and is directed at identifying the neural mechanisms that constitute a given whole system capacity. Once these mechanisms are identified or hypothesized, the third step is taken, and the mechanism is implemented in robots and tested in an environment as a means of confirming it. Again, if the robots are not able to produce the behaviors that the insects are able to, the model is then revised and iterated until virtual behavioral identity is achieved, or it is concluded that a distinct cognitive mechanism is at play (see Figure 7 below).



[158] This means that the iterative process in this instance was not needed.

Figure 7. Model of Webb's approach in the context of the 'validity cycle'

3.1.9 Lessons in Complementarity

Webb's methodology obviously cannot currrently serve as an exact model for the study of every cognitive capacity in every species. The same technologies that are available for testing insect brains are not available for the study of vertebrate brains. The background information that is available to researchers will inevitably vary across taxa. Nevertheless, this method can function as a guide that highlights the advantages that come with adopting a variety of approaches simultaneously. That is, they can serve as a broader lesson in the advantages of complementarity.

The progress that Webb's lab has achieved is due at least in part to the fact that the approaches that have been adopted in the research program are seemingly 'top-down' and 'bottom-up' in a variety of ways; they exceed any one characterization. The research program is 'bottom-up' in the sense that it is orienting itself around a specific type of animal cognition (i.e., bottom-up zoo-centric; starting point), but it is also 'top-down' in so far as it draws on research on spatial navigation in more complicated animals (i.e., top-down complex). It is 'bottom-up' in the sense that the cognitive capacities that are being studied are comparatively simple and this process exists within a system that is also comparatively simple (i.e., bottom-up simple). It is also 'bottom-up' in the sense that the capacity is thought to be evolutionary widespread in so far as the pathfinding mechanism is plausibly instantiated in several insects. And finally, once again, the research program is 'bottom-up' in the sense that it targets subsystem mechanisms, but it is also 'top-down' in so far it requires an understanding of whole-organism behavior in order have any sort of orientation regarding what the sub-system is subordinate to.[159] In a somewhat removed way, it is also 'anthropocentric' (starting point) top-down in so far as the research is shaped indirectlyby the study of navigation in humans. In this way, a variety of top-down and bottom-up approaches are at play simultaneously and it is plausible that this is in part responsible for the success of the research program.

[159] This is congruent with an argument put forward by Krakauer et al. (2017), which emphasizes neuroscience's need for behavioral research. This is sometimes put under the header of 'behavior first' neuroscience.

At this point an objection might be made that it is in fact exclusively the various senses of bottom-up that are doing all the work here. If this were the case, then there would not be a significant sense in which the productivity of either research program could be attributed to their complementarity. However, such an objection misses the mark in a couple of ways.

First, isolated from any version of the top-down down approach, the various senses of bottom-up are themselves complementary in character. This remains the case even if the role of the various top-down approaches in the research is in some sense negligible. **Second**, to frame the role of top-down approaches in this research as negligible would be to significantly understate their role in the research and the iterative nature that they take on. For example, whole system top-down approaches do not merely provide a fixed orientation point around which mechanistic explanation orients itself. Rather, bottom-up approaches iteratively refine an understanding of the top-down whole-system behaviors by providing a firmer grasp on the bounds of the phenomenon. A similar point can be made about anthropocentric (starting point) top-down approaches. Intuitions about cognitive capacities in humans are capable of being iteratively refined as a part of complementary research.

To hold the bigger picture in perspective, above all, the primary advantage of adopting a complementary approach can be found in its potential ability to deal with some of the problems that were repeatedly introduced by the validity cycle. Notice the progress that has been made in these two case studies. There is no longer an extensive process of questioning the characterization of the capacity or the task-capacity relationship in a deep and significant way. And importantly, this is not because the fundamental explanatory target is merely behavioral while the cases that are responsible for giving rise to the validity cycle are in some way genuinely cognitive. In both cases, the mental is only accessible through the behavioral. There is no shortcut that could be taken. And in both cases the project is to study a genuine cognitive capacity.[160] There are indeed significant differences between the background knowledge that is available to the two research programs, but this is distinct from an argument that would claim that that the case studies discussed above somehow fall beyond the scope of comparative cognitive science. The task of

[160] Once again, there is a debate in the literature around what counts as cognitive with recent trends becoming ever more permissive (e.g., see Lyon et al. 2021). However, this debate is orthogonal to the issues at hand.

overcoming these differences in background knowledge should not be understated, and again, the proposal is not simply to tell the scientists working on primates to just do what the 'worm people' do. Obviously, such a proposal is untenable. Rather, the proposal is to look more closely at a high level of abstraction at what has made these cases genuinely progress generating and to use this as a guide to move significant portions of comparative cognitive science out of the 'validity cycle' that has ensnared it for so long.

In these cases, the mechanisms that are being tested for are richly embedded and able to be reliably tested for. The hypotheses are specified in a way that leaves no doubt regarding what is being investigated. As has been argued, the typical approach to comparative cognition attempts to study capacities that are characterized in folk psychological terms that are typically intuitively rooted (Penn 2011). Attempts to study these capacities leads to hypotheses that are underspecified, and debates about how to best think about or characterize the capacity at hand. This poses a problem because it is often not clear what exactly is being tested for in a given research paradigm, which again, has led to several clashing empirical targets. By contrast, complementary approaches provide a step towards leaving these problems behind. This is in part because the subsystem low-level mechanisms that constitute the cognitive capacity are posited as part of the hypotheses that are being tested. This gives the hypotheses that are being tested a degree of specification that is not typically had in the standard methods that are implemented in comparative cognitive science. This also leaves little, or at least less, room for equivocation regarding how to characterize a capacity, which means that objections and contestations regarding characterization will typically be targeted at a local rather than global level. Features of a model of the capacity that is being studied might be drawn into question rather than the entire understanding of the capacity.

Moreover, in positing hypotheses that make detailed claims about the relationship between structure and function, complementary approaches can place a variety of constraints on inquiry which are provided by evidence from across the cognitive and biological sciences. Stone et al. (2017) were able to model path integration at a very fine-grained level by making use of existing established evidence in neuroscience and biology. Rather than starting at an extremely high level of abstraction, they were able to exploit the rich existing body of knowledge from other disciplines to fruitfully guide investigations and place strong constraints on inquiry. It is this type

of approach that cuts across a variety of top-down and bottom-up approaches that holds the promise of driving comparative cognitive science out of the validity cycle and the problems that have given rise to it.

Given this, one promising way forward given the epistemic and material circumstances of the discipline would be to back off the study of sophisticated cognitive capacities in larger animals and instead focus, at least in the short term. on capacities that are more tractable in more 'simple' animals like bees and C. elegans. That is to adopt an 'invertebrates first' approach to the study of non-human minds. The human brain has ~86 billion neurons and the chimpanzee has ~28 billion. the honeybee has ~960 thousand (White et al. 1986) and nematodes have ~302 neurons. It is clearly the case that an extensive cellular level of analysis that is able to be performed on insects and C. elegans will not be able to be done on primates anytime soon, and even if this was able to be done, it is unclear what could be taken from it given the severe degree of complexity. Substantial progress can be made by shifting comparative inquiry towards invertebrates, which have been historically neglected. In many ways this is low hanging but valuable fruit, particularly given that many of the strategies that have been proposed to ameliorate the problems facing comparative cognition involves fine graining the problem.

3.2 Break the Target Capacity into Dimensions

Another ameliorative strategy that has been put forward for dealing with the 'validity cycle' involves breaking the cognitive capacity into independent dimensions that are able to come in degrees. There are a variety of reasons for adopting such an approach to scaling cognition. First, if the cognitive capacity in question is a disputed one, breaking it into several less controversial composing parts can be a way of initially gaining traction in the research, avoiding verbal disputes, and sharpening the actual site of contention. Another reason that one might adopt a dimensional approach to cognition is that there is reason to think aspects of the cognitive capacity in question are underappreciated or underinvestigated. In general, the approach is directed at breaking the capacity down in a way that intends to make it more tractable, and to place it within an evolutionary gradualist context.

In this section, two recent examples will be discussed in which the 'dimensional' strategy has been adopted: Starzak and Gray (2021) and Brown (2021). The problems that arise in each specific case will be identified, however, a number of these issues are shared between the cases. In doing so, the virtues and vices of adopting this strategy within that context will be highlighted with an eye to giving an overarching diagnosis of how coarse-grained target capacities can be broken down to make the relevant questions more tractable. In the more general discussion of the dimensional approach to cognitive capacities, the analysis will partially draw on dimensional analysis as it is used in other scientific contexts. This will set a number of preliminary constraints around how dimensional analyses should be performed and what types of they should possess in order to achieve the goals behind adopting this strategy.

To be clear from the outset, breaking down complex and seemingly intractable problems into more manageable parts can be a promising strategy in many research programs in comparative cognition; the intention here is in no way to refute this claim. However, there are innumerable ways in which this can be done and some of these are more promising and aligned with the stated goals than others.

3.2.1 Dimensions of Physical and Causal Cognition

Many animals fabricate and use tools to achieve their goals. For example, some chimpanzees have been observed using stones to crack nuts and sticks to fish termites out of holes (Goodall 1970; McGrew 1992; Nishida & Hiraiwa 1982). The extent to which animals understand the higher order physical principles that underly this type of action has occupied much of the debate around folk physics, which is also called naïve physics or physical cognition in the animal cognition literature.[161] It is clear that many animals understand that tools work. The question driving this research program, however, is whether they understand *why* they work (Povinelli 2000). The concept of folk physics was originally indexed to a common human intuitive understanding of the

[161] The folk physics research program has its roots in Jean Piaget's (1963) developmental work which sought to uncover how causal mechanisms are understood in development. While some causal realist such as Michotte (1963) held that causal mechanisms were indeed directly observable, Piaget, although also a causal realist, thought that causal mechanisms needed to be inferred from observable physical processes. In many ways, this mirrors the way in which the debate in animal cognition would come to focus on second-order relational reasoning (Penn and Povinelli 2007).

physical world. It is not to be confused with a scientific conception of physics, which most humans do not understand with any degree of precision. It is meant to cover concepts such as gravity, transfer of force, size–shape interactions, physical connection, causal interactions, and object transformation, among others.

Figure 8. A corvid participating in a version of the trap tube task.[162]

The trap tube test is one of the most well-known contemporary physical cognition tasks (see Figure 8 above for an image of a variation of this experiment). It was originally designed by Visalberghi and Limongelli (1994) and was inspired by chimpanzees that were observed 'fishing' termites from a mound with a blade of grass (Goodall 1968). In the standard version of this task, the subject uses a tool to put push a reward out of an end of a transparent tube. However, the task is complicated by the fact that there is a hole in the center of the tube. If the subject does not use the tool to push the reward it in the right direction, it will be trapped by the hole, and therefore not obtained. Visalberghi and Limongelli's (1994) original tests on capuchin monkey were unsuccessful, meaning the monkeys failed the task, which cast doubt on their ability to understand this aspect of folk physics. However, in Limongelli et al.'s (1995) subsequent tests, two of their five chimpanzees were successful. After performing several variations on this task, Limongelli et al. argued that these chimpanzees were able to engage in causal reasoning and understood the higher order causal interactions between the tool, the reward, the tube and the trap.

[162] This photo was taken from Nathan Emery's website https://featheredape.com/

Among others, Reaux and Povinelli (2000) criticized these claims and argued that the chimpanzee's successful behavior was also consistent with an explanation that appealed to associative learning. On this alternative explanation, the chimpanzees learned a behavioral rule such as 'use the tool to push the reward away from the trap', without having to engaging in any sort of genuine higher-order causal reasoning at all. Additionally, when Reaux and Povinelli (2000) attempted to replicate this experiment, only one of their four chimpanzees performed above chance, and when the hole was rotated to the top of the tube, meaning that it was no longer able to trap the reward, the chimpanzee continued pushing the reward away from the hole.[163] From this, and a number of other conditions and control conditions that Reaux and Povinelli tested, they claimed that a procedural rule was being used that did not involve the invocation of higher-order reasoning. In addition to this, the various trap tube tasks were thought to contain several potentially confounding elements that made it difficult to isolate higher-order causal understanding as a variable in the test subjects. Higher-order causal understanding here is meant to imply the presence of a 'genuine' or 'rich' concept of causation and not just a context bound association. Given this, Povinelli and Reaux (2003) introduced what was called the trap table task, in which the subject is confronted with a choice of pulling one of two rakes. The first brings the reward to the subject, while second pulls it to fall into the trap. Only one of the seven chimpanzees in this trial were able to successfully perform the task at above chance levels, and it was the same one that performed above chance in their replication of the trap tube experiment (Povinelli and Reaux 2003). This result in turn lead to several variations on the task that were intended to separate out higher-order causal reasoning from an explanation that was rooted in associative learning.[164] In these variations, the chimpanzees failed to generalize their understanding to novel situations. From these results of the tasks, Povinelli argued that although chimpanzees and orangutans seem to make use of tools more than other non-human primates, this has nothing to do with their understanding of unobservable causal principles, or their possession of a concept of folk physics. That is, their concept formation is limited to a perceptual basis, which allows for the formation of a *rule-based strategy*.

[163] Despite Reaux and Povinelli's claims, this replication is not particularly diagnostic. See the previous chapter for details on why this is the case.

[164] This includes the inverted and broken rake problems, the flimsy tool problem, the tool insertion problem, and the rope hook and touching stick problems, the support problem, and the bendable tool and tool construction problems. For more details of these see Povinelli (2003).

This conclusion, however, was not met with univocal agreement, and it was challenged, for example, by the Aesop's fable experimental paradigm (Bird and Emery 2009; Cheke et al. 2011; Jelbert et al. 2014; Logan et al. 2014; Taylor et al. 2011), which was inspired by Aesop's classic crow and pitcher story. In these experiments, mostly involving corvids, the subject is intended to access a reward at the bottom of a tube that is partially filled with water. By dropping objects into the tube that will displace the water and cause the floating reward to rise, the subject would be able to access it. The premise behind this paradigm is that the subject will only be able to do this if they understand the higher-order principle that underlies water displacement. In one experiment, Jelbert et al. (2014) found that New Caledonian crows (1) dropped stones into the tubes filled with water before they dropped them into the tubes with sand, (2) they chose to use buoyant objects before non-buoyant objects, and (3) they chose to drop objects into tubes with more rather than less water, all of which gave them easier access to the reward. From this, and similar studies, it was argued that corvids possess an understanding of higher-order principles underlying water displacement. This claim was surprising and novel because it suggested that New Caledonian crows have an understanding of causation and/or folk physics that would not be able to be achieved in human children until at least 5 years of age.

It should be clear that this research program constitutes a case of the 'validity cycle' in which the characterization of the capacity that is being tested for and the ability of the experiment to test for the capacity have been repeatedly drawn into question in a way that follows predictable paths. Moreover, the heyday of work on physical cognition in comparative psychology seems to be in the past, with very little empirical work being done on this question since 2014.[165] In this way, this fits the pattern presented in the 'validity cycle' above in so far as the research program can be characterized as having been largely abandoned. The reasons for this are surely multiple. There has been a general squeeze on funding in addition to the emergence of a trend towards phasing out research on great apes. In addition to this, there might be reason to think that the low

[165] This is not to say that a revival could not occur but given the general patterns that research of this type seem to follow, there is good reason to be skeptical that this will happen.

hanging fruit had been obtained and that further development would require a type of research that would far exceed the resources that are or will be available in the near future.[166]

Against this backdrop, Starzak and Grey (2021) propose to generate progress in the research program by breaking the target capacity down into dimensions. The idea of doing so they say is that, "rather than asking whether an animal fulfills the criteria for causal understanding, we should shift our attention to the question of *how to conceptualize causal understanding*" (Starzak and Grey 2021, pg. 9).[167] This means that we should shift our attention to how the cognitive capacity is being conceptualized, and in doing so, on this account, we need to break causal understanding into the following three dimensions; (1) sources, (2) integration, and (3) explicitness.[168] The idea of breaking the target down in this way is rooted in the goal of providing an evolutionarily plausible concept of causal cognition that can account for the space between representing higher order causal understanding, and explanations that have their roots in associative learning.

The first dimension covers the variety of (1) **sources** that an animal can draw on to make use of causal information. Drawing directly on Woodward's (2011) analysis of the different types of causal cognition, there are thought to be three sources of causal information: ego-centric, social, and observational causal learning. Ego-centric causal learning relies on a subject's own actions to extract causal information, social causal learning relies on observing the effects of others to extract causal information, and observational causal learning relies on variations that occurs in the natural world. It follows from this, that any given animal will be restricted in any one of the domains to a certain degree and that no animal will be able to exploit every source to the maximum degree possible. Moreover, it might turn out that some animals are only able to make use of one of the sources such as ego-centric causal information and in failing to acknowledge this possibility means that the research is not able to be sufficiently sensitive to the various ways in which causal

[166] This is not to say that work on this issue has ceased all together. For example, Hennefield et al. (2018) used meta-analytic techniques to account for success on the tasks by appealing to trial-and-error learning. In this way, even though empirical work has slowed to a virtual halt, a core feature that drives the validity cycle manages to persist, however, without new empirical work to drive it forward, the characterization given above is seemingly an accurate one.
[167] The italics in this quote are mine.
[168] This contrasts with accounts such as Gärdenfors and Lombard's (2018), which offers seven grades of causal cognition.

understanding can manifest. This is consequential in so far as it will be easy to bypass or overlook instantiations of causal understanding that are not extremely sophisticated.

The second dimension covers the degree to which an animal can (2) **integrate** causal information that is derived from different sources. For example, in some contexts, egocentric causal information may be able to be clearly integrated with social causal information, which would presumably result in a more robust understanding on a given causal mechanism. The restrictions or biases that might shape how an animal is able to integrate information across sources will in turn shape the types of causal understanding they are able to acquire.

The third dimension covers the degree to which an organism's causal representations are thought to be (3) **explicit**. This is thought to determine what an animal is able to do with the causal information in various contexts for various ends. They see this as determining things like how flexible behavior is. This means that if a causal relation is more explicitly represented, then more can be done with it, and it will in turn have less of an automatic character. The thought here seems to loosely track whether the causal representation is situated within in the system 1 versus system 2 distinction.[169]

The core of the proposal is then that by tracking causal understanding along these dimensions, a more fine-grained conceptualization can be obtained by allowing for studies to take on a degree of precision that they would not otherwise be able to. Moreover, as Starzak and Grey (2021, pg. 16) highlight, the approach is meant to highlight the way in which these three dimensions can dissociate from one another to form a conceptual space of causal cognition. In doing so, the thought is that an evolutionary realistic model of causal cognition that comes in degrees can be had, which is a step towards ending what they term the 'animal cognition wars.' In short, they see this as a promising strategy for achieving progress in the research program.

Getting clear on the concept that is being tested, or the target phenomenon of interest, is ultimately essential for an advancing research program, and this is explicitly the intention that

[169] A worry that one might have here is that these dimensions are in no obvious way distinct from one another and that this might have significant consequences for evaluating causal cognition along these dimensions.

underlies Starzak and Grey's approach. Moreover, having an operationalization of the concept that is sensitive to its various manifestations of it is essential. However, there does not seem to be agreement about what is being tested for in a coarse-grained sense, and this makes the project of homing in on an account of the phenomenon even more difficult than it might otherwise be.[170]

The reader might have noticed this in the shifts back and forth between 'folk physics' and 'causal cognition' as the phenomenon of interest. While understanding 'folk physics' might include aspects of 'causal cognition', collapsing them into each other, or claiming that they are co-extensive is seemingly not a productive direction for the research program to go in. Moreover, 'causal cognition' is plausibly too wide in scope to be a concept that is going to be tractable in the science of comparative cognition, particularly given that it potentially covers an extremely wide swath of cognitive capacities. For example, capacities like mindreading and other aspects of social cognition can be plausibly construed in terms of causation, but it does not seem like this is the best way to go about framing them in so far as important aspects of the capacity are lost under this more general and abstract framing. The same goes for the study of folk physics. If the original goal of the research program was to understand the extent to which non-human animals have an understanding of concepts like gravity that are covered by folk physics, then it is seemingly a mistake to reframe these problems under a concept that has a higher degree of abstraction like causation, or at the very least it is a confounding contribution to the research.[171] Rather, more traction could plausibly be obtained by asking detailed questions about concepts like gravity, and assuming that a concept of gravity can be broken down into dimensions, assigning dimensions to it so as to adequately capture that concept relative to the relevant goals. Starzak and Grey (2021) should not solely saddle the blame for this oscillation between targets of inquiry, given that it has occurred in the empirical literature as well, but going forward the research program would be well served not to reproduce it and to be clear regarding what the phenomenon of interest is.[172]

However, even if it is assumed that causal cognition and not an understanding of folk physics is the relevant phenomenon of interest, and the above concerns are put aside, it is still not clear that

[170] To be clear, not in a fine-grained sense involvng the precise calibration of the concept.
[171] Povinelli's book after all was titled, "Folk physics for apes."
[172] Given the type of empirical work that has been performed, the phenomenon seems to be better captured by the concept of folk physics.

this version of a dimensional analysis put forward by Starzak and Grey will be able to adequately address the problems that the research program is facing. A more detailed analysis of the capacity is surely a step in the right direction. It is hard to imagine how it could not be. However, it is not clear that Starzak and Grey's dimensional approach to causal cognition is the same thing as giving a more detailed analysis of the target phenomenon. This is initially evidenced by the relationship between causal cognition and the proposed dimensions. The intended metaphysical relationship between a capacity and the dimensions that it has been broken into is unclear and left implicit by Starzak and Grey, although they do say that it is meant to be a conceptualization of the target concept. One starting point for determining this might be to look at dimensional analysis as it is used in several other theoretical and practical sciences. In doing so, this can be used as a preliminary guide for thinking through the use of dimensions in the sciences of the mind and identifying their virtues and vices.

For example, in the physical sciences the concept SPEED can be usefully broken into length travelled divided by time; two dimensions with a defined relationship between them. Regardless of where a given object is located within these dimensions, it can be said to have a speed. This remains the case even if all the dimensions are bottomed out or instantiated to the maximum degree. In this way, what it means to be traveling at a certain speed is captured conceptually by these two dimensions. This makes it a prototypical instance of dimensional analysis that is commonly implemented across the sciences (Gibbings 2011). These dimensional analyses are accepted to obey several constraints such as the principle of dimensional homogeneity (Lemons 2017), which determines the types of transformations that can be performed on the respective dimensions and their relations. Most importantly, however, in doing so, they can capture the higher-level target concept in a more precise way by subjecting it to detailed analysis, relative to the relevant goals.

Can the same be said about Starzak and Grey's (2021) dimensions? The to answer to this question is seemingly *no*. If the project of giving a dimensional analysis of causal cognition is aimed at carving out a middle ground between associative learning explanations and explanations that invoking higher-order representations of causal relationships by analyzing the concept in more detail as they state, then the analysis has explicitly failed. What the three dimensions (again:

source, integration, and explicitness) could potentially be used for is to give a more fine-grained analysis of what types of higher-order representations are being invoked in each context. For example, a higher-order causal representation could be constrained to a certain number of sources, or it could be integrated with other information in a limited way, or it could be invoked more or less explicitly. However, if this is the case, then one is already located within the space of higher-order representations.

In this sense, no middle path has been carved and it is not clear that the relevant concept has been captured at all. Instead, only a specific more detailed explanation of the higher-order representation has been obtained. Moreover, it is not entirely clear that these details are the relevant ones. If we instead adopt an associative learning explanation of the data, then these dimensions can potentially provide a more detailed explanation of how that has taken place, although it might not be obvious that these would be the right dimensions to do so. Again, and importantly, however, Starzak and Grey's dimensional analysis is unable carve a middle ground between these two types of explanations as they intend, nor does it provide the tools that would be needed to adjudicate between these two types of explanations that have given rise to what they call 'the animal cognition wars', which they claim to take a step towards ending.

It seems that a prerequisite for performing this type of dimensional analysis is getting clear on limit cases. If again the original idea is to capture the space between representing higher-order causal understanding, and associative learning within the dimensional analysis, at a minimum, the dimensions should represent these two poles that it purports to thread its analysis between. But these two poles are nowhere to be found in this analysis. In this way, it may seem that the call is to stop asking this original question that has vexed researchers for so long. But unfortunately, that question remains securely in place and as central as ever.

This is because Starzak and Grey's dimensions merely capture the some of the ways in which the cognitive capacity could be constrained. In this sense, the point is well taken. No cognitive capacity is fully promiscuous in scope and flexibility comes with context-bound limitations. Even if an animal can be thought to possess a certain cognitive capacity such as associative learning or theory of mind or causal understanding, this always comes with certain biases and limitations.

However, this point is largely accepted in the discipline and in the philosophy that targets it. Being able to pick up on the fine-grained ways in which capacities can be instantiated in the world is something that many, but of course not all, comparative psychologists are already sensitive to. For example, within the context of a debate on animal rationality and concept possession, Hurley (2003) argued that an overly intellectualized account of the mind has been routinely used in this debate and that conceptual abilities that are fully promiscuous and freely re-combinable are not needed for animals to have reasons for acting. Instead, they require only holism and normativity. In this way, animals can have 'islands of practical rationality', which is meant to be opposed to a fully continuous 'space of reasons' (Sellars 1956). Hurley's intervention highlights the way in which many in the philosophy of mind has repeatedly used an overly demanding conception of concept possession. For a period, this led to a highly anthropocentric picture in the philosophy that neglected the empirical literature on animal cognition and left non-human conceptual capacities under-theorized. However, its value as an analysis goes beyond this context. This is because it provides a way to give a fine-grained analysis of the way in which cognitive capacities such as concept possession can be highly context bound. On this approach, cognitive abilities have a structure that integrates biases and limitations that aspects like environmental constraints place on it. However, again, this a solution to a different problem than the one that the causal cognition research program is facing, which is dealing with how to determine whether higher order causal representations of any form can be attributed to non-human animals.

To summarize, Starzak and Grey's proposal is best interpreted as an indirect contribution to a project initiated by Hurley (2003) and others, and it should be seen as a welcome addition or extension. However, their proposal does little to solve the problem that they set out to tackle and it in no way plausibly capable of 'ending the animal cognition wars' in the way that they claim.

3.2.2 Dimensions of Behavioral Innovation

In the most general sense, behavioral innovation has been theorized as the capacity to produce novel or modified behaviors that were not previously found in the relevant population (Reader and Laland 2002). Some examples that have been candidates for behavioral innovation have

included potato washing by Japanese macaques (Kawai 1965), a behavior that was not observed in previous generations, and blue tit milk bottle opening (Hinde and Fisher 1951), in which British birds learned to open and drink from milk bottles with aluminum caps and were able to pass this down generations.

However, perhaps one of the most frequently discussed cases of what might be considered to be behavioral innovation can be found in a study done by Weir, Chappell, and Kacelnik (2002). This study was motivated by observations of New Caledonian Crows in the wild, in which they would manufacture and use two types of hook tools as a part of foraging; a hooked twig and a stepped-cut barbed pandanus leaf (Hunt 1996). In this study, the author's New Caledonian Crow named Betty was able to manufacture a hooked tool from a piece of malleable wire in order to lift a reward from a plastic tube. In addition to this, Betty was also able to make hooked tools to complete this task from a variety of other media such as aluminum in order to solve similar foraging tasks. These surprising results shifted debates around the distribution of behavioral innovation away from primates alone and spawned a watershed of new research.

However, the typical markers of the 'validity cycle' emerged in this context as the characterization of the target capacity and the claims that were based in data derived from the experiment were both challenged, albeit in a perhaps less heated way than has taken place in other research programs.[173] While the various definitions have a family resemblance with one another they also have important differences between them. Kummer and Goodall (1985), for example, emphasize the way in which behavioral innovation can involve old behaviors being mobilized to solve new problems or new behaviors being mobilized to solve old problems.[174] Reader & Laland (2002) emphasize that the fact that the behavior should not have previously existed in the relevant population, and Ramsey et al. (2007) focus on the fact that behavioral innovation cannot be the consequence of social learning or environmental induction. By contrast, Carr et al. (2016) argue that behavioral innovation can involve social learning, but that if it does, it needs to have an

[173] If research programs never fully establish themselves so that open debates cannot be had, in some instances they might move seemingly very quickly to retirement. Moreover, the behavioral innovation debates are complicated by the fact that they are so closely connected to several other research programs that are highly contested such as social normativity, cumulative culture, and traditions, among others.

[174] While not mentioned, this should also include the logical possibility of new behaviors being mobilized to solve new problems.

asocial element as well. Moreover, these definitions have invariably been the target of critique.[175] In keeping with the 'validity cycle', these definitions, however, have not been the sole site of dispute, and in many ways, they have made space for the data to be more explicitly contested. This can be seen in Logan et al. (2016) who offer a social learning interpretation of the Weir, Chappell, and Kacelnik (2002) study. In addition to this, Rutz et al. (2016) offered a more deflationary explanation of the same data, by showing that New Caladonian Crows engaged in bending and other similar behaviors in the wild when fashioning their hooked twig tools. It is deflationary in the sense that it takes a case that was held up by many as one that demonstrated remarkable behavioral innovation and shows that the data can also be explained in terms of already existing behavior. That is, nothing new necessarily needed to be appealed to in order to account for this behavior.

It is against this backdrop, and in response to Arbilly and Laland's (2017) account which orders innovations from low to high magnitude, that Brown (2021) introduces her dimensional account of behavioral innovation. The driving motivation of this approach is to move these debates beyond the binary in which a behavior is framed as being either innovative or not, and instead to give a more fined grained and tractable account of the capacity that is being studied. This lets us potentially avoid agreeing upon a specified definition. However, it does require that some constraints be placed on the characterization of the cognitive capacity in the form of the following dimensions: (i) experience with being manipulated, (ii) novelty of the motor action being performed, (iii) novelty of the problem being solved, (iv) spontaneity of the behavior, and (v) robustness of the innovation.[176]

Brown (2021) makes three claims regarding the virtues of her multidimensional approach to behavioral innovation. Each of these claims will be addressed and it will be argued that given the current state of the research program, these claims should not be expected to be made good on. **First**, she claims that it allows us to make more accurate and nuanced comparisons of cases of behavioral innovation. Assuming that this is the case, this would indeed be an excellent feature of

[175] For example, Rendell et al. (2007), claim that Ramsey et al.'s (2007) definition will include and exclude behaviors it should not. This type of objection is typical in the debates involving cultural innovation.
[176] There is clearly a lot to say about each of these, however, Brown does not further specify these dimensions.

such an approach. One of the paradigmatic struggles that arise in comparative cognitive science centers around the project of comparing cases of a single capacity, particularly across species. If a multi-dimensional approach to behavioral innovation would in fact fulfill this role, that would indeed represent progress. **Second**, she claims that a multi-dimensional approach allows for the comparison of cases of behavioral innovation within a species. **Third**, she claims that this will allow for comparisons of innovativeness across species and clades and that doing this will provide support for inferring the mechanism that underlies the cognitive capacity in question.

There is reason to be skeptical that this dimensional approach will be able to do the work that Brown intends. **First**, the comparisons of cases of behavioral innovation will not be so easy to come by under this proposal in so far as the accuracy of any comparison relies on the accuracy of the individual components. This is primarily because of the nature of the dimensions that have been proposed. Many of the dimensions are at least as mysterious as the higher-order phenomenon; that is, the thing that the dimensions are dimensions of. This is a problem. For example, novelty is a central feature of dimensions (ii) and (iii). However, determining exactly what counts as novel will require an expansive understanding of the capacities that are underlying the behavior under investigation. Until this is had, the dimensions will be highly subject to a largely intuitive framing effect. The novelty of any given problem is contingent upon how the animal approaches it with the capacities that it has, and absent this constraint, any motor action or problem can be plausibly framed as maximally or minimally novel. The same applies to dimensions (i), (iv), and (v), albeit in distinct way. This is again to say that terms like spontaneity and robustness and highly contested in the science, or even outright neglected, and there is no straightforward way to apply them to any given situation.

The situation becomes even more complicated as comparisons are attempted to be made across species. This is because it is unclear what interspecies comparisons of behavioral innovation are indexed against. For an individual organism, this is at least somewhat straightforward in so far as we might be able to standardize a dimension such as novelty of the motor action being performed (ii) to allow for meaningful comparisons. This is even conceivable on the species level. But beyond that the measures cease to be obviously meaningful absent a trans-taxa baseline to index these measurements against. It is unclear where such a standard would originate from. Given this,

it seems that attempts to use this dimensional approach to make interspecies comparisons faces some serious obstacles.

In addressing Brown's **second** point, interspecies comparisons likewise face seriously obstacles, albeit of a distinct type. For example, novelty, which again is central to dimensions (ii) and (iii), is something that should be indexed relative the capacities of the species under investigation as was alluded to above. However, this task might not be as straightforward as it sounds. Throughout ontogeny, individuals of a given species are presented with a variety of dynamic experiences and learning opportunities that shape their cognitive capacities, and depending on the species in question there will plausibly be a large amount of individual variation. This in turn determines what might count as novel and places a strong epistemic barrier to intra-species comparisons. This becomes even more complicated when lab animals are being tested. Not only does there need to be an established baseline of the cognitive capacities, but that baseline also needs to be able to be shifted and tracked across ontogeny.

Finally, Brown's **third** point primarily encounters difficulties considering the problems with previous two. If the capacity in question cannot be accurately assessed considering the severe uncertainty that characterizes the terms used to define the capacities, it is not clear that this generates a better situation when attempting to identify the mechanism underlying the capacity.

Given this, in its current form, it is not clear that this dimensional analysis will be productive. Moreover, while it might seem that such an approach avoids the problem of agreeing on a definition in advance, in many ways, Brown's (2021) conception is a prototypical instance of contesting the definition of a capacity that is being tested in so far as it lays claim to the relevant dimensions. In a certain way, the types of debates and objections that currently take place in the research program remain the same, but with a multiplicity of new disputes. Rather than contesting what might count as behavioral innovation in some general sense, instead the dimensions, their relations, and how they relate to the phenomenon of interest will be the site of contestation. In certain contexts, this might be a step in the right direction. Again, however, there is nothing automatic about dimensionsal analyses that guarantees that they will be progress generating, and in many cases, they can be progress inhibiting. Which of the two is achieved

depends on the dimensions and their relations, and relative to the relevant goals, how these capture the higher order concept.

In addition to this, it seems like some version of the binary evaluation might end up being unavoidable even in this type of dimensional framework in so far as entry into the dimensional space will require a sort of binary evaluation along the specific dimensions, even if they end up being graded. In so far as dimensions are chosen, and one is located within those dimensions that are meant to account for some cognitive capacity, one might think that several choices need to be made that have a binary-like characteristic in order to locate a given behavior within the dimensions. For example, should the problem/task that the animal is confronted with be framed as being novel full-stop, regardless of degree?[177] Once this binary question is settled, it can then be located within the space, and one might think that it is exactly this type of question that is going to be particularly challenging to answer.

To take another example, if force is defined as mass times length divided by time squared (i.e., three dimensions with a defined relationship), regardless of the values that these dimensions take, an object can be said to have a force. It is not degrees of force. It is force, full stop. Can the same be said of the dimensions of behavioral innovation? The answer is seemingly no. Innovation is not something that is intuitively an aspect of every behavior. The conceptualization is meant to (1) capture the extent to which a behavior is, for example, novel along dimensions (i)-(iii), which will enable the possibility of (2) holistically capturing varying magnitudes of innovation, globally evaluated. However, by breaking down the target in this way, the question of how the dimensions relate to one another to produce an overall magnitude is left unanswered, despite the importance that Brown places on being able to represent this (Brown 2021; pg. 7). Presumably, these dimensions interact in a certain way so as to produce something that can be quantified and in turn ordered when performing comparisons; this after all was this intention. But it is ultimately unclear how this can be achieved with the resources available and the current state of the science. Particularly given the flexibility of dart charts, which Brown uses to represent the dimensional mapping of behavioral innovation, and the lack of conditions for quantitively evaluating these

[177] One way out of this might be to argue that every behavior is novel to some degree. Establishing that degree, however, might be an insurmountable task.

dimensions and comparing them, there is a lot of ambiguity and a lot of space for misunderstanding. In this way, it is not clear that this instance represents a promising way forward.

3.2.3 Broader Lessons for Dimensional Approaches

Capacity	**Behavioral Innovation (Brown 2022)**	**Causal Cognition (Starzak and Gray 2021)**
Dimensions	Experience	Integration
	Novelty of problem	Explicitness
	Novelty of motor action	Sources
	Robustness	
	Spontaneity	

Table 21. Summary of the two recent dimensional approaches discussed above.

Dimensional approaches are generally implemented in order to navigate the variety of conceptual problems and disputes that emerge in the scientific study of animal minds (see Table 21. above for a summary of two of these which were addressed here). Moreover, they purportedly allow some of these problems that arise such as underdetermination and the dearth of constraints, to be bypassed by forcing specification and making potential sites of debate more explicit. In this way, dimensional approaches make space to account for the many ways that one could possess a given cognitive capacity while simultaneously doing this with as much precision as possible. This is particularly attractive in the many contexts in which there is no clearly agreed upon definition of the cognitive capacity that is being tested for. This has been proposed as one of the ways of breaking out of the many issues caused by what was characterized as the validity cycle above.

However, as demonstrated in the analysis of the two cases above, dimensional approaches must meet certain criteria if they are going to deal with the problems that they were intended to take on. That is, there is nothing about breaking a capacity into dimensions alone, that guarantees a productive outcome. This means that a set of criteria are needed that will direct dimensional approaches and make it clear what allows them to be successful and what causes them to be confounding. Below, several normative criteria that intended to guide dimensional analyses of target concepts are presented. These are novel and represent a significant contribution to the

literature in so far as dimensional analyses have been implemented in several contexts and endorsed by several leading thinkers in the field (e.g., see Halina forthcoming, Dacey forthcoming).

First, the goals of dimensional analyses should be made explicit and contestable. These goals should not be assumed to be obvious. In highlighting that dimensional analyses can be introduced for a variety of reasons and underwritten by a variety of values, they can be more productively analyzed and developed, and the problem of critiquing a moving target can be avoided. For example, is the goal to open the potential diversity that is contained within the capacity in an almost exploratory manner (e.g., as in Birch et al. 2020)? Or is it to make a stronger metaphysical claim about the constitution of the capacity by the various dimensions? For example, the dimensions of behavioral innovation might be thought to capture in some full sense just what that capacity amounts to. Or maybe the dimensions are introduced solely to resolve some sort of local dispute. Again, the goals and values of dimensional analyses as they have been implemented thus far in the scientific study of animal minds have been at best implicit in form. This needs to be changed.

Second, assuming the appropriate goals, the lower-level dimensions that are intended to account for the higher-level capacity should be at least less mysterious than the higher-level capacity. Otherwise, the controversy is simply being displaced down a level and potentially multiplied. If this is the case, it is not clear in what sense the introduction of dimensions is contributing to the research program. This is a serious problem for animal behavior research in so far as there does not seem to be an uncontroversial foundation that most capacities could possibly be reduced to. At least in part, this is what has made controversies so prominent and the capacities themselves so difficult to get a good hold of. And again, the claim is that reduction seems to be what is taking place and is being targeted in many of these scenarios. For example, what it means to have causal cognition means to be located somewhere along these three dimensions. What it means to perform a behavioral innovation is to behave in a way that is located along these five dimensions.

Third, assuming the appropriate goals, the capacity that is being broken down into dimensions should be retained and properly represented in the lower-level dimensions. If the capacity of

interest disappears, or there is no obvious or uncontroversial relation between the dimensions and the higher-level capacity, then this is plausibly a case in which the subject has been changed. This is important given the conceptual work that capacities are meant to be doing. Breaking capacities down into dimensions opens the possibility of their being analyzed in more detail. However, this can only be done by committing to the definitional project.

Fourth, assuming the appropriate goals, absent a defined relation between the dimensions, any possible combinations of values of the dimensions should be seen as capturing the higher-level capacity. Take an analogy with two-dimensional space. No matter which point is chosen along the X and the Y axes, it will be within 2D space. The same, however, does not seem to apply to the examples discussed above. For example, if one has 0 sources within the conceptual space of causal cognition, they no longer seem to be within that space at all. Likewise, if one 'bottoms out' along each of the dimensions except for spontaneity in the dimensional model of behavioral innovation, then it no longer seems like behavioral innovation is being addressed at all. One way out of this problem would be to start setting demarcations within the dimensional space. This would entail drawing lines within dimensional space on which one could say that the capacity is present to some degree. It would be consistent with this strategy to allow for fuzzy boundaries or vague cases, but the boundaries, would need to be introduced, nonetheless. One of the problems with this is that it returns in the problem to some of these debates were meant to avoid in the first place. By breaking the problem down into its constituent parts, it was thought that we would be able to obtain a more fine-grained method for investigating these cognitive capacities. That might turn out to be the case, but if we no longer know when the cognitive capacity is present or not that is being explicitly targeted, the plot seems to have been lost.[178]

Part of the appeal of the approach is that several coarse-grained capacities are tightly connected with one another. For example, Brown (2022, pg. 1) writes, "More proximately, comparative psychologists associate behavioural innovativeness with a range of cognitive capacities including insight, problem solving, causal cognition, and creativity" (Reader and Laland 2003). In this way,

[178] This is not to say that the project could not be outright abandoned and that these questions could not be replaced by more fine-grained profiling projects. However, relevant to the current goals and values and the projects that have been adopted something essential has seemingly been lost in these scenarios.

dimensions are potentially networked into tight dependencies with one another. With sufficient data the ways in which various dimensions hang together could be tracked and mapped. This is to say, when things go well, dimensional analyses can be incredibly fruitful. However, nothing guarantees this. As has been argued, in principle, dimensional analyses have the potential to yield detailed knowledge of our cognitive capacities and their variations. But the precondition for benefiting from these analyses is that the above conditions for dimensional analyses be met. Otherwise, the problems that the approach was meant to overcome will be preserved and reproduced.

3.3 Formalize the Theories

As argued previously in this part of the dissertation (see Part 2 Section 2.1), part of what gives rise to the 'validity cycle' is the multifaceted problem of 'weak' theory. Given this, a set of proposals have been made that are intended to amend this situation and produce 'strong' theory through the implementation of formal and computational methods that target these theories. Proponents of this approach generally see it as an essential part of the solution to the theory crisis that the sciences of the mind are increasingly confronting. Here, a specified approach to the proposal within the context of comparative cognitive science is introduced. The goal of this section is to specify when and why formal and computational methods might be implemented to produce stronger theory in comparative cognitive science, and to highlight to some of the risks of doing so.

Within the sciences of the mind, there are several authors who have argued that formalization has the potential to ameliorate the so-called theory crisis. For example, Smaldino (2019) argues that strong theory should (i) be able to be formalized mathematically and/or computationally to generate testable hypotheses, and (ii) that it needs to acknowledge and work through any contradictions contained in the theory, assuming there are any. Oberauer and Lewandowsky (2019) argue that the researcher's degrees of freedom need to be reduced on the theoretical level, which is done by either performing what they call discovery oriented research (research that searches for effects to form hypotheses about) or theory oriented research (research that explicitly tests hypotheses), with the latter being bolstered by the formal and computational modelling of

theories.[179] Likewise, the broad and urgent need for the sciences of the mind to adopt formal modelling is also shared by van Rooij (2019), who views the use of formal, computational, and mathematical modeling through the lenses of complex systems theory and complexity analysis, and sees these methods as essential elements in escaping the theory crisis.

While this position is the dominant one in the literature, the idea that the formalization of theories, and the use of computational methods, is key to addressing the theory crisis has also received push back. Maatman (2021), for example, makes the strong claims that (i) formal methods cannot solve the theory crisis, (ii) that it is neither a necessary nor sufficient measure, and (iii) that adopting formal methods could make the problem worse. In addition to this, within the context of sociology, Besbris and Kahn (2017) argue that what is needed is less theory and more description, and that the outsized focus on theory generation, including through the implementation of formal and mathematical methods, has led the discipline to further impoverishment. This position is consistent with Borsboom's (2014) diagnosis of the problem, who argues that because the psychological and cognitive sciences have a dearth of theory, they are required to lean heavily on empirical work, and that this is something that simply should be accepted and even embraced given the current state of the research.

In short, there is consensus among these authors that the sciences of the mind are faced with a serious problem that centers around theory. However, there is an absence of agreement regarding what exactly the problem is and what role formal and computational methods should have in ameliorating this situation. While there is strong pushback against the exclusive use of verbal theories among some of the stronger proponents of formalization, this type of proposal sometimes implies starting from zero and/or abandoning verbal theories all-together. This is blatantly a mistake. Afterall, it is not even clear what this would mean given that verbal theories are seemingly always the starting point from which formal theories are developed as a part of an iterative process. Moreover, even the most formal mathematical models require an interpretation.

[179] This is terminology that Oberauer and Lewandowsky lift from Simmons et al. (2011). This trades on an analogy with the 'degree of freedom' in statistics, however, it does not have that technical meaning. Rather, it simply seems to imply that constraints be placed on the researcher.

In this way, the verbal is an inevitable part of any part of formal inquiry.[180] The question then is a somewhat straightforward one: when should theories be formalized?

Many of the stronger proponents of formalization, such as van Rooij, Fried, and Smaldino have argued that the answer is essentially the following: *always*. They hold that the best way to achieve a good or 'strong' formal theory, is to initially form a 'bad' formal theory that can eventually be made 'strong' through a process of iteration. Within the broader discourse, the general case for formalization and its appeals can be seen perhaps most clearly in van Rooij and Blokpoel's (2020) attempt to provide a guide to how verbal theories can be converted into formal theories. In one example, they set up a problem involving a decision that needs to be made regarding who to invite a dinner party. If some of the friends do not get along, this creates a situation in which who likes whom can be formalized, which will allow for recommendations to be made regarding the combination of individuals that the host might want to invite to the party. As van Rooij and Blokpoel (2020, pg. 6) write, "Say a host knows six people they all like, i.e., P = L = {A,B,C,D,E,F}. We can depict their like and dislike relations in a figure. The color of the lines indicate for each pair of persons p_i, $p_i \in P$ the value of like (p_i, p_i)[…]" (see Figure 9. below).

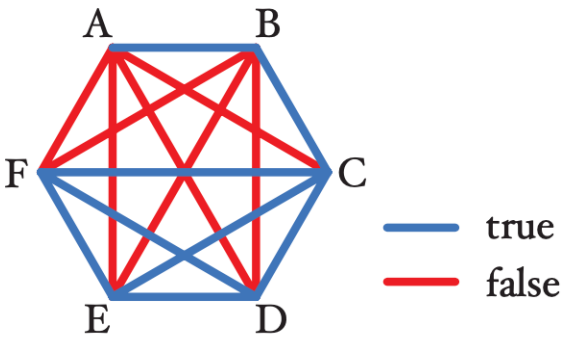


Figure 9. A model of the dinner party example employed by van Rooij and Blokpoel (2020, pg. 6)

[180] Strangely, this point has been missed by some in the literature (e.g., Smaldino 2019, Muthukrishna and Heinrich 2019), however, the potential of an iterative back and forth between the verbal and the formal in theory development has been repeatedly highlighted by van Rooij (2021), among others.

From this, a prediction about who the host would invite to the party can be made. Namely either C, D, E, and F or A and B. Under idealized assumptions, any other combination of guests can be excluded.

The appeal of this approach is that all of this can be done by making a series of simple formalizations and then performing basic logical operations on them. Moreover, formalization in this case clearly facilitates achieving the stated goal. Although some theories are weak in a way that will make it difficult to formalize them, formalization will always be a virtue of a theory, on van Rooij and Blokpoel's account.

One of the initial limitations of this proposal lies in the fact that it is fundamentally unclear how to implement formal and computational methods in many of the research programs in the psychological and cognitive sciences. This is because many theories in the sciences of the mind simply do not even remotely resemble the dinner party example. This challenge exists against a background in the sciences more broadly, where any physical situation can be formalized in a variety of ways (Frigg and Nguyen 2020). While Rooij and Blokpoel's examples are clear cut and serve as good case studies on formalization broadly construed, particularly for scientists in comparative cognitive science that are not typically accustomed to formal methods, they are not clearly or straightforwardly applicable to many cases in the psychological or cognitive sciences. The local question is then, what contribution can these proposals realistically make to comparative cognitive science, particularly when it comes to overcoming the challenges that the validity cycle introduces?

Comparative cognition has had its own, albeit very limited, version of this larger debate around the status of theory in the discipline and the role of formal and computational methods in addressing this problem. Allen (2014, pg. 76), for one, has argued that comparative cognition targets what he terms the "wrong kind of theory." Generally, the argument is in line with the points made within the human psychology literature– the theories tested in the discipline are too verbal, metaphorical, and vague to gain significant traction, and given their almost exclusive implementation in the discipline we should not be surprised to see extremely limited development in the near future. While verbal theories have managed to do some groundwork,

Allen's overarching argument is that they can take the work no further, which is evidenced by the numerous purportedly stalled research programs. The promise is that with the help of the tools of formal and computational modelling the discipline will be able to overcome its 'trophy-hunting' phase which will allow it to generate 'strong theory.'

Against this, Lurz (2014, pg. 99), argues against any sort of radical change, such as the one proposed by Allen, that might entail shifting the explanandum of comparative cognition to something other than the "reasons for which animals do what they do—folk psychological questions about animal behavior." In doing so, he appeals to the purported progress that the standard theories and approaches to comparative cognition have yielded over the past forty years by employing concepts like reason, knowledge, belief, and desire, and argues that abandoning these in a type of 'eliminativist' move will result in both impoverished explanations in particular research programs and a misguided discipline more broadly that is returned to the so-called 'darkest days of behaviorism.'

This debate is a potentially rich one. Yet over eight years have passed since the publication of these papers, and they have seen no further direct uptake either in in the philosophy of science or in the science of comparative cognition, even though the theory crisis literature has since exploded within the broader sciences of the mind. Nevertheless, the use of formal and computational methods in the discipline, has gained more traction, and has been used to generate theory. With the recent surge in the implementation of machine learning methods across the sciences, this trend seems only likely to increase. Below, two cases are discussed and the broader lesson regarding formal and computational methods for theory development in the discipline are extracted.

This first comes from van der Vaart et al. (2012), who introduce a formal model of scrub jay re-caching, which shows that mindreading is not required for the completion of two experimental tasks.[181] The appeal of the model can be found in the fact that it purportedly does not require an ad-hoc behavioral rule to be formed and instead requires only one behavioral rule: prefer to cache

[181] The reader will remember that this case was discussed in Part 1 Section 5.1 and in Part 2 Section 1.1.2 of this dissertation.

not close to conspecifics.[182] The model then has the following three parameters, "[…] *d*, governing how quickly memories decay, *n*, specifying the noise in their likelihood of recall, and *st*, determining the stress threshold at which recovering 'virtual birds' re-cache (Vaart et al. 2012; pg. 2)." Every time the virtual bird cached or recovered, this would be encoded in the memory system at a certain strength. More frequently or more recently caching or recovering at a site, will increase its strength. The stronger the memory of caching or recovering at a certain site, the less likely the virtual bird will be to cache or recover there again. Based on work that links stress and caching behavior (Pravosudov 2003), Vaart et al. assume that stress increases both caching and recaching, and that missing caches is a source of stress that the birds want to minimize. This implies that if there is a spectator, particularly if it is a dominant one, stress levels will increase, which will cause the virtual bird to recache when its stress levels surpass a certain threshold. Based on this model, Vaart et al. ran their simulation of two experiments: Emery and Clayton (2001) and Dally et al. (2004). They found that the virtual birds behaved 'similarly' to that of the scrub jays in these experiments, although mindreading was explicitly not a part of the model. Moreover, Vaart et al. describe a sort of feedback loop that appears to have occurred to produce this effect. The more stress that the virtual bird experienced, the more it re-cached, which later in the experiment caused its memories to degrade, which led to more recovery failures, and finally more stress (See Figure 10 below for an illustration of this).

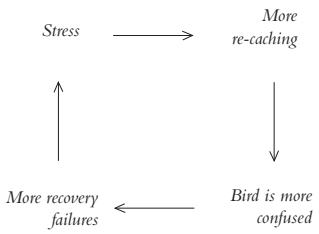


Figure 10. Depiction of the feedback loop in Vaart et al.'s computational model

Through computational methods, Vaart et al. (2012) produce a formal specified model that relies neither on the invocation of ad hoc behavioral rules nor mindreading to achieve its results in a simulated environment. Still, there might be reason to think something important is being lost in

[182] This is a common response to the alternative hypothesis explanation that has been given by Fletcher and Carruthers (2013), among many others.

this model. According to Lurz (2014), Vaart et al. produce exactly the type of cognitive modelling that promises to lead comparative cognitive science astray, in so far as it abandons folk psychological terminology (i.e., the reasons for which animals act).[183] What is interesting about this charge, however, is that Vaart et al.'s model is still at least implicitly shot through with plausibly folk psychological terminology. For example, stress can be measured physiologically through markers like cortisol levels, but it is also a folk psychological state. Moreover, the other parameters such as memories and their use seem to be able to be reframed in the more straight forward folk-psychological terms like beliefs and desires. The difficulty that this model faces then is not that 'the reasons that animals act' are disappearing from the study of minds, but that the relationship between the formal aspects of the model and the interpretation of the model is highly underdetermined. Even though Vaart et al. appeal repeatedly to the 'simplicity' of the model as one of its core virtues (i.e., only one behavioral rule is relied upon as opposed to many), one might worry that this comes with concerns about its external validity and the unconstrained nature of the interpretation of the model. While it is clearly possible to construct a theory that is mathematically precise and ripe for computational manipulation, this is no way entails that the theory is an accurate or even useful one that can bear on the target system, particularly when the verbal theory is highly disputed. In this way, sensitivity to both the precision and accuracy of the model must be held up, and the relevant degree of uncertainty must be centered. Given the case that has been made, mindreading and the other alternative hypotheses remain live hypotheses in these experiments.[184]

Given this, it remains unclear what contribution formal and computational methods can make to theory development in the non-human animal mindreading research program. Within the context of research on mindreading and autism in humans, Galitsky (2016) produced a formal model of theory of mind. Galitsky's (2016) model shines in its ability to take a large set of complicated linguistic information to produce a contribution to a consistent theory of mental state

[183] Contra Lurz, folk psychological explanations should be appealed to as long as they are useful, and holding onto them beyond this is simply to resort to dogma. There is no fundamental reason to think that folk psychological explanations are in any way essential to the study of animal minds, although overcoming them is not going to be as easy as just taking up formal modelling.

[184] It is interesting to contrast Vaart et al.'s exclusively behavioral computational model with Webb's mechanistic approach to computational and robotic modelling discussed earlier in this part of the dissertation. It seems plausible that what made Webb's approach more clearly successful is its explicitly complementary character.

reasoning. However, it is not obvious that this is the challenge facing the non-human animal mindreading research program at all, given that the model is solely based around natural language usage.[185] This places a block on its potential utility for addressing the types of presence v. absence questions that have occupied the non-human animal research program.

Given this, it is not entirely obvious that formalization is *always* the best or most obvious way to develop a theory or to advance a research program as Rooij and Blokpoel (2020) have argued. The question then remains: When can formal and computational methods assist in theory development and when might they be distracting or confounding?

The second case has a somewhat different character and comes from the intentional communication research program. Although work on great ape intentional communication aims to unlock the evolutionary origins of human communication (Fitch 2010, Corballis 2010), inquiry into this capacity is fractured with the vocal, gestural, and facial lines of research employing incongruous theoretical and methodological approaches. Recently, however, Bohn et al. (2021a), have generated a multimodal cognitive model of great ape intentional communication as a multifaceted social inference process, which cuts across the previously disparate lines of research. Taking inspiration from the RSA framework that has been used to analyze human infant language understanding (Bohn et al. 2021b) and cluster analyses on behavioral streams of socio-communicative behavior (Altman 1965), Bohn et al. developed a model that relies on three parameters; the information contained in the signal, the relationship between communicative partners, and the social context. This model is precise and makes strides towards unifying the research program under a richer more 'realistic' theory that has greater predictive capacity than its mono-modal (i.e., solely vocal, gestural, or facial) counterparts.

In constrast to Vaart et al. (2012), this is an example in comparative cognitive science in which formal and computational methods in the discipline are successfully implemented to generate a

[185] This case is meant to clarify the fact that although there has been a 'successful' formal model of ToM, the conditions that contributed to its success will not be available to the non-human animal mindreading research program. Making this clear serves to define the challenging task of implementing formal and computational models to study ToM in non-human animals and why work done in the human research program cannot be obviously drawn upon to do so.

genuinely novel and 'stronger' theory. Part of what explains its success is the nature of the data that is being dealt with. In contrast to the first case, Bohn et al. (2021a) do not simply take a weak verbal theory, formalize it, and run it through a simulation. Instead, they used formal and computational methods to synthesize a wide variety of data and to draw inferences that would have otherwise evaded researchers that were relying on observations of one modality alone. In this way, this provides model case for the implementation of this formal and computational methods in comparative cognitive science.

So where does this leave the science of comparative cognition in relation to formal and computational methods? There is undoubtedly a place for these methods in the future of comparative cognitive science and their implementation will likely increase exponentially in the coming years.[186] However, while these approaches can be useful in some contexts, formalization and computational methods are not a panacea for the cluster of problems that are unified under the header of the theory crisis.[187]

Formal and computational methods hold the following potential advantages, among others[188]:

1. They can identify non-obvious inconsistencies in theories.
2. They can resist the problem of theoretical and hypothesis under-specification.
3. They can produce non-obvious testable predictions.
4. They can analyze and synthesize large amounts of disparate information.

However, benefiting from these advantages requires the relevant research program to be in a state to do so, and not all research programs are in that state. For example, identifying inconsistencies can be useful, but in the early stages of theory development eliminating inconsistencies should perhaps not be a priority. Moreover, if theories are implemented in a non-iterative dialogical way with verbal theory, they also risk cementing inaccurate implicit assumptions about the target

[186] This is the case even though most comparative cognitive scientists typically lack training in formal and computational methods. There is good reason to think that training in the discipline will be transformed in coming years.
[187] This remains the case both within the science of comparative cognition and within the sciences of the mind more broadly construed.
[188] The first three are highlighted by Rooij and Blokpoel (2020).

system, such as a certain model of cognition or representation. In addition to this, in taking on an unwarranted degree of precision they risk distracting from the broader project of theory development that should be situated in a more exploratory context. The widespread invocations of the 'free energy principle' in the sciences of the mind might be seen as an instance of this (e.g., see Friston 2010).

In sum, as part of an iterative back and forth with verbal theorizing, in specific contexts, formal and computational methods clearly have the potential to contribute to theoretical development in comparative cognitive science. However, the progress that can be obtained from this strategy is in no way straightforward and requires substantial local analysis in order to determine the correct context of application. Doing so will allow for the potential benefits of computational and formal methods to be held in view and for this approach to be maximally progress generating.

3.4 Change the Testing

Another proposal for dealing with challenges posed by the validity cycle involves moving away from traditional testing methods. In what Taylor et al. (2022) characterize as 'success testing', an animal is evaluated with regards to its ability to complete a task or solve a problem in a predominantly binary sense. Whether or not the species has that capacity hangs predominately on a single test. It is this approach to testing that is responsible for the persistence of what they term the many to one mapping problem; many cognitive mechanisms can produce any single behavior. In contrast to this, they propose what they call 'signature testing', which is meant to identify relevant 'patterns' of cognition with the intention of constraining the space of hypotheses through multiple 'signatures.' They highlight identifying things like biases and information processing errors, but a signature is really just a, "[…] a pattern of behaviour (or other pattern of observable evidence) that constrains the hypothesis space for the cognition under investigation […]" (Taylor et al. 2022; pg. 4). By changing the type of testing that is done, and not overburdening individual tests, Taylor et al. argue that the inferential gap between behavior and cognitive mechanism can be closed, and the many to one mapping problem can be overcome.

The approach takes its inspiration from way in which unobservable phenomena have been studied in other sciences. Taylor et al.'s (2022) core example involves the identification of black holes in astrophysics which, like cognitive properties, are not directly observable. Imagine that a region of space is observed with an accretion disc that is emitting x-rays. These properties alone are not necessarily indicative of the presence of a black hole in so far as many objects in space are capable of producing them. However, if this observation is combined with another measurement of the object's mass, the hypothesis space becomes more constrained. This, in addition to other 'signatures' such as low luminosity, a lack of a boundary layer, and x-ray bursts with a particular profile, can all be combined to narrow the space of hypotheses in a way that will warrant a strong inference regarding the presence of a black hole.

This case is thought to provide a useful point of contrast with 'success testing' or pass/fail practices that Taylor et al. frame as being central to the science of comparative cognition as it is currently practiced. What is thought to be wrong with these approaches on this account is that they either make use of a single signature, or they do not even frame it as a signature, which amounts to overburdening a single source of evidence with the task of purportedly fully constraining the space of hypotheses.[189] However, focusing instead on the numerous information-processing errors, limits, and biases that are obtained through several different sources, are thought to be able to compose a cluster of signatures which would be able to constrain the space of hypotheses in the way that a single signature fundamentally would not be able to. To return to the black hole example, this would be akin to using any one test in isolation and expecting it to settle the debate. If low luminosity was the only source of evidence available, the hypothesis space would remain insufficiently constrained. Likewise, if a lack of a boundary layer was the only source of evidence available, the hypothesis space would remain insufficiently constrained. However, in combination, the search becomes at least more constrained. The set of possible hypotheses that can account for the evidence is reduced. Essentially, this amounts to making use of all the available evidence in order to draw the strongest possible inference.

[189] To be clear, there is surely a space between these two positions, although there is not a reason to think that it is currently being enacted in comparative cognitive science.

For Taylor et al. (2022), the main appeal of signature testing is its ability to accelerate progress in the discipline rather than merely generating debate. This is a familiar problem from the previous discussion of the validity cycle in this part of the dissertation (see Part 2 Section 1.1). Their core example of the implementation of signature testing involving animal cognition comes from a study done by Bastos and Taylor (2020) on kea parrots (N=6) in which they claim to have identified three signatures of domain general statistical inference as it is understood in humans: (1) the use of relative frequencies, and the integration of information from both (2) the physical as well as (3) the social domains. These were identified over three experiments, which comprised the study.

This first experiment involved the kea observing a trainer sampling two types of tokens (orange neutral rectangles and a black rewarding rectangles) from two jars. They found that kea's behavior in attempting to obtain the reward correlated with them tracking relative rather than absolute frequencies. From this Bastos and Taylor (2020; pg. 2) claimed that this, "provide[d] conclusive evidence that kea show true statistical inference using the relative frequency of item." Taylor et al. (2022; pg. 743) go on to interpret this as an information processing bias.[190] The second experiment was a modification of the first and was meant to test the keas' ability to integrate a physical constraint into their predictions. In this experiment, a barrier was placed halfway down each of the jars, which prevented the lower half from being sampled from. The relative frequency of the orange and black tokens varied below the barrier of each jar. After training, and under a variety of conditions, the kea, on Bastos and Taylor's account, were able to successfully integrate this physical information into their predictions when attempting to obtain the reward. The third experiment provided a further modification and was intended to test the kea's ability to integrate social information about the bias of the sampler into their predictions. Again, from their findings it was claimed that the kea were able to use this information about the bias of the sampler, and allow it to override information about relative frequency, in order to make successful predictions. Taken together, these three experiments, interpreted independently as providing evidence of three distinct signatures, were claimed by Bastos and Taylor to collectively provide evidence of

[190] Presumably these are distinct but not contradictory claims.

domain general statistical inference in keas.[191] Their performance on these tasks was also claimed to mirror that of infants (Wellman 2016) and chimpanzees (Eckert et al. 2018).

Taylor et al. (2022; pg. 746) highlight four current limitations to their approach: (i) generating signatures, (ii) nesting, (iii) weighting, and (iv) context and reliability. These are important challenges that the approach faces. In the following, however, several additional challenges facing the approach are introduced and analyzed. What initially stands out in this study is the certainty with which the signatures themselves are presented. Language like, "conclusive evidence" is difficult to justify or even interpret in this context. Apart from the many alternative hypotheses that were not considered a sample size of 6 alone seems to warrant at least chopping the word "conclusive" out of this claim. There is reason to think, however, that is not merely an inflated claim, particularly when placed within the context of introducing signature testing. This is because it reveals a certain sense in which the description of the signature testing approach can be somewhat misleading. One question that might be asked is whether signature testing reproduces many of the problems of success testing, albeit under a different guise. As Taylor et al. (2022; pg. 742) write, "signature testing subsumes success testing." But if this is the case, one might think that the foundation of the approach is already on shaky grounds. This is because if pass/fail or success testing is not abandoned then neither are many of the problems or debates that come with it.

For example, assume that Bastos and Taylor's (2020) first experiment, despite what is claimed, does not provide incontrovertible evidence that kea are using relative frequencies to make these predictions and that there are still a number of live hypotheses on the table that go over and beyond the one alternative they consider which was focused on absolute frequencies.[192] Instead assume that the experiment is significantly underdetermined in a way that has been shown to be common in many research programs in comparative cognitive science. Does this count as a signature of domain general statistical inference? Seemingly not in any sort of straightforward way.

[191] Bastos and Taylor claim to have controlled for associative learning by analyzing first trial performance on the three experiments, and also claim to have ruled out evidence of the presence of learning effects or the use of low-level associative strategies. While these are interesting controls, they in no way exhaust the space of possible associative learning strategies.

[192] There are of course many other challenges that this experiment faces that prevents it from providing "conclusive evidence."

The broader problem is that what counts as a signature will frequently be a site of dispute because success testing is still being relied upon to produce evidence of the presence of signatures. And importantly, as Taylor et al. (2022) admit, success testing typically engenders dispute and not progress. This seems somewhat uncontroversial, and in a certain way, this should come as no surprise. This is because in a somewhat indirect way, signature testing entails doubling down on the existing methods of the discipline. This is not necessarily a bad thing, but progress cannot be achieved by skipping over potentially contentious claims or bypassing unavoidable debates, regardless of how messy and heated they might be. How can a collection of underdetermined and underspecified hypotheses and theories be used to produce clear inferences that can constrain the space of hypotheses at the higher level?[193] If signature testing cannot do this, then there is good reason to think that the debates that success testing faced, at least in many cases, will reemerge.[194]

However, this is not the only possible interpretation of this approach. Another might not rely on success testing alone but instead would attempt to make broad use of the data that is produced in any given experiment. Rather than solely engaging in significance testing where data regarding an animal's ability to pass or fail a task is used, broad swaths of data regarding these experiments could be drawn upon, in additional to observational data, in order to identify patterns and biases and produce as many constraints as possible. If there was a computational model constructed from an overarching theory that was able to predict the presence of a pattern of bias or error, this would be a promising approach. However, this in no way represents the epistemic situation in most or even any research programs in comparative cognition. If instead data on patterns of bias or error is collected and then a model is fit with unconstrained free parameters to that data, then there is a serious concern about overfitting or fudging. Addressing concerns about wasting data in this way would explicitly come at too high of a price.

[193] It is assumed here that the collection of hypotheses and theories do not bear directly on one another.
[194] One can also forecast new controversies arising in so far as there is no a-theoretical way to weight individual signatures. This is particularly the case insofar as signatures simply are either solely individual experiments or observations, or they are composed of them. As has been shown, the discipline largely struggles to derive clear evidence from these individual experiments.

Another complication can be found in Taylor et al.'s core diagnosis. The many-to-one mapping problem that they identify is experimental in origin and it focused on the underdetermination of theory and hypotheses by data. The idea is that behavior that is exhibited in any given experimental context can be plausibly mapped onto numerous cognitive capacities in so far as the experiments that are run in comparative cognition are characteristically underdetermined. As has been argued in the discussion of the validity cycle above, however, this is only one part of the problem. The theories regarding the cognitive capacities that are being tested are underspecified and subject to persistent debate around their characterization. This means that even if there was a one-to-one correspondence between behavior and the cognitive capacities that are currently on offer, and admittedly it is hard to imagine what this might look like, problems would remain. Again, this is because of the status of theory in the discipline, and it is a problem that experimental evidence alone cannot possibly solve. In this way, debate, however annoying it might be, cannot be simply bypassed.

Signature testing is vaguely reminiscent of other approaches to the study of cognition such as cognitive marker approaches, cluster approaches, and the natural kind approaches. These share the property of abandoning the search for a single necessary condition.[195] This aspect of the approaches is one that should be clearly embraced. Moreover, any approach that provides the tools to perform more detailed ways of analyzing the available evidence is surely a welcome one. However, there might be reason to think that this does not fully capture the problem that the discipline is facing. While some comparative cognitive scientists hang too much on individual experiments, and overinflated claims are undoubtedly commonplace, not everyone in the discipline is guilty of doing this. Doing so risks transmorphing the discipline into an unrepresentative monolith. In particular, the philosophy of animal minds has at times held a laser like focus on Daniel Povinelli and his often heated and exagerated commentaries. While there is nothing wrong with critiquing influential individual scientists who make strong skeptical claims, it is a mistake to do this at the risk of homogenizing a diverse discipline. For example, in the mindreading debates, Tomasello, as well as his colleagues at the time in Leipzig, were looking at

[195] These approaches are likewise seemingly congruent with recent work by Dacey (forthcoming) who argues against the overburdening of individual experiments, and that by separating out statistical from substantial hypotheses (see Bausman and Halina 2018), we can make the evidential basis for our inferences more explicit.

evidence from a variety of sources simultaneously and assigning a particular weight to these collections of experiments as a whole, in addition to drawing on constraints from adjacent sciences, in order to infer the presence of mindreading. While they were also referring to extra empirical theoretical virtues, this weighting and interpretation of the 'total evidence' was exactly what was at stake in that debate and flickers of this can be seen other research programs. When Tomasello and colleagues talk about a variety of sources of evidence providing confirmatory evidence, in a certain way they are doing exactly what Taylor et al. are proposing. However, this does not seem to have been obviously progress generating in itself. Instead, there was a lot of debate around how to weigh the evidence and interpret the experiments. In this way, the path forward is slow, long, and hard fought but it might end up being the only option.

3.5 Four Modified Complementary Proposals

In their modified versions, each of these proposals have something substantial to offer the science. These modifications are briefly recapped below and are best summarized as a series of recommendations and lessons that hold the potential to be run simultaneously in a complementary manner with an aim to generating maximum progress in the discipline.

(1) 'Top-down' and 'bottom-up' approaches need to be disambiguated into five types in order to be intelligible. Moreover, the set of 'top-down' and/or 'bottom-up' approaches that can be adopted in any given context needs to be sensitive to the epistemic state of the research program in question, and multiple approaches must be adopted simultaneously in a complementary manner. It does very little to say that a research program should investigate the cognitive mechanism underlying a capacity if it is in no way epistemically situated to do this and it might turn out that work on higher primates should be stepped back from at least temporarilty.

(2) Dimensional approaches have the potential to generate more fine-grained analyses of the target cognitive capacities and to generate progress. However, for them to be able to do this they need to have certain features. (a) The goals of dimensional analyses should be made explicit and contestable. (b) Assuming the appropriate goals, the lower-level

dimensions that are intended to account for the higher-level capacity should be at least less mysterious than the higher-level capacity. (c) Assuming the appropriate goals, the capacity that is being broken down into dimensions should be retained and properly represented in the lower-level dimensions. (d) Assuming the appropriate goals, absent a defined relation between the dimensions, any possible combinations of values of the dimensions should be seen as capturing the higher-level capacity. (e) Assuming the appropriate goals, dimensional analyses of cognitive capacities need to behave like dimensional analyses in other scientific contexts.

(3) While the formalization of theories in comparative cognition holds the potential to contribute to theory development that is much needed in particular research programs, the blanket application of this strategy to the entire science is not particularly promising. Instead, formalization needs to be performed with a sensitivity to contexts that are amenable to it and by working in a back-and-forth iterative manner with verbal theories.

(4) Changing how testing is done in the science of comparative cognition can be a promising approach in certain research contexts. However, this can only be done by taking on the challenges that success testing faces. Moreover, given that there is no agreed upon or systematic way to evaluate bodies of evidence, this approach should be employed with caution.

There are obvious commonalities between these approaches. In the most general sense, the idea is to introduce more precision into the science so as to generate more determinate evidence. There seems to be broad agreement on the part of critics and those that are interested in reforming the discipline that the current research is too coarse and/or vague and that what is needed is generally more detail and precision. This could be achieved in terms how we theorize the object of study or the tools that are used to perform the inquiry. Given this, there seems to be a natural affinity between these proposals and given that they each have shortcomings in their initial states, there is the opportunity for them to complement one another in their modified states.

The advantage of running each of these approaches simultaneously, and in a complementary manner should be clear. However, some approaches will be more complementary than others and determining this is going to be a highly contextual task. While the validity cycle continues to pose substantial challenges to the discipline of comparative cognitive science, there is no reason to think that moving beyond it is fundamentally out of reach. However, this progress will not come for free or automatically, particularly given the current state of the recommendations for doing so. By implementing these proposals and running them, in so far as possible in a complementary manner, there is good reason to think that incremental progress can be achieved, and that the science can be securely placed on a path towards advancement.

4. Conclusion

In this section, a novel diagnosis of the problems that the discipline of comparative cognitive science has been repeatedly confronted with has been introduced: the validity cycle. It has been argued that on many in the discipline's own lights, these problems represent a substantial threat. A series of cases studies in comparative cognition were then introduced. Three were plausibly developed research program and one was still in the process of developing. It was argued that these represent instances of the validity cycle. The question of why this cycle arises was then investigated in more detailed and it was argued that every step of the empirical process in comparative cognition plausibly contributed to the introduction of the validity cycle; namely the theory that is employed in the discipline is weak, the hypotheses that are tested are underspecified, and experiments that are ill equipped to engage in hypothesis selection. Taken together, this places the science of comparative cognition in a plausibly tenuous state.

Then, four ameliorative proposals were analyzed and modifications to them were introduced. The **first** involves taking a bottom-up approach to the study of animal minds. The **second** involves breaking the capacity into dimensions. The **third** involves adopting formalization techniques in the discipline. The **fourth** involves shifting the way experimentation is done towards attempting to identify signatures of cognitive capacities. After critiquing and modifying each of these it was argued that running these approaches simultaneously and in a complementary manner is the most promising way forward.

Inferring the cognitive mechanisms underlying the behavior of non-human animals has proven to be extremely difficult; perhaps even more so than the scientists who initially turned towards the study of sophisticated cognitive capacities in animals could have ever imagined. The challenges posed by the validity cycle and its causes are serious, and a quick and clean resolution should not be expected or even pursued. Nevertheless, there is good reason to remain cautiously optimistic about the prospect of achieving substantial progress in the science of comparative cognition, even in the short term. While certain research programs will surely face more substantial challenges than others, adopting the various modified approaches that have been introduced, and running them simultaneously in a complementary manner, is far and away the most promising way forward for this nascent science.

3

CONCLUSION

Throughout this dissertation, a cautiously optimistic view of the state of comparative cognitive science has been developed and defended. It has been argued that with each step of the empirical process calibrations need to be performed and that this initial and necessarily high degree of uncertainty needs to be faced head-on if it is ever going to be adequately downgraded. Importantly, progress is not guaranteed and the path forward forks in innumerable ways. The difficulty lies in the fact that it is often not easy to determine which paths are dead-ends before going down them. Over the course of this dissertation, various contributions have been made to the task of separating dead-ends from open-roads through the introduction of several ameliorative changes to the practice of comparative cognitive science.

To recap, Part 1 addressed the issue of replication: both the looming threat of a replication crisis descending upon comparative cognitive science as well as the broader issue of what a replication is. A deflationary account of replication was introduced which holds direct implications for the practice of replication, demarcating replicability as a criterion, evaluating the replication crisis, and evaluating replication experiments and research programs. One upshot for comparative cognitive science is that it is not mature enough to experience a genuine replication crisis, but that replication experiments can still genuinely contribute to the generation of progress in particular research programs.

In Part 2, it was argued that comparative cognitive science, particularly the part of it that attempts to study 'sophisticated' cognitive capacities exclusively with behavioral methods, is generally burdened by what was termed the 'validity cycle'. Several causes of the validity cycle were identified: the theory that is employed in the discipline is weak, the hypotheses that are tested are

underspecified, and experiments that are performed are underdetermined. Four proposals that have been made and their shortcomings were then investigated more closely. The first involves taking a bottom-up approach to the study of animal minds. The second involves breaking the target capacity into dimensions. The third involves adopting formalization techniques in the discipline. The fourth involves shifting the way experimentation is done towards attempting to identify signatures of cognitive capacities.

Of course, these proposals in no way plausbily solve all the problems or threats that hang over the discipline, which are undeniably numerous. Although several claims have been defended regarding how to improve the science, part of the contribution of this work is located in laying bare several questions that need to be given more focused treatment in future research. In an ideal world, the science of comparative cognition would see a flood of resources, that it currently does not have access to, that would be put to use in a cautious and well-measured manner. In this world, many research programs could greatly benefit from larger sample sizes. A greater variety of experiments on a greater variety of species with a greater variety of background conditions could converge to form a stronger evidential base. While addressing this problem would in no way solve all the problems facing the discipline, it might be a necessary condition for obtaining the type of substantial progress that has long been sought after. Tragically, there is no reason to think that anything like this world will emerge in the near future. This leaves both the science, and those making proposals for how to proceed, in a hard spot.

On top of these empirical constraints, almost every aspect of this discipline requires deep philosophical work. This means that the general epistemic situation would likewise benefit from a massive number of resources being poured into this area as well. While the philosophy of comparative cognitive science has gained increased attention recently, this somewhat esoteric subdiscipline of the philosophy of science is still very much a modest one. This change, however, is well within the realm of possibility. Nevertheless, at a certain point, progress on the philosophical front is contingent upon progress on the empirical front. There is no obvious way to sever one from the other.

Given this, a realistic outlook on the types of research that should be expected from the discipline in the near future should be adopted give that there is no reason to think that an extreme number of resources will suddenly be poured in the comparative cognitive science. This is particularly the case regarding studies of primates and the larger mammals that have been the target of experiments regarding sophisticated cognitive capacities in so far as ethical considerations have in recent times increasingly placed blocks on this work. If anything, the heyday of this type of research seems to have already passed, and even if it were possible, pouring a lot of resources into the study of sophisticated cognitive capacities in very complicated animals that have bizarre histories might not be the most promising way to most quickly advance the science, as was argued in Part 2 of the dissertation.

Moreover, given the high degree of uncertainty that characterizes this research, and the types of inferences that can be drawn from it, there might be good ethical reasons to avoid this type of work all together; at least temporarily.[196] This comes down to the cost of performing such research and the ethical risks that comes with holding these animals in captivity. If the quality of the research that is gained from such studies is weak in the way that has been argued throughout this dissertation, it may turn out that taking on these ethical risks is unjustifiable, given the prospective benefits.

Fundamental theoretical and methodological issues have repeatedly held comparative cognition back from making more substantial contributions to the sciences of the mind. The discipline has too often been snared in false starts and verbal disputes. Rather than continuing down the same path experimentalists have been steadfastly driving towards, a more grounded and promising approach that has been defended in this dissertation should be adopted. In all too many cases, experimentalists have asked more of behavioral research than it is plausibly able to give. However, its force is only truly unlocked when it is placed within a large and diverse context of research. In the 70 years since the cognitive revolution, scientists have developed a wide variety of methods in a variety of disciplines have been used to study cognitive phenomenon. Going forward, comparative cognitive science must locate itself within the broad and diverse network that

[196] There are many complications and barriers to studying higher primates, particularly as regulations have been introduced (Suran and Wolinsky 2009).

consitutes the cognitive sciences. Doing this will allow the black box to eventually be cracked opened, for structure to be mapped to function, and for increasingly probative evidence to be obtained. It is in this way that substantial progress can be slowly and steadily achieved.

References:

Aarts H, Dijksterhuis A (2002) Category activation effects in judgment and behaviour: The moderating role of perceived comparability. British Journal of Social Psychology 41: 123–138.

Adam, T., Agafonova, N., Aleksandrov, A., Altinok, O., Sanchez, P. A., Anokhina, A., ... & Sahnoun, Z. (2012). Measurement of the neutrino velocity with the OPERA detector in the CNGS beam. Journal of High Energy Physics, 2012(10), 93.

Allen, C. (2014). Models, mechanisms, and animal minds. The Southern Journal of Philosophy, 52, 75-97.

Altmann SA. 1965 Sociobiology of rhesus monkeys. II: Stochastics of social communication. Journal of Theoretical Biology 8, 490–522.

Amodio, P., Farrar, B. G., Krupenye, C., Ostojic, L., & Clayton, N. S. (2021). Little evidence that Eurasian jays protect their caches by responding to cues about a conspecific's desire and visual perspective. ELife, 10, e69647.

Andrews, K. (2016). Pluralistic folk psychology in humans and other apes. In The Routledge handbook of philosophy of the social mind (pp. 133-154). Routledge.

Andrews, K. (2020). The animal mind: An introduction to the philosophy of animal cognition. routledge.

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. Psychological Methods, 21(1), 1.

Arbilly, M., & Laland, K. N. (2017). The magnitude of innovation and its evolution in social animals. Proceedings of the Royal Society B: Biological Sciences, 284(1848), 20162385.

Anjum, R. L., & Mumford, S. (2018). Causation in science and the methods of scientific discovery. Oxford University Press, USA.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. European journal of personality, 27(2), 108-119.

Atance, C. M., & O'Neill, D. K. (2001). Episodic future thinking. *Trends in cognitive sciences*, *5*(12), 533-539.

Bacon, F. (2000). Novum organon (Vol. 1620). London.

Baillargeon, R. (2004): 'Infants' Reasoning about Hidden Objects: Evidence for Event- General and Event-Specific Expectations', Developmental Science, 7, pp. 391–414.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature News, 533(7604), 452.

Bargh JA, Chen M, Burrows L (1996) Automaticity of social behavior: direct effects of trait construct and stereotype-activation on action. Journal of Personality and Social Psychology 71: 230–244.

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

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. Behavioral and Brain Sciences, 30(3), 241-254.

Barker, K. B., & Povinelli, D. J. (2019). Anthropomorphomania and the Rise of the Animal Mind: A Conversation. Journal of Folklore Research, 56(2-3), 71-90.

Barone, P., Corradi, G. and Gomila, A. [2019]: 'Infants' Performance in Spontaneous- Response False Belief Tasks: A Review and Meta-analysis', Infant Behavior and De- velopment, 57, p. 101350.

Barrett, L. (2010). Too much monkey business. Grounding sociality, 219-236.

Basso, A. (2017). The appeal to robustness in measurement practice. Studies in History and Philosophy of Science Part a., 65–66, 57–66.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? Journal of Personality and Social Psychology, 74(5), 1252–1265. https://doi.org/10/bszcmm

Bausman, W., & Halina, M. (2018). Not null enough: pseudo-null hypotheses in community ecology and comparative psychology. Biology & Philosophy, 33(3), 1-20.

Beach, F. A. (1950). The snark was a boojum. American Psychologist, 5(4), 115.

Bennett, J. (1978). Some remarks about concepts. Behavioral and Brain Sciences, 4, 557–560.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Johnson, V. E. (2018). Redefine statistical significance. Nature human behaviour, 2(1), 6-10.

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. Nature, 483(7391), 531-533.

Bermúdez, J. L. (2005). The phenomenology of bodily awareness. Phenomenology and philosophy of mind, 295-322.

Besbris, M., & Khan, S. (2017). Less Theory. More Description. Sociological Theory, 35(2), 147-153.

Birch, J., Schnell, A. K., & Clayton, N. S. (2020). Dimensions of animal consciousness. Trends in cognitive sciences, 24(10), 789-801.

Birch, J. (2022). The search for invertebrate consciousness. Noûs, 56(1), 133-153.

Bird, A., and E.Tobin. (2012). Natural kinds. In E. Zalta (Ed.), The stanford encyclopedia of philosophy (Winter 2012 Edition). Online at: http://plato.stanford.edu/archives/win2012/entries/natural-kinds/.

Bird, A. (2020). Understanding the replication crisis as a base rate fallacy. The British Journal for the Philosophy of Science.

Bird, C. D., & Emery, N. J. (2009). Rooks use stones to raise the water level to reach a floating worm. Current Biology, 19(16), 1410-1414.

Bischof-Köhler, D. (1985). Zur phylogenese menschlicher motivation.

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

Boeckle, M., Schiestl, M., Frohnwieser, A., Gruber, R., Miller, R., Suddendorf, T., ... & Clayton, N. S. (2020). New Caledonian crows plan for specific future tool use. Proceedings of the Royal Society B, 287(1938), 20201490.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. The philosophical review, 97(3), 303-352.

Bohn M, Tessler MH, Merrick M, Frank MC. 2021 How young children integrate information sources to infer the meaning of words. Nature Human Behaviour 5, 1046–1054.

Bohn, M., Liebal, K., Oña, L., & Tessler, M. H. (2022). Great ape communication as contextual social inference: a computational modelling perspective. Philosophical Transactions of the Royal Society B, 377(1859), 20210096.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. Psychological review, 111(4), 1061.

Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. Journal of Educational Measurement, 50(1), 110-114.

Boyle, A. (2018). Mirror Self-Recognition and Self-Identification. Philosophy and Phenomenological Research, 97(2), 284-303.

Boyle, A. (2020). The impure phenomenology of episodic memory. *Mind & Language*, *35*(5), 641-660.

Boyle, A. (2021). Replication, uncertainty and progress in comparative cognition.

Bradley, B. (2015). Is death bad for a cow? In T. Višak & R. Garner (Eds.), The ethics of killing animals (pp. 51–63). New York: OUP.

Bradley, R. (2017). Decision theory with a human face. Cambridge University Press.

Braude, S. E. (1979). ESP and Psychokinesis: a Philosophical Examination. Philadelphia, PA: Temple University Press.

Brooks, J. X., & Cullen, K. E. (2019). Predictive sensing: The role of motor signals in sensory processing. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 4(9), 842-850.

Brown, T. L. (2009). Chemistry: the central science. Pearson Education.

Brown, R. L. (2019). Infer with care: A critique of the argument from animals. Mind & Language, 34(1), 21-36.

Brown, R. L. (2021). Mapping Out the Landscape: A Multi-dimensional Approach to Behavioural Innovation. Philosophy of Science, 1-23.

Brownstein, Michael, "Implicit Bias", The Stanford Encyclopedia of Philosophy (Fall 2019 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/fall2019/entries/implicit-bias/

Brunswik E. (1955). Representative design and probabilistic theory in a functional psychology. Psychological Review, 62, 193–217.

Buckingham, B. R. (1921). Intelligence and its measurement: A symposium--XIV. Journal of Educational Psychology, 12(5), 271.

Buckner, C. (2014). The semantic problem (s) with research on animal mind-reading. Mind & Language, 29(5), 566-589.

Bugnyar, T., Reber, S. A., & Buckner, C. (2016). Ravens attribute visual access to unseen competitors. Nature communications, 7(1), 1-6.

Buttliere, B., & Wicherts, J. (2018). Opinions on the value of direct replication: A survey of 2,000 psychologists.

Carney, D. R., Cuddy, A. J., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. Psychological science, 21(10), 1363-1368.

Carr, K., Kendal, R. L., & Flynn, E. G. (2016). Eureka!: What is innovation, how does it develop, and who does it?. Child development, 87(5), 1505-1519.

Cartwright, N. (2009). Evidence-based policy: what's to be done about relevance?. Philos Stud 143, 127–136 https://doi.org/10.1007/s11098-008-9311-

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. New England Journal of Medicine, 299(18), 999-1001.

Cesario J, Plaks JE, Higgins ET (2006) Automatic social behavior as motivated preparation to interact. Journal of Personality and Social Psychology 90: 893.

Chang, H. (2004). Inventing temperature: Measurement and scientific progress. Oxford University Press.

Cheke, L. G., Bird, C. D., & Clayton, N. S. (2011). Tool-use and instrumental learning in the Eurasian jay (Garrulus glandarius). Animal Cognition, 14(3), 441-455.

Cheeseman, J. F., Millar, C. D., Greggers, U., Lehmann, K., Pawley, M. D., Gallistel, C. R., ... & Menzel, R. (2014). Way-finding in displaced clock-shifted bees proves bees use a cognitive map. Proceedings of the National Academy of Sciences, 111(24), 8949-8954.

Chittka, L., Rossiter, S. J., Skorupski, P., & Fernando, C. (2012). What is comparable in comparative cognition?. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1603), 2677-2685.

Clayton, N. S., & Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. Nature, 395(6699), 272-274.

Clayton, N. S., & Wilkins, C. (2018). Seven myths of memory. Behavioural processes, 152, 3-9.

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, H. M. [1984]: 'When Do Scientists Prefer to Vary Their Experiments?', Studies in History and Philosophy of Science Part A, 15, pp. 169–74.

Collins, H. M. [1985]: Changing Order, London: Sage.

Collins, H. M. (1994). A strong confirmation of the experimenters' regress. Studies in History and Philosophy of Science part A, 25(3), 493-503.

Corballis, M. C. (2010). Mirror neurons and the evolution of language. Brain and language, 112(1), 25-35.

Corballis, S. T., & Suddendorf, T. (1997). Mental time travel and the evolution of the human mind. Genet. Soc. Gen. Psychol. Monogr, 123, 133-167.

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.

Crapse, T. B., & Sommer, M. A. (2008). Corollary discharge across the animal kingdom. Nature Reviews Neuroscience, 9(8), 587-600.

Craver, C. F. (2007). Explaining the brain: Mechanisms and the mosaic unity of neuroscience. Clarendon Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological bulletin, 52(4), 281.

Cronbach, L. J. (1982). In Praise of Uncertainty. New directions for program evaluation.

Dacey (forthcoming) Of a Different Mind

Dacey, M. (2022). "Naturalistic Epistemology," The Internet Encyclopedia of Philosophy, ISSN 2161-0002, https://iep.utm.edu/associationism-in-philosophy-of-mind/

Dacey, M. (2023). Evidence in default: Rejecting default models of animal minds. The British Journal for the Philosophy of Science, 74(2), 000-000.

Dally, J. M., Emery, N. J., & Clayton, N. S. (2004). Cache protection strategies by western scrub–jays (Aphelocoma californica): Hiding food in the shade. Proceedings of the Royal Society of London. Series B: Biological Sciences, 271(suppl_6), S387-S390.

Dally, J. M., Emery, N. J., & Clayton, N. S. (2005). Cache protection strategies by western scrub-jays, Aphelocoma californica: implications for social cognition. Animal Behaviour, 70(6), 1251-1263.

Dahman, Y. (2017). Nanotechnology and functional materials for engineers. Elsevier.

Darwin, C. (1873). Origin of certain instincts.

DeGrazia, D. (1996). Taking animals seriously: Mental life and moral status. Cambridge, MA: CUP.

Dennett, D. C. Content and Consciousness. London: Routledge and Kegan Paul, 1969.

Dennett, D. (1978). Beliefs about beliefs. Behavioral and Brain Sciences, 4, 568–570.

Dennett, D. C. (1983). Intentional systems in cognitive ethology: The "Panglossian paradigm" defended. Behavioral and Brain Sciences, 6(3), 343-355.

Derksen, M. (2019). Putting Popper to work. Theory & Psychology, 29(4), 449-465.

De Nigris, A., Piccardi, L., Bianchini, F., Palermo, L., Incoccia, C., & Guariglia, C. (2013). Role of visuo-spatial working memory in path integration disorders in neglect. Cortex, 49(4), 920-930.

De Waal, F. B., & Ferrari, P. F. (2010). Towards a bottom-up perspective on animal and human cognition. Trends in cognitive sciences, 14(5), 201-207.

Dominus, S. (2017). When the revolution came for Amy Cuddy. The New York Times, 29.

Dorrenberg, S., Rakoczy, H. and Liszkowski, U. [2018]: 'How (Not) to Measure Infant Theory of Mind: Testing the Replicability and Validity of Four Non-verbal Measures', Cognitive Development, 46, pp. 12–30.

Douglas, H. (2009). Science, policy, and the value-free ideal. University of Pittsburgh Press.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind?. PloS one, 7(1), e29081.

Dretske, F. (1997). Naturalizing the mind. mit Press.

Duhem, P. M. M. (1954 [1906]). The aim and structure of physical theory (trans: Weiner, P.P.). Princeton: Princeton University Press.

Drummond, C. (2009). Replicability is not reproducibility: nor is it good science.

Dyer, F. C., & Dickinson, J. A. (1996). Sun-compass learning in insects: Representation in a simple mind. Current Directions in Psychological Science, 5(3), 67-72.

Eaton, T., Hutton, R., Leete, J., Lieb, J., Robeson, A., & Vonk, J. (2018). Bottoms-up! Rejecting top-down human-centered approaches in comparative psychology. International Journal of Comparative Psychology, 31.

Eckert, J., Rakoczy, H., Call, J., Herrmann, E., & Hanus, D. (2018). Chimpanzees consider humans' psychological states when drawing statistical inferences. Current Biology, 28(12), 1959-1963.

Emery, N. J., & Clayton, N. S. (2001). Effects of experience and social context on prospective caching strategies by scrub jays. Nature, 414(6862), 443-446.

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.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. Elife, 10, e71601.

Evangelista, C., Kraft, P., Dacke, M., Labhart, T., & Srinivasan, M. V. (2014). Honeybee navigation: critically examining the role of the polarization compass. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1636), 20130037.

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.

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.

Farrar, B., & Ostojic, L. (2019). The illusion of science in comparative cognition

Farrar, B. G., Ostojic, L., & Clayton, N. S. (2021). The hidden side of animal cognition research: Scientists' attitudes toward bias, replicability and scientific practice. PloS one, 16(8), e0256607.

Feest, U. (2012). Exploratory experiments, concept formation, and theory construction in psychology. Scientific concepts and investigative practice, 3, 167-189.

Feest, U., & Steinle, F. (2016). Experiment. In Paul Humphreys (ed.), The Oxford Handbook of Philosophy of Science. Oxford University Press. pp. 274-295 (2016)

Feest, U. (2016). The experimenters' regress reconsidered: Replication, tacit knowledge, and the dynamics of knowledge generation. Studies in History and Philosophy of Science Part A, 58, 34-45.

Feest, U. (2019). Why replication is overrated. Philosophy of Science, 86(5), 895-905.

Feest, U. (2020). Construct validity in psychological tests–the case of implicit social cognition. European Journal for Philosophy of Science, 10(1), 1-24.

Feest, U. (2022). Data quality, experimental artifacts, and the reactivity of the psychological subject matter. European Journal for Philosophy of Science, 12(1), 1-25.

Fidler, F., & Wilcox, J. (2018). Reproducibility of scientific results.'

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

Fitch WT. 2010 The evolution of language. Cambridge University Press.

Fleck, L. (1935). Genesis and Development of a Scientific Fact, Chicago and London, The university of Chicago Press, 1979, (translation by Fred Bradley & Thaddeus J. Trenn).

Fletcher, L., & Carruthers, P. (2013). Behavior-reading versus mentalizing in animals. Agency and joint attention, 82-99.

Fletcher, S. C. (2021). The role of replication in psychological science. European Journal for Philosophy of Science, 11(1), 1-19.

Fletcher, S. C. (2021). How (not) to measure replication. European Journal for Philosophy of Science, 11(2), 1-27.

Franklin, A. (1999). Can that be right?: Essays on experiment, evidence, and science. Boston: Kluwer.

Franklin, L. R. (2005). Exploratory experiments. Philosophy of Science, 72(5), 888-899.

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. Psychological Inquiry, 31(4), 271-288.

Frigg, R., & Nguyen, J. (2020). Modelling nature: An opinionated introduction to scientific representation. Cham: Springer.

Friston, K. (2010). The free-energy principle: a unified brain theory?. Nature reviews neuroscience, 11(2), 127-138.

Frolov, S. (2021). Quantum computing's reproducibility crisis: Majorana fermions.

Galitsky, B., & Galitsky, B. (2016). Formalizing theory of mind. Computational Autism, 95-176.

Gallistel, C. R. (1990). The organization of learning. The MIT Press.

Gallup, G. G. (1970). Chimpanzees, self-recognition Science 167. BE-87.

Gallup Jr, G. G., & Anderson, J. R. (2018). The "olfactory mirror" and other recent attempts to demonstrate self-recognition in non-primate species. Behavioural Processes, 148, 16-19.

Gärdenfors, P., & Lombard, M. (2018). Causal cognition, force dynamics and early hunting technologies. Frontiers in Psychology, 9, 87.

Gershman, S. J. (2019). How to never be wrong. Psychonomic Bulletin & Review, 26(1), 13-28.

Gibbings, J. C. (2011). Dimensional analysis. Springer Science & Business Media.

Glen, William (1982), The Road to Jamarillo. Stanford: Stanford University Press.

Glenberg, A. M. (1997). What memory is for. Behavioral and brain sciences, 20(1), 1-19.

Godfrey-Smith, Peter, 2007, "The Strategy of Model-Based Science", Biology & Philosophy, 21(5): 725–740. doi:10.1007/s10539-006-9054-6

Godfrey-Smith, P. (2016). Philosophy of biology (Vol. 8). Princeton University Press.

Gómez, O. S., Juristo, N., & Vegas, S. (2010). Replications types in experimental disciplines. In Proceedings of the 2010 ACM-IEEE international symposium on empirical software engineering and measurement (pp. 1-10).

Good, I. J. (1960). The paradox of confirmation. The British Journal for the Philosophy of Science, 11(42), 145-149.

Goodall, J. (1968). Behaviour of free-living chimpanzees of the Gombe Stream area. Animal Behaviour Monograph, 1, 163–311.

Goodall, J. (1970). Tool-using in primates and other vertebrates. In D. S. Lehrmann, R. A. Hinde & E. Shaw (eds.) Advances in the Study of Behavior, Vol. 3 (pp. 195–249). New York: Academic Press

Goodman, N. (1972). Seven strictures on similarity.

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

Grah, G., Wehner, R., & Ronacher, B. (2005). Path integration in a three-dimensional maze: ground distance estimation keeps desert ants Cataglyphis fortis on course. Journal of experimental biology, 208(21), 4005-4011.

Greenwood, J. D. (2016). All the way up or all the way down?: Some historical reflections on theories of psychological continuity. Journal of Comparative Psychology, 130(3), 205.

Grice, H. P. (1975). Logic and conversation. In Speech acts (pp. 41-58). Brill.

Griffiths, D., Dickinson, A., & Clayton, N. (1999). Episodic memory: what can animals remember about their past?. *Trends in cognitive sciences*, *3*(2), 74-80.

Gulz, A. E. (1992). The planning of action as a cognitive and biological phenomenon.

Guttinger, S. (2020). The limits of replicability. European Journal for Philosophy of Science, 10(2), 1-17.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., … Zwienenberg, M. (2016). A Multilab Preregistered Replication of the Ego- Depletion Effect. Perspectives on Psychological Science, 11(4), 546–573. https://doi.org/10.1177/1745691616652873

Halina, M. (2015). There is no special problem of mindreading in nonhuman animals. Philosophy of Science, 82(3), 473-490.

Halina, M. (2021). Replications in comparative psychology. Animal Behavior and Cognition, 8(2), 263-272.

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. Animal Behaviour, 59(4), 771-785.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know?. Animal behaviour, 61(1), 139-151.

Harman, G. (1978). Studying the chimpanzee's theory of mind. Behavioral and Brain Sciences, 4, 576–577.

Harman, E. (2011). The moral significance of animal pain and animal death. In T. Beauchamp & R. Frey (Eds.), The Oxford handbook of animal ethics. New York: OUP.

Haugeland, J. (1989). Artificial intelligence: The very idea. MIT press.

Heinze, S., & Homberg, U. (2007). Maplike representation of celestial E-vector orientations in the brain of an insect. Science, 315(5814), 995-997.

Hemelrijk, C. K., & Bolhuis, J. J. (2011). A minimalist approach to comparative psychology. Trends in Cognitive Sciences, 15(5), 185-186.

Hempel, C. G. (1945). Studies in the Logic of Confirmation (I.). Mind, 54(213), 1-26.

Hempel, C. G. (1966). Philosophy of Natural Science. Printice Hall. Inc., Englewood Cliffs.

Hendricks, M., Ha, H., Maffey, N., & Zhang, Y. (2012). Compartmentalized calcium dynamics in a C. elegans interneuron encode head movement. Nature, 487(7405), 99-103.

Hendricks, M., & Zhang, Y. (2013, July). Complex RIA calcium dynamics and its function in navigational behavior. In Worm (Vol. 2, No. 3, pp. 99-103). Taylor & Francis.

Hennefield, L., Hwang, H. G., Weston, S. J., & Povinelli, D. J. (2018). Meta-analytic techniques reveal that corvid causal reasoning in the Aesop's Fable paradigm is driven by trial-and-error learning. Animal cognition, 21(6), 735-748.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. Nature, 466(7302), 29-29.

Heesen, R., Bright, L. K., & Zucker, A. (2019). Vindicating methodological triangulation. Synthese, 196(8), 3067–3081.

Heyes, C. M. (1994). Reflections on self-recognition in primates. Animal Behaviour, 47(4), 909-919.

Heyes, C. M. (1995). Self-recognition in primates: further reflections create a hall of mirrors. Animal Behaviour, 50(6), 1533-1542.

Hinde, R. A., and J. Fisher. 1951. Further observations on the opening of milk bottles by birds. Br. Birds 44:393-396.

Hirvonen, I., & Karisto, J. (2022). Demarcation without Dogmas. Theoria.

Hüffmeier, Joachim, Jens Mazei, and Thomas Schultze. "Reconceptualizing replication as a sequence of different studies: A replication typology." Journal of Experimental Social Psychology 66 (2016): 81-92.

Hodos, W., & Campbell, C. B. G. (1969). Scala naturae: why there is no theory in comparative psychology. Psychological Review, 76(4), 337.

Hunt, G. R. (1996). Manufacture and use of hook-tools by New Caledonian crows. Nature, 379(6562), 249-251.

Hurley, S. (2003). Animal action in the space of reasons. Mind & Language, 18(3), 231-257.

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis.

Inoue, S., & Matsuzawa, T. (2007). Working memory of numerals in chimpanzees. Current Biology, 17(23), R1004-R1005.

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

Ioannidis, J. P. (2018). The proposal to lower P value thresholds to. 005. Jama, 319(14), 1429-1430.

Ioannidis, J. P. (2018). Why replication has more scientific value than original discovery. Behavioral and Brain Sciences, 41.

Inoue, S., & Matsuzawa, T. (2007). Working memory of numerals in chimpanzees. Current Biology, 17(23), R1004-R1005.

Irvine, E. (2021). The role of replication studies in theory building. Perspectives on Psychological Science, 16(4), 844-853.

Jelbert, S. A., Taylor, A. H., Cheke, L. G., Clayton, N. S., & Gray, R. D. (2014). Using the Aesop's fable paradigm to investigate causal understanding of water displacement by New Caledonian crows. PloS one, 9(3), e92895.

Johnson, V. E., Payne, R. D., Wang, T., Asher, A. and Mandal, S. [2017]: 'On the Reproducibility of Psychological Science', Journal of the American Statistical Association, 112, pp. 1–10.

Johnstone, D. J. (1989). On the necessity for random sampling. The British Journal for the Philosophy of Science, 40(4), 443-457.

Kabadayi, C., & Osvath, M. (2017). Ravens parallel great apes in flexible planning for tool-use and bartering. *Science*, *357*(6347), 202-204.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological review, 80(4), 237.

Kamat, P., Vandenberghe, S., Christen, S., Bongoni, A. K., Meier, B., Rieben, R., & Khattab, A. A. (2016). Dexrazoxane shows no protective effect in the acute phase of reperfusion during myocardial infarction in pigs. Plos one, 11(12), e0168541.

Kawai, M. (1965). On the system of social ranks in a natural troop of Japanese monkeys (1): Basic rank and dependent rank. - In: Japanese monkeys, a collection of translations (S. A. ALTMANN, ed.). Atlanta, p. 66-86.

Kellert, S. H., Longino, H. E., & Waters, C. K. (Eds.). (2006). Scientific pluralism (Vol. 19). U of Minnesota Press.

Kiontke, K., & Fitch, D. H. (2013). Nematodes. Current Biology, 23(19), R862-R864.

Klein, S. B. (2013). Making the case that episodic recollection is attributable to operations occurring at retrieval rather than to content stored in a dedicated subsystem of long-term memory. Frontiers in behavioral neuroscience, 7, 3.

Klein, S. B. (2014). Autonoesis and belief in a personal past: An evolutionary theory of episodic memory indices. Review of Philosophy and Psychology, 5, 427-447.

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Hilgard, J., ... & Ratliff, K. A. (2022). Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. Collabra: Psychology, 8(1), 35271.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. Behavioral and brain sciences, 19(1), 1-17.

Krachun, C., Lurz, R., Mahovetz, L. M., & Hopkins, W. D. (2019). Mirror self-recognition and its relationship to social cognition in chimpanzees. Animal cognition, 22(6), 1171-1183.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. Neuron, 93(3), 480-490.

S. Kripke (1980) Naming and Necessity Basil Blackwell Oxford

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. Science, 354(6308), 110-114.

Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. Wiley Interdisciplinary Reviews: Cognitive Science, 10(6), e1503.

Kuhn, T. S. (1969). The structure of scientific revolutions (2, enlargedth ed.). Chicago & London: University of Chicago Press (1st ed. 1962).

Kuorikoski, J., & Marchionni, C. (2016). Evidential diversity and the triangulation of phenomena. Philosophy of Science, 83(2), 227–247.

Kummer, H., & Goodall, J. (1985). Conditions of innovative behaviour in primates. Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 308(1135), 203-214.

Kulke, L., von Duhn, B., Schneider, D. and Rakoczy, H. [2018]: 'Is Implicit Theory of Mind a Real and Robust Phenomenon? Results from a Systematic Replication Study', Psychological science, 29, pp. 888–900.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In Lakatos, I., &Musgrave, A. (Eds.) Criticism and the growth ofknowledge (pp. 91–196). Cambridge: Cambridge University Press

Lavelle, J. S. (2020). When a crisis becomes an opportunity: The role of replications in making better theories.

Lo, Y. H., Liao, C. T., Zhou, J., Rana, A., Bevis, C. S., Gui, G., ... & Miao, J. (2019). Multimodal x-ray and electron microscopy of the Allende meteorite. Science advances, 5(9), eaax3009.

Legg, E. W., & Clayton, N. S. (2014). Eurasian jays (Garrulus glandarius) conceal caches from onlookers. Animal Cognition, 17(5), 1223-1226.

Legg, E. W., Ostojic, L., & Clayton, N. S. (2016). Caching at a distance: a cache protection strategy in Eurasian jays. Animal Cognition, 19(4), 753-758.

Leibovici, L. (2001). Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial. Bmj, 323(7327), 1450-1451.

Leonelli, S. (2018). Rethinking reproducibility as a criterion for research quality. In Including a symposium on Mary Morgan: curiosity, imagination, and surprise. Emerald Publishing Limited.

Lind, J. (2018). What can associative learning do for planning?. Royal Society open science, 5(11), 180778.

Logan, C. J. (2016). How far will a behaviourally flexible invasive bird go to innovate?. Royal Society Open Science, 3(6), 160247.

Logan, C. J., Avin, S., Boogert, N., Buskell, A., Cross, F. R., Currie, A., ... & Montgomery, S. H. (2017). Beyond brain size. BioRxiv, 145334.

Lewis, D. (1979). Counterfactual dependence and time's arrow. Noûs, 455-476.

Lewis D. (1973) Counterfactuals. Oxford: Blackwell.

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional.

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. Meta-Psychology, 3.

Lemons, D. S. (2017). A student's guide to dimensional analysis. Cambridge University Press.

Limongelli, L., Boysen, S. T., & Visalberghi, E. (1995). Comprehension of cause€ ffect relations in a tool-using task by chimpanzees (Pan troglodytes). Journal of Comparative Psychology, 109(1), 18

Logan, C. J., Jelbert, S. A., Breen, A. J., Gray, R. D., & Taylor, A. H. (2014). Modifications to the Aesop's Fable paradigm change New Caledonian crow performances. PLoS One, 9(7), e103049.

Lurz, R. W. (2011). Mindreading animals: the debate over what animals know about other minds. MIT press.

Lurz, R. W., Kanet, S., & Krachun, C. (2014). Animal mindreading: A defense of optimistic agnosticism. Mind & Language, 29(4), 428-454.

Lurz, R. (2014). Does Comparative Animal Cognition Need to Be Saved by Cognitive Modeling?. The Southern Journal of Philosophy, 52, 98-108.

Lurz, R., Krachun, C., Mahovetz, L., Wilson, M. J., & Hopkins, W. (2018). Chimpanzees gesture to humans in mirrors: using reflection to dissociate seeing from line of gaze. Animal Behaviour, 135, 239-249.

Lykken, D. T. (1991). What's wrong with Psychology, anyway.> In D. Ghiccetti 8c W Grove (Eds.), Thinking clearly about psychology (pp. 3-39).

Lynch Jr, J. G., Bradlow, E. T., Huber, J. C., & Lehmann, D. R. (2015). Reflections on the replication corner: In praise of conceptual replications. International Journal of Research in Marketing, 32(4), 333-342.

Lyon, P., Keijzer, F., Arendt, D., & Levin, M. (2021). Reframing cognition: getting down to biological basics. Philosophical Transactions of the Royal Society B, 376(1820), 20190750.

Maatman, F. O. (2021). Psychology's Theory Crisis, and Why Formal Modelling Cannot Solve It.

Machery, E. (2017). Philosophy within its proper bounds. Oxford University Press.

Machery, E. (2020). What is a replication?. Philosophy of Science, 87(4), 545-567.

Machery, E. (2021). A mistaken confidence in data. European Journal for Philosophy of Science, 11(2), 1-17.

Machery, E. (2021). The alpha war. Review of Philosophy and Psychology, 12(1), 75-99.

Martin, C. B., & Deutscher, M. (1966). Remembering. *The Philosophical Review*, *75*(2), 161-196.

Martin, J.P., Guo, P., Mu, L., Harley, C.M., and Ritzmann, R.E. (2015). Central-complex control of movement in the freely walking cockroach. Curr. Biol. 25, 2795–2803.

Mayo, D. (2000). Experimental practice and an error statistical account of evidence. Philosophy of Science, 67(3), S193–S207.

Mayo, D. G. (2018). Statistical inference as severe testing. Cambridge, UK: Cambridge Univ. Press Access provided by Katholieke Universiteit Leuven-KU Leuven on, 10(25), 21.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean?. American Psychologist, 70(6), 487.

McGrew, W. C. (1992). Tool-use by free-ranging chimpanzees: the extent of diversity.

Meehl, 1978 P.E Meehl Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology Journal of Consulting and Clinical Psychology, 46 (1978), pp. 806-834

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. Psychological reports, 66(1), 195-244.

Merton, R. K. (1963). Resistance to the systematic study of multiple discoveries in science. European Journal of Sociology/Archives Européennes de Sociologie, 4(2), 237-282.

Michaelian, K. (2016). Mental time travel: Episodic memory and our knowledge of the personal past. MIT Press.

Michaelian, K. (2017). The Routledge Handbook of Philosophy of Memory. Routledge.

Michotte, A. (2017 [1963]). The perception of causality. Routledge.

Mikhalevich, I., Powell, R., & Logan, C. (2017). Is behavioural flexibility evidence of cognitive complexity? How evolution can inform comparative cognition. Interface focus, 7(3), 20160121.

Mari Mikkola. Gender concepts and intuitions. Canadian Journal of Philosophy, 39(4):559–583, 2009

Mitchell, R. W. (2015). A critique of Stephane Savanah's "mirror self-recognition and symbol-mindedness". Biology & Philosophy, 30(1), 137-144.

Mittelstaedt, M. L., & Mittelstaedt, H. (1980). Homing by path integration in a mammal. Die Naturwissenschaften, 67(11), 566-567.

Monsó, S. (2019). How to tell if animals can understand death. Erkenntnis, 1-20.

Monsó, S., & Osuna-Mascaró, A. J. (2021). Death is common, so is understanding it: the concept of death in other species. Synthese, 199(1), 2251-2275.

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.

Moore, R. (2018). Gricean communication, language development, and animal minds. Philosophy Compass, 13(12), e12550.

Morrison, S. J. (2014). Reproducibility project: cancer biology: time to do something about reproducibility. Elife, 3, e03981.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., ... & Ioannidis, J. P. (2017). A manifesto for reproducible science. Nature human behaviour, 1(1), 1-9.

Muskens, R. (2000). Underspecified semantics. In Reference and Anaphoric Relations (pp. 311-338). Springer, Dordrecht.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. Nature Human Behaviour, 3(3), 221-229.

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

Naqshbandi, M., & Roberts, W. A. (2006). Anticipation of future events in squirrel monkeys (Saimiri sciureus) and rats (Rattus norvegicus): tests of the Bischof-Kohler hypothesis. Journal of Comparative Psychology, 120(4), 345.

Nishida, T., & Hiraiwa, M. (1982). Natural history of a tool-using behavior by wild chimpanzees in feeding upon wood-boring ants. Journal of Human Evolution, 11(1), 73-99.

Ness, D., & Calabrese, P. (2016). Stress effects on multiple memory system interactions. Neural plasticity, 2016.

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. Psychological methods, 18(3), 301.

Noble, W., & Davidson, I. (1996). Human evolution, language and mind: A psychological and archaeological inquiry. CUP Archive.

Norton, J. D. (2007) 'Causation as Folk Science', in H. Price and R. Corry (eds), Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited, Oxford: Oxford University Press.

Norton, J. D. (2015). Replicability of experiment. THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia, 30(2), 229-248.

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. Trends in cognitive sciences, 15(4), 152-159.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Yarkoni, T. S. C. I. E. N. T. I. F. I. C. S. T. A. N. D. A. R. D. S. (2015). Promoting an open research culture. Science, 348(6242), 1422-1425.

Nosek, B. A., & Errington, T. M. (2020). What is replication?. PLoS biology, 18(3), e3000691.

Nosek, B. A., & Errington, T. M. (2020). The best time to argue about what a replication means? Before you do it.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., ... & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. Annual review of psychology, 73, 719-748.

Nuijten, M. B., Bakker, M., Maassen, E., & Wicherts, J. (2018). Verify original results through reanalysis before replicating: a commentary on "Making replication mainstream" by Rolf A. Zwaan, Alexander Etz, Richard E. Lucas, & M. Brent Donnellan.

Nguyen, J. P., Shipley, F. B., Linder, A. N., Plummer, G. S., Liu, M., Setru, S. U., ... & Leifer, A. M. (2016). Whole-brain calcium imaging with cellular resolution in freely behaving Caenorhabditis elegans. Proceedings of the National Academy of Sciences, 113(8), E1074-E1081.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. Psychonomic bulletin & review, 26(5), 1596-1618.

Ogawa, H., & Miller, J. P. (2013). In vivo Ca 2+ imaging of neuronal activity. In Methods in Neuroethological Research (pp. 71-87). Springer, Tokyo.

Okasha, S. (2011). Experiment, observation and the confirmation of laws. Analysis, 71(2), 222-232.

Olsson, A., Knapska, E., & Lindström, B. (2020). The neural and computational systems of social learning. Nature Reviews Neuroscience, 21(4), 197-212.

Onishi, K. H. and Baillargeon, R. [2005]: 'Do 15-Month-Old Infants Understand False Beliefs?', Science, 308, pp. 255–8.

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

Osgood, J. M. (2017). Effect of Ego-Depletion Typing Task on Stroop Does Not Extend to Diverse Online Sample. Journal of Articles in Support of the Null Hypothesis, 13(2).

Ostojić L, Shaw RC, Cheke LG, Clayton NS. 2013. Evidence suggesting that desire-state attribution may govern food sharing in Eurasian jays. PNAS 110: 4123–4128. DOI: https://doi.org/10.1073/pnas.1209926110, PMID: 23382187

Ostojić L, Legg EW, Shaw RC, Cheke LG, Mendl M, Clayton NS. 2014. Can male Eurasian jays disengage from their own current desire to feed the female what she wants. Biology Letters 10: 20140042. DOI: https://doi.org/10.1098/rsbl.2014.0042, PMID: 24671829

Ostojić L, Legg EW, Dits A, Williams N, Brecht KF, Mendl M, Clayton NS. 2016. Experimenter expectancy bias does not explain Eurasian jays' (Garrulus glandarius) performance in a desire-state attribution task. Journal of Comparative Psychology 130: 407–410. DOI: https://doi.org/10.1037/com0000043, PMID: 27709968

Ostojić L, Legg EW, Brecht KF, Lange F, Deininger C, Mendl M, Clayton NS. 2017. Current desires of conspecific observers affect cache-protection strategies in California scrub-jays and Eurasian jays. Current Biology 27: R51–R53. DOI: https://doi.org/10.1016/J.CUB.2016.11.020, PMID: 28118584

Osvath, M., & Kabadayi, C. (2018). Contrary to the gospel, ravens do plan flexibly. Trends in Cognitive Sciences, 22(6), 474-475.

Ouellette, M. H., Desrochers, M. J., Gheta, I., Ramos, R., & Hendricks, M. (2018). A Gate-and-Switch Model for Head Orientation Behaviors in Caenorhabditis elegans. Eneuro, 5(6).

Patil, P., Peng, R. D., & Leek, J. T. (2016). A statistical definition for reproducibility and replicability. BioRxiv, 066803.

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.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. Perspectives on Psychological Science, 7(6), 531-536.

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1480), 731-744.

Penn, Derek C. (2011) "How Folk Psychology Ruined Comparative Psychology and What Scrub Jays Can Do about It." In Animal Thinking: Contemporary Issues in Comparative Cognition, ed. Menzel, Randolf and Fischer, Julia, 253–65. Cambridge, MA: MIT Press.

Peng, R. D. (2011). Reproducible research in computational science. Science, 334(6060), 1226-1227.

Penn, D. C., & Povinelli, D. J. (2013). The comparative delusion: the behavioristic/mentalistic dichotomy in comparative theory of mind research. In J. Metcalfe & H. S. Terrace (Eds.), Agency and Joint Attention (pp. 62–78). New York: Oxford University Press.

Piaget, J. (1963). The attainment of invariants and reversible operations in the development of thinking. Social research, 283-299.

Pirri, J. K., McPherson, A. D., Donnelly, J. L., Francis, M. M., & Alkema, M. J. (2009). A tyramine-gated chloride channel coordinates distinct motor programs of a Caenorhabditis elegans escape response. Neuron, 62(4), 526-538.a

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

Polanyi, M. (1958). Personal knowledge: Towards a post-critical philosophy. London: Routledge & K. Paul.

Polanyi, M. (1966). The tacit dimension. London: Routledge & K. Paul.

Popper, K. (1959 [2002]) The Logic of Scientific Discovery. London: Routledge

Poston, T. (2021). Explanatory coherence and the impossibility of confirmation by coherence. Philosophy of Science, 88(5), 835-848.

Povinelli, D. (2000). Folk physics for apes: The chimpanzee's theory of how the world works.

Povinelli, D. J., & Reaux, J. E. (2000). The trap-table problem. Folk physics for apes, 132-148.

Pravosudov, V. V. (2003). Long-term moderate elevation of corticosterone facilitates avian food-caching behaviour and enhances spatial memory. Proceedings of the Royal Society of London. Series B: Biological Sciences, 270(1533), 2599-2604.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. Behavioral and brain sciences, 1(4), 515-526.

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.

Radder, H. (Ed.). (2003). The philosophy of scientific experimentation. University of Pittsburgh Pre.

Ramsey, G., Bastian, M. L., & Van Schaik, C. (2007). Animal innovation defined and operationalized. Behavioral and Brain Sciences, 30(4), 393-407.

Rajala, A. Z., Reininger, K. R., Lancaster, K. M., & Populin, L. C. (2010). Rhesus monkeys (Macaca mulatta) do recognize themselves in the mirror: Implications for the evolution of self-recognition. PLoS One, 5(9), e12865.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. Psychological science, 26(5), 653-656.

Reader, S. M., & Laland, K. N. (2002). Social intelligence, innovation, and enhanced brain size in primates. Proceedings of the National Academy of Sciences, 99(7), 4436-4441.

Regan, T. (2004). The case for animal rights. (Updated with a New (Preface ed.). Berkeley: University of California Press.

Reaux, J. E., & Povinelli, D. J. (2000). The trap-tube problem. Folk physics for apes, 108-131.

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.

Redshaw, J., Taylor, A. H., & Suddendorf, T. (2017). Flexible planning in ravens?. *Trends in cognitive sciences*, *21*(11), 821-822.

Reed, Baron, "Certainty", The Stanford Encyclopedia of Philosophy (Spring 2022 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2022/entries/certainty/>.

Reichenbach, H. (1938). Experience and prediction: An analysis of the foundations and the structure of knowledge.

Rendell, L., Hoppitt, W., & Kendal, J. (2007). Is all learning innovation?. Behavioral and Brain Sciences, 30(4), 421-422.

Romero, F. (2019). Philosophy of science and the replicability crisis. Philosophy Compass, 14(11), e12633.

Romero, F. (2020). The division of replication labor. Philosophy of Science, 87(5), 1014-1025.

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

Roy, J. E., & Cullen, K. E. (2003). Brain stem pursuit pathways: dissociating visual, vestibular, and proprioceptive inputs during combined eye-head gaze tracking. Journal of neurophysiology, 90(1), 271-290.

Rutz, Christian, Shoko Sugasawa, Jessica E. M. van der Wal, Barbara C. Klump, and James J. H. St Clair. 2016. "Tool Bending in New Caledonian Crows." Royal Society Open Science 3 (8). https://doi.org/10.1098/RSOS.160439.

Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2010). Can there ever be too many options? A meta-analytic review of choice overload. Journal of consumer research, 37(3), 409-425

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.

Schliesser, E. (2019). Synthetic philosophy. Biology & Philosophy, 34(2), 1-9.

Seelig, J. D., & Jayaraman, V. (2015). Neural dynamics for landmark orientation and angular path integration. Nature, 521(7551), 186-191.

Sellars, W. (1956). Empiricism and the Philosophy of Mind. Minnesota studies in the philosophy of science, 1(19), 253-329.

Shapin, S. (2022). Hard science, soft science: A political history of a disciplinary array. History of Science, 60(3), 287-328.

Shaw, R. C., & Clayton, N. S. (2012). Eurasian jays, Garrulus glandarius, flexibly switch caching and pilfering tactics in response to social context. Animal Behaviour, 84(5), 1191-1200.

Smaldino, P. (2019). Better methods can't make up for mediocre theory. Nature, 575(7783), 9-10.

Sober, E. (2015). Ockham's razors. Cambridge University Press.

Soler, L. (2011). Tacit aspects of experimental practices: analytical tools and epistemological consequences. European Journal for Philosophy of Science, 1(3), 393-433.

Shettleworth, S. J. (2009). Cognition, evolution, and behavior. Oxford university press.

Shettleworth, S. J. (2010). Clever animals and killjoy explanations in comparative psychology. Trends in cognitive sciences, 14(11), 477-481.

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.

Spaulding, S. (2018). How we understand others: Philosophy and social cognition. Routledge.

Srinivasan, M., Zhang, S., Lehrer, M., & Collett, T. S. (1996). Honeybee navigation en route to the goal: visual flight control and odometry. The Journal of experimental biology, 199(1), 237-244.

Srinivasan, M.V. (2014). Going with the flow: a brief history of the study of the honeybee's navigational 'odometer'. J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol. 200, 563–573.

Stanford, P. K., & Kitcher, P. (2000). Refining the causal theory of reference for natural kind terms. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 97(1), 99-129.

Stanford, K. (2017). Underdetermination of Scientific Theory. In E. N. Zalta (ed.), The Stanford Encyclopedia of Philosophy (Winter 2017 Edition). The Metaphysics Research Lab. https://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination/

Starzak, T. B., & Gray, R. D. (2021). Towards ending the animal cognition war: a three-dimensional model of causal cognition. Biology & Philosophy, 36(2), 1-24.

Stegenga, J., & Menon, T. (2017). Robustness and independent evidence. Philosophy of Science, 84(3), 414-435.

Stevens, J. R. (2017). Replicability and Reproducibility in Comparative Psychology. Frontiers in Psychology, 8, 862. https://doi.org/10.3389/fpsyg.2017.00862

Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. Judgment and Decision making, 10(5), 479-491.

Stone, T., Webb, B., Adden, A., Weddig, N. B., Honkanen, A., Templin, R., ... & Heinze, S. (2017). An anatomically constrained model for path integration in the bee brain. Current Biology, 27(20), 3069-3085.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. Perspectives on Psychological Science, 9(1), 59-71.

Strunz, F. (1993). Preconscious mental activity and scientific problem-solving: A critique of the Kekulé dream controversy. Dreaming, 3(4), 281.

Suddendorf, T., & Busby, J. (2003). Mental time travel in animals?. *Trends in cognitive sciences*, *7*(9), 391-396.

Suddendorf, T., & Busby, J. (2005). Making decisions with the future in mind: Developmental and comparative identification of mental time travel. *Learning and Motivation*, *36*(2), 110-125.

Sullivan, J. A. (2009). The multiplicity of experimental protocols: a challenge to reductionist and non-reductionist models of the unity of neuroscience. Synthese, 167(3), 511-539.

Suran, M., & Wolinsky, H. (2009). The end of monkey research? New legislation and public pressure could jeopardize research with primates in both Europe and the USA. EMBO reports, 10(10), 1080-1082.

Tal, E. (2017). Measurement in science. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy (Fall 2017 ed.). https://plato.stanford.edu/archives/fall2017/entries/measurement-science/

Taylor, A. H., Elliffe, D. M., Hunt, G. R., Emery, N. J., Clayton, N. S., & Gray, R. D. (2011). New Caledonian crows learn the functional properties of novel tool types. PloS one, 6(12), e26887.

Taylor, A. H., Bastos, A. P., Brown, R. L., & Allen, C. (2022). The signature-testing approach to mapping biological and artificial intelligences. Trends in Cognitive Sciences.

Templer, V. L., & Hampton, R. R. (2013). Episodic memory in nonhuman animals. *Current Biology*, *23*(17), R801-R806.

Tosoni, A., Galati, G., Romani, G. L., & Corbetta, M. (2008). Sensory-motor mechanisms in human parietal cortex underlie arbitrary visual decisions. Nature neuroscience, 11(12), 1446-1453.

Tulving, E. (1972). Organization of memory. Episodic and semantic memory.

Tulving, E. (1983). Elements of episodic memory.

Tulving, E. (1986). What kind of a hypothesis is the distinction between episodic and semantic memory?.

Tulving, E., Schacter, D. L., Mclachlan, D. R., & Moscovitch, M. (1988). Priming of semantic autobiographical knowledge: A case study of retrograde amnesia. *Brain and cognition*, *8*(1), 3-20.

Tulving, E. (2002). Episodic memory: From mind to brain. Annual review of psychology, 53(1), 1-25.

van Rooij, I. (2019). Psychological science needs theory development before preregistration. See https://featuredcontent. psychonomic. org/psychological-science-needs-theory-development-before-preregistration.

van Rooij, I., & Blokpoel, M. (2020). Formalizing verbal theories: A tutorial by dialogue. Social Psychology, 51(5), 285.

Vickerstaff, R. J., & Cheung, A. (2010). Which coordinate system for modelling path integration?. Journal of Theoretical Biology, 263(2), 242-261.

Visalberghi, E., & Limongelli, L. (1994). Lack of comprehension of cause€ ffect relations in tool-using capuchin monkeys (Cebus apella). Journal of Comparative Psychology, 108(1), 15.

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

Vonk, J., & Povinelli, D. J. (2006). Similarity and difference in the conceptual systems of primates: The Unobservability hypothesis. In E. Wasserman & T. Zentall (Eds.), Oxford handbook of comparative cognition: Experimental explorations of animal intelligence (pp. 363–387). Oxford: Oxford University Press.

Vonk, J., & Shackelford, T. K. (2013). An introduction to comparative evolutionary psychology. Evolutionary psychology, 11(3), 147470491301100301.

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., ... & Zwaan, R. A. (2016). Registered replication report: strack, martin, & stepper (1988). Perspectives on Psychological Science, 11(6), 917-928.

Wallace, D. G., Hines, D. J., Pellis, S. M., & Whishaw, I. Q. (2002). Vestibular information is required for dead reckoning in the rat. Journal of Neuroscience, 22(22), 10009-10017.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05".

Wachtel, P. L. (1980). Investigation and its discontents: Some constraints on progress in psychological research. American Psychologist, 35(5), 399.

Weaver, I. C., Cervoni, N., Champagne, F. A., D'Alessio, A. C., Sharma, S., Seckl, J. R., ... & Meaney, M. J. (2004). Epigenetic programming by maternal behavior. Nature neuroscience, 7(8), 847-854.

Weir, A. A., Chappell, J., & Kacelnik, A. (2002). Shaping of hooks in New Caledonian crows. Science, 297(5583), 981-981.

Webb, B., & Consilvio, T. (Eds.). (2001). Biorobotics. Mit Press.

Wehner, R., & Srinivasan, M. V. (1981). Searching behaviour of desert ants, genus Cataglyphis (Formicidae, Hymenoptera). Journal of comparative physiology, 142(3), 315-338.

Wellman, H. M., Cross, D. and Watson, J. [2001]: 'Meta-analysis of Theory-of-Mind Development: The Truth about False Belief', Child Development, 72, pp. 655–84.

Wellman, H. M., Kushnir, T., Xu, F., & Brink, K. A. (2016). Infants use statistical sampling to understand the psychological world. Infancy, 21(5), 668-676.

Westra, E. (2017). Spontaneous mindreading: A problem for the two-systems account. Synthese, 194(11), 4559-4581.

Whiten, A. (1996). When does smart behaviour-reading become mind-reading? In P. Carruthers & P. K. Smith (Eds.), Theories of Theories of Mind (pp. 277–292). Cambridge: Cambridge University Press.

Whiten, A., Goodall, J., McGrew, W. C., Nishida, T., Reynolds, V., Sugiyama, Y., ... & Boesch, C. (1999). Cultures in chimpanzees. *Nature*, *399*(6737), 682-685.

White, J. G., Southgate, E., Thomson, J. N., & Brenner, S. (1986). The structure of the nervous system of the nematode Caenorhabditis elegans. Philos Trans R Soc Lond B Biol Sci, 314(1165), 1-340.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition, 13(1), 103-128.

Wimsatt, W. (1981). Robustness, Reliability, and Overdetermination. In M. Brewer & B. Collins (Eds.), Scientific Inquiry in the Social Sciences (pp. 123–162). Jossey-Bass.

Woodward, J. (1989). Data and phenomena. Synthese, 79, 393–472.

Woodward, J. (2000). Data, phenomena, and reliability. Philosophy of Science, 67, S163-S179.

Woodward, J. (2011). Psychological studies of causal and counterfactual reasoning. Understanding counterfactuals, understanding causation. Issues in philosophy and psychology, 16-53.

Yarkoni, T. (2022). The generalizability crisis. Behavioral and Brain Sciences, 45.

Zwaan, R.A., Etz, A., Lucas, R.E., Donnellan, M.B. (2018). Making replication mainstream. Behavioral and Brain Sciences, 41, e120.