When a Crisis Becomes an Opportunity: The Role of Replications in Making Better Theories

Jane Suilin Lavelle

Abstract

While it is widely acknowledged that psychology is in the throes of a replication 'crisis', relatively little attention has been paid to the role theory plays in our evaluation of replications as 'failed' or 'successful'. This paper applies well-known arguments in philosophy of science about the interplay between theory and experiment to a contemporary case study of infants' understanding of false belief (Onishi and Baillargeon [2005]), and attempts to replicate it. It argues that the lack of consensus about over-arching theories informing both the concepts under study and the methodologies used to track them means that researchers disagree over which experiments constitute replications of the original. The second part of the paper places this specific debate within a broader discussion of the replication crisis as a crisis of 'theory', developing work by Muthukrishna and Henrich ([2018]) and Alexander Bird ([2018]). Bird argues that the lack of agreed over-arching theories in psychology means that a high rate of replication failure is to be expected; this paper agrees with his diagnosis but challenges his proposal that more replication will resolve the problem.

1 Introduction

The psychological sciences are currently in the throes of a replication 'crisis' (Lilienfeld and Waldman [2017], McNutt [2014], Open Science Foundation [2015], Pashler & Wagenmakers [2012]). In the broadest sense, replicating an experiment consists in repeating the original experiment and gaining the same results. A failed replication is one where the results do not match those of the original. Failed replications can cast doubt on the veracity of the original findings, leading to allegations of bad practice such as HARK-ing (hypothesising after the results are known), p-hacking, selective data inclusion, and outright fabrication of the data (Cooper [2013]; Fanelli [2009]; Fang et al. [2012]; Kerr [1998]; Miller [2010]; Peterson [2016]). If failing to successfully repeat the results of another's experiment exclusively points to bad methods on behalf of the original experimenters, then the current run of failed replications certainly suggests a crisis in the psychological sciences.

While malpractice can be a causal factor for experiments failing to replicate (Fanelli [2009]), it certainly is not the only one. This has long been acknowledged by philosophers of science, who point out that the process of deciding whether Experiment B successfully replicates Experiment A is by no means straightforward (Collins [1985, 1989]; Feest [2016]). There are questions to be asked about how closely the methods of B follow those of A, and decisions to be made about how similar the results of the two experiments need to be in order for B to constitute a successful replication of A. Even if one has settled these questions satisfactorily, there are the further complications of assessing the theoretical consequences of B failing to replicate A. Were A's results a false positive or coincidence, or were B's? Which experiment gives the more accurate picture

of the phenomenon in question? Some of these complications are captured in what Harry Collins terms the 'Experimenters' Regress' (Collins [1985]). Imagine that researchers are working to discover whether an effect E exists. Research group R believe that E does exist, and conduct experiment A, the results of which appears to support their hypothesis that E exists. Research group R' does not believe in the existence of E, and conduct experiment B which is intended to repeat the methods of A. The results of B do not support the hypothesis that E exists. Further replications are conducted, some of which support the hypothesis that E exists, others contradicting it. The problem is that there is no consensus about what the right result of the experiment should be, as no-one yet knows for sure that E exists (or doesn't exist). And if we do not know what the right result should be, we cannot easily judge which experiments should be considered successful in getting the correct outcomes. This puzzle remains even when all the experiments are performed well (i.e. there is no malpractice in any of the research groups), demonstrating that failed replications are not always indicative of bad scientific practice (either on behalf of the original experimenters, or the replicating teams) (Stroebe and Strack [2014]).

There are deep philosophical issues to be explored concerning how we characterise replications and how they impact discussions about objectivity in science and the role of experiments in the scientific method (Sprenger [2018]; Romero [forthcoming]), tactics for breaking the experimenters' regress, and non-scientific influences on hypotheses (Feest [2016]). This paper develops one aspect of this debate, namely, how experimenters' prior theoretical commitments affect their evaluation of a replication as 'failed' or 'successful', through the lens of a contemporary case study: the ongoing debate about infants' understanding of false beliefs. Section two describes the case study in detail, with

sections three and four highlighting points where researchers' theoretical commitments about the both methodology and the concepts it tracks, affect their evaluation of replication work. The second part of the paper shows how the issues covered in sections three and four are specific instantiations of more general explanations that have been given for psychology's replication crisis. Section six examines an argument given by Michael Muthukrishna and Joe Henrich ([2019]) (among others) that new areas of science lack over-arching theories to inform their hypotheses and guide intuitions, and shows how this applies to the case study; and section seven applies these observations to a recent contribution by Alexander Bird ([2018]), wherein he argues that we should expect a high rate of false positives in the psychological sciences due to a low prior probability that the hypotheses under consideration are true. While I believe Bird's diagnosis is correct, I disagree with his analysis of how replication can help improve the situation.

Although bad scientific practices, as well as sociological factors such as the 'file-drawer' effect and the incentives associated with publishing positive findings, are partially responsible for the current replication crisis, these factors are by no means exhaustive and will not be discussed further in the paper. Unfortunately, failing to replicate a finding is often perceived as a black-mark against the original lab, threatening to spiral to suspicion, animosity and even funding cuts. Furthermore, this solely negative view of failed replications threatens public trust in science: how can we trust what one research team says when another will show them to be wrong (Ferguson [2015])? I do not believe that any of the results discussed in the paper stem from bad science. Thus, a further aim of this paper is to serve as a call to arms for philosophers of science: there are many philosophically interesting and illuminating reasons for why experiments do not replicate which have nothing to do with bad scientific practice. There is so much for

philosophers of science to do in developing and evaluating these reasons, which in turn has the potential to better our understanding of the phenomena in question as well as reinforcing that contradictory results are part of the scientific journey, and not just bad data.

2 Infants' Understanding of False Beliefs

This section gives a brief history of the false belief tasks, before explaining in detail one particular infant false belief (IFB) task and attempts to replicate it.

2.1 Elicited response false-belief tasks

In 1983 Hans Wimmer and Josef Perner set out to see whether children could attribute false-beliefs to other people. They devised the 'Maxi' task, the format of which has been the foundation for hundreds of subsequent tasks. Children watch a puppet, Maxi, hide some chocolate in one of two cupboards. Maxi leaves the chocolate in cupboard X and goes out to play. In his absence, his mother enters and moves the chocolate from cupboard X to cupboard Y. She leaves and Maxi returns, then the child is asked where Maxi will look for his chocolate. Three-year olds overwhelmingly respond that Maxi will look in cupboard Y, that is, where the chocolate really is and not where Maxi believes the chocolate to be. Around four-years of age children correctly answer that he will look in cupboard X. The authors explained their result with the hypothesis that three-year old children are limited in their ability to attribute psychological states to other people and are unable to attribute false beliefs to others, whereas four-year old children have developed this ability. This task, and those like it, are referred to as 'Elicited response'

tasks as they require the child to respond to a question asked by the experimenter: 'Where will Maxi look for his chocolate'.

In a meta-analysis of 178 elicited response false-belief tasks, Wellman *et al* ([2001]) report that 'the "medium" in which the false-belief task is presented has no significant effect. That is, it makes no difference if the protagonist is presented as a real person, a puppet, a doll, a pictured storybook character, or a videotaped person' (p.664). And while there is some variation in the age at which children pass,[1] overall the trend is robust, and has even been referred to as 'developmental dogma' (Low et al. [2016]): four-year olds succeed, whereas three-year olds fail.

2.2 Spontaneous response false-belief tasks

In 2005 Kristine Onishi and Renée Baillargeon published their paper 'Do 15-month-old infants understand false beliefs?' wherein they argued that 'Yes, they do' . They gained their results through a completely different paradigm to the Maxi task, using measurements of infants' spontaneous looking times rather than verbal, elicited responses.

Onishi and Baillargeon's experiment used the well-established 'Violation of Expectation' (VoE) paradigm. This paradigm exploits the fact that infants look longer

[1]For example, children with siblings pass the test earlier than only children (Perner *et al* [1994]); children from families with low SES (socioeconomic status) are slower to develop mastery of false belief tasks (Holmes *et al* [1996]); children with low social status within a group struggle with some false belief tasks (Rizzo & Killen [2018]); and the frequency of mental state talk within the family affects false belief performance (Brown *et al* [1996]).

at events that surprise them. Surprising events are those which the infant did not expect to happen, and therefore a longer looking time at event B in contrast to event A indicates that the infant expected A but not B to occur. The theory is that results build up a picture of what infants expect from the world, which in turn gives insight into their knowledge about how the world works. Originally used to test infants' understanding of physical knowledge (Baillargeon [2004]), the paradigm has been extended to test infants' understanding of psychological concepts like goals, agency and, in this experiment, belief. The VoE paradigm is referred to as a 'spontaneous response' method, as it measures an involuntary response from the infant. This contrasts with 'elicited response' tasks which require the child to give a voluntary response in the form of pointing or speaking.

The first stage of the experiment is to familiarise the infant to a series of events until she loses interest. In the first familiarisation event the infant watched an experimenter play with a watermelon toy, before placing it in one of two boxes in front of her (green or yellow). In the second and third familiarisation trials the experimenter approached the apparatus and reached into the box where she left the watermelon, as if to grab it.

The second stage of the experiment is the 'belief induction'. Here the infants watch one of four conditions. In the critical false belief conditions, the infant saw the watermelon move, in the experimenter's absence, from the box where it had been left into the other box. In order to counteract the possibility that infants just had a preference for boxes of a particular colour, there were two distinct false belief conditions. In one the toy moved from the yellow box, where the experimenter had left it in the familiarisation trials, to the green box. In the other, the experimenter was initially present, and watched the toy move from the yellow box, where she had left it, to the green box. The experimenter was then occluded by a screen, at which point the toy

moved from the green box back into the yellow box. In the two true belief conditions the experimenter was present throughout and watched the movement of the toy. The third stage of the experiment is the test trial. Here, each infant was shown one of two events: the experimenter reaching into a green box or the experimenter reaching into the yellow box, and the looking time of the infant during this event was recorded.

The critical finding was that infants would look longer at those test trials where the experimenter did not act in accordance with her belief about the watermelon's location, regardless of whether that belief was true or false. For example, in the condition where the experimenter leaves her toy in the yellow box, and the toy moves in the experimenter's absence to the green box, during the test trial infants who saw the experimenter reach into the green box (where the toy actually is) looked considerably longer at this event than those infants who watched the experimenter reach into the yellow box. This indicates that those infants who saw the experimenter act on her false belief about the toy's location were not at all surprised by this outcome, in contrast to those infants who saw the experimenter act in a way that did not match with her belief. This led Onishi and Baillargeon to conclude that

> 'Whether the actor believed the toy to be hidden in the green or the yellow box and whether this belief was in fact true or false, the infants expected the actor to search on the basis of her belief about the toy's location. These results suggest that 15-month-old infants already possess (at least in a rudimentary and implicit form) a representational theory of mind: They realize that others act on the basis of their beliefs and that these beliefs are representations that may or may not mirror reality.' ([2005], p.257)

Another frequently used spontaneous method is that of 'Anticipatory Looking' (Surian et al. [2007]; Southgate et al. [2007]; Luo and Baillargeon [2007]). In this paradigm, researchers measure where an infant looks at a scene, before the protagonist comes back to retrieve her object. If the infant looks to where the protagonist should search for her object in the light of her beliefs, then this is taken to support the hypothesis that the infant is correctly attributing beliefs to the protagonist. Since Onishi and Baillargeon's 2005 publication there have been more than thirty published papers reporting success in spontaneous response false belief tasks in infants aged two years-old and younger (see (Scott and Baillargeon [2017]) for a review). While VoE and Anticipatory Looking are the most used paradigms, researchers have also claimed success using other spontaneous methods (e.g. EEG measurements (Southgate and Vernetti [2014]))[2]. This forms a significant body of work used to support the view that infants can attribute false beliefs to other people.

2.3 The developmental gap

These two sets of data, from elicited and spontaneous false belief tasks, have given rise to what is known as the 'Developmental Gap'. Why is it that infants' looking times appear to support the hypothesis that they can attribute false beliefs to others, while three-year olds' performances on elicited task consistently fail to show any such ability? Since the publication of Onishi and Baillargeon's work in 2005 this question has dominated developmental psychology and attracted a great deal of attention from philosophers of cognitive science. But recent attempts to replicate the results of

[2]EEG: Electroencephalogram

spontaneous response tasks, in particular VoE and Anticipatory Looking tasks, have met with very mixed success, causing some to question the robustness of the data and, in turn, the support they offer for the hypothesis that infants can attribute false beliefs to others (Kulke et al. [2018]).

2.4 Replication

This paper will focus on replication attempts of Onishi and Baillargeon's infant false belief task for two reasons: first, in a recent meta-analysis of 33 spontaneous response tasks, Pamela Barone and colleagues ([2019], p.13) reported that correct performance on the VoE task was 'about 3.58 times more likely than incorrect performance', whereas no significant effect size was found across Anticipatory Looking tasks; second, as an analysis of both VoE and AL paradigms is beyond the scope of this paper, focusing on what the scientific community takes to be the more reliable paradigm seems a reasonable starting point (Barone et al. [2019]; Poulin-Dubois et al. [2018]).

Kulke and Rakoczy ([2018]) report nine replication attempts of Onishi and Baillargeon's VoE false-belief task. Each replication attempt deviated from the original in small ways, e.g. using different stimuli or not including true-belief as well as false-belief trials. However, the 'primary information focus' (Schmidt [2009], p. 94) remained the same through the experiments: babies watch an actor hide an object in one location; in the actor's absence the object is moved to a different location; when the actor returns she either looks in the location where she left her toy (belief-congruent condition: behaviour matches belief), or she looks to where the toy is now (belief-incongruent condition: behaviour does not match belief). The critical measure is

how long infants look at each condition.

Of the nine replication attempts, four are reported as successful, and three of these were conducted by members of Baillargeon's lab with Baillargeon as a co-author (a point that will be discussed in section four). Kulke and Rakoczy do not report the results from the non-replication studies (beyond saying that they did not replicate the original findings). However, two published papers shed light on this. Lindsey Powell and colleagues ([2018]) reported that there was no interaction between infants' (17 months) looking times and the trials, concluding that 'we failed to find evidence that 1.5-year-old infants expected the experimenter to reach for an object either where she falsely believed it to be, or in its true location' ([Powell et al., 2018, p. 5]). Sebastian Dörrenberg and colleagues ([2018]) had a partial replication with 24 month old infants: while infants looked significantly longer at belief-incongruent trials in contrast to belief-congruent trials, they only did so in the false-belief condition. That is, there was no difference in looking time between conditions where an actor behaved in a way that matched her true-belief, versus those where she acted in a way that did not match her true-belief. This is a problematic result as it suggests that infants can correctly recognise when someone's behaviour does not match her false-belief, but not when that behaviour does not match her true belief! Dörrenberg suggests that this data supports a non-psychological explanation of the results, indicating that infants simply look longer at trials when someone reaches for a full, rather than an empty box ([Poulin-Dubois et al., 2018, p. 309]). Finally, Yott and Poulin-Dubois ([2012]) found 18 month old infants looked longer when an actor's search behaviour did not accord with her beliefs. However, as they did not run a true-belief condition these data can be considered only a partial replication of Onishi and Baillargeon's original findings, for reasons just rehearsed in

Dörrenberg's case.

3 Theory and the 'Phenomena in Question'

Allan Franklin observed that a key factor in evaluating an experiment is

> '[...] the question of whether there is sufficient evidence showing that the apparatus is actually measuring the quantity or phenomenon of interest, and not some experimental artefact.' ([Franklin, 1981, 370])

Harry Collins responded to Franklin by maintaining that his observation does not go far enough, for in order to gain community consensus about the proper functioning of the equipment and method '[t]here must be sufficient agreement over the nature of the phenomenon in question to allow for there to be agreement over what is to count as a well-performed experiment' ([Collins, 1984, 173]). This section illustrates how different researchers' theoretical commitments affect the construal of the phenomena under consideration, both of what is tracked by the VoE and how it is tracked.

3.1 Phenomenon one: belief attribution

Since Onishi and Baillargeon's seminal paper researchers have argued about what kind of conceptual competence is tracked by spontaneous methodologies. Three types of theory dominate, and while all three accept that infants can have beliefs, they disagree about what sorts of things infants can have beliefs about. The 'Conceptually rich' account maintains that infants have the 'capacity to attribute propositional representations of various basic types, including belief' to other people ([Carruthers, 2013, p.143]). In other

words, infants can have beliefs about other people's beliefs, i.e. the infant believes that Sally believes that the ball is in the basket. Onishi and Baillargeon's interpretation of their data is a conceptually rich one (see section 2.2).

A second, psychologically-sparse group of views, maintains that infants can attribute psychological states to others, but that the states they attribute are less complex than 'propositional representations' (Apperly [2010]; Butterfill and Apperly [2013]; Carey [2009]; Low et al. [2016]; Wellman [2014]). These positions claim that infants track relations that stand between an agent and an object, and that infants do not need to represent the other's psychological state in the form of a propositional attitude. Where a conceptually rich account says 'The baby believes that Sally believes that the ball is in the basket', psychologically-sparse account says 'The baby believes 'that Sally stands in the 'registering' relation to that object' (where 'registration' is a relational state). Because psychologically-sparse accounts maintain that infants cannot attribute propositional attitudes to others, they predict that infants should fail false-belief tasks that require representing how an object appears to another person (Low et al. [2016]).

Finally there are the associationist, non-psychological accounts. These maintain that infants are behaviourists and do not attribute anything psychological at all to the protagonist (Heyes [2014*a*,*b*]; Santiesteban et al. [2014]). Instead, infants use behavioural-rules to predict where the protagonist will look, e.g. the infant believes that 'people return to where they last left an object'.

If the 'phenomenon in question' is 'infants ability to attribute false beliefs to others' then these hypotheses quite clearly disagree about whether Onishi and Baillargeon's experiment adequately measures it. It could, so the argument goes, equally well be measuring an infants' ability to track another's registration, or her ability to apply rules

about behaviours. Importantly, a failure to replicate Onishi and Baillargeon's results would affect all three hypotheses equally, as they each take for granted that the experiment demonstrates an ability to discriminate between belief-congruent and incongruent behaviours. Should this discrimination fail to replicate and subsequent doubts are raised about its existence, then any theory whose aim is to explain it runs the risk of redundancy.

3.2 Phenomenon two: expectation

In the VoE paradigm there are the methodological assumptions that if infants look longer at event A, which is unexpected, in contrast to event B which is expected, then this indicates that they...

> '...[1] Possess the expectation under investigation, [2] detect the violation in the unexpected event, and [3] are 'surprised' by this violation.' ([Baillargeon, 2004, 392])

These assumptions have not gone unchallenged. Bruce Hood and colleagues ([2004]) have argued against [1] on the grounds that one can detect that something has gone awry without necessarily knowing what that is. For example, Kaiser and colleagues ([1985]) demonstrated that when college students were asked to draw the path of a falling object (e.g. the trajectory of a ball emerging from a spiral tube) their drawing was usually inaccurate. But when the same participants watched a movie simulation of the event which mimicked the path they had drawn, they immediately recognised it as somehow wrong. The participants detected that something was amiss with what they had seen, but their drawings would suggest they did not possess an expectation of how

the ball would fall.

In response, Baillargeon has argued that achieving [2] and [3] require that the participant has an expectation 'in some manner and at some level' of what should occur, therefore if there is agreement that these conditions have occurred then [1] stands by necessity ([Baillargeon, 2004, 392]). Her response implies that this expectation is a specific one, but one could offer a weaker reading of 'expectation', arguing that it is fuzzy: sufficient to cause recognition of something as being wrong, without being specific enough to pinpoint what, exactly, that may be. The bigger point here is that there is disagreement about what an 'expectation' is, amounting to disagreement about the 'phenomenon in question' tracked by the method: is it a specific expectation or a fuzzy one? If it is a fuzzy expectation, then we may expect more variance in the data, which in turn affects the magnitude of the effect size, itself a key factor in how we evaluate replication attempts as successful or unsuccessful.[3]

3.3 Phenomenon three: looking time

The VoE methodology requires a measurement of how long an infant looks at a particular event. Typically, infants need to look at an event for at least two seconds to be included in a data set. The methods for measuring infants' looking times vary across laboratories.

[3]There is, in fact, huge variance in looking times across infants. A recent meta-analysis by Barone and colleagues reported that 'Looking times considerably varied from one experimental condition to another: they stretched from 4.52 to 29.5 s when unexpected outcomes are presented and from 3.27 to 18.8 s in the case of expected events' ([Barone et al., 2019, 17]).

In some cases, remote corneal-reflection eye trackers are used: cameras which track the reflection from a participant's pupils to determine what they are looking at. These trackers give automatic measurements of the participant's looking time, and some can implement 'participant controlled trials', meaning that the trial ends automatically when the baby looks away from the stimulus for more than two seconds. For the VoE false belief studies discussed here, though, the more common method to measure eye-gaze is through on- or off-line coding by human observers. Onishi and Baillargeon used on-line coding, where two hidden observers watch the infant, and judge when she begins and finishes looking at the event, with the trial ending when the primary observer determined that the baby had looked away for more than two consecutive seconds, or had looked at the scene for 30 cumulative seconds (Onishi and Baillargeon 2005, supplmentary material). Powell and colleagues ([2018]) used 'offline coding', where two coders watch video footage of the trial and measure how long the infant looks at each event. This coding is 'trial blind', which means that coders do not know which condition the baby is in; they simply record how long the baby looks at that event.[4] It is standard practice to report inter-observer reliability, and in published papers this is usually above 95%.

On- and off-line human coding are based on the same theoretical principle: human observers are accurate enough in ascertaining when an infant starts and ceases to look at a particular event. Here, at least, there is some consensus that the apparatus (human coders) for measuring eye-gaze succeeds in tracking the phenomenon of interest (an infant's looking time). But it is noteworthy that experimental reports rarely, if ever, justify why they use one method of measurement rather than another, and that Powell

[4]On-line coding is also assumed to be trial-blind; anecdotal evidence suggests that this very difficult to implement in practice.

and colleagues clearly felt that using a different measure of looking time did not mark a significant difference between their 'replication' of Onishi and Baillargeon's work and the original. However, the possibility remains open that one laboratory may consider another's experiment to be less 'well-performed' should they choose a different method of measuring looking time than the original experiment.

These three phenomena highlight points in any IFB experiment where the researchers' prior theoretical commitments shape how they measure the phenomenon of interest, and what they take these phenomena to be. In the case of measurement there appears to be some consensus that human observers can accurately measure looking time; in the cases of the specificity of the expectation tracked by the VoE and the type of psychological concepts required to succeed on the experimental task, the picture is more complex. I return to these issues in section 6.

4 Experimenter Skill

Experimenter skill is a recurring theme in the replication literature (Collins [1985]; Leonelli [2018]; Polanyi [1967]). In a published paper, it is impossible to list all the variables in any one experiment; the best one can do is list all the factors believed necessary for bringing about the effect. However, as Collins documents in his case study on TEA[5] lasers ([1985]), when an experiment purports to show a new phenomenon, it is almost never the case that an experimenter aiming to repeat the finding can reproduce the original results based on the published report alone. Even when the two experimenters meet to discuss their techniques, or the original experimenter visits the

[5]Transversely Excited Atmospheric pressure CO_2 laser

new lab to examine their set-up, it is not always clear what, in the knowledge exchange, was the piece of information critical to getting the replication to work. When the experiment is new, neither the original experimenter, nor the one aiming to replicate the results, may know what is missing in the unsuccessful case; only that something is missing.

Renée Baillargeon is one of the most skilled people in the world with the VoE methodology, using it since the mid-1980's in her work on infants' understanding of the physical world. She has undoubtedly embodied a huge amount of tacit knowledge relevant to her trade (Polanyi [1967]), and not all of this knowledge will be presented in a published paper.[6] These skills, in turn, give her an unique perspective on what constitutes a 'well-performed' replication of her work, a point that comes out clearly in her responses to 'failed' replications. While she has offered several reasons for why her original data failed to replicate, ranging from methodological changes to details of the analysis of the infants' looking times ([Buttelmann et al., 2018, pp. 113-5]), I focus here on the two responses that most clearly illustrate her embodied experimental skill.

First, she observes that infants need to have an expectation of how the actor will behave in order for that expectation to be violated. Her concern with Powell and colleagues' experiment is that the familiarisation trials are too complex, involving the actor putting the toy in one box, leaving, and then returning and switching the toy to another box. This occurs before the actor leaves for the second time and the belief-induction trial commences (with the toy moving from where it has been left to the

[6]For example, in presentations she explains that she uses real actors whenever possible because babies are far more interested in watching real people than videos, and so they are more likely to attend to the stimuli.

other location). Baillargeon writes

> 'One factor that might have contributed to the task's negative results is that these events were somewhat confusing (e.g., why did the agent switch the toy to the opposite box when she returned?) and made it difficult for children to form a clear expectation about what the agent sought to accomplish.' ([Buttelmann et al., 2018, 113])

Her second concern is that Powell and colleagues did not allow the infant sufficient time to form an expectation about the actor's behaviour. In the original study, infants watched the toy move from one box to another in the actor's absence (belief-induction phase), and this scene was then paused for a few seconds before they watched the test phase. In Powell's study, however, infants watched the familiarisation phase, followed immediately by the belief-induction phase and then the test phase. Baillargeon maintains that the 'children did not have sufficient time to process the novel information they had received and to form an expectation about what the agent would do before she completed her actions.' (ibid). That the pause is not explicitly mentioned in published experimental reports suggests that its inclusion in the method stems from tacit knowledge of how babies best respond in the lab: knowledge gained from many years of experience. When the pause is missing (as occurs in the Powell failed replication), its significance takes on a new importance. Here is a case where a lack of shared knowledge in the community about how a particular methodology works affects replication success, or, in Collins' words, where 'ignorance of the nature of the experimental conditions makes it impossible to guarantee results' ([1985], p.171). Replication experiments serve an essential role in making tacit knowledge explicit, by encouraging a more rigorous

investigation of the methods and how they work (Strack [2017]).

As noted above, of the nine replication attempts reported in Kulke and Rakoczy ([2018]), three of the four successful replication attempts of the VoE came from Baillargeon's laboratory. But maybe this outcome unsurprising, if this indeed is the laboratory with the most knowledge and skill in applying the VoE method in the false-belief case. One innovative way of overcoming these limitations is shown by the 'Many Babies Consortium' (Bergelson *et al* [forthcoming]), which conducts large scale replication projects across infant laboratories. They ask each lab to upload a 'video-walkthrough' of the infant's experience, from entering their laboratory through to their departure. Each laboratory has its own way of welcoming and settling the infant, which in turn are techniques the experimenters have learned through years of experience, and a video is much more likely to show these than the most comprehensive of written reports. However, it remains the case, as Collins points out ([1984], p.171) that the strength of a replication often relies on who performed it, with research communities giving more weight to successful replications from laboratories other than the original. More rigorous forms of reporting how a method is applied, e.g. video-walkthroughs, have the potential to share the knowledge embodied in a particular laboratory, de-mystifying and thus increasing confidence in, the original experiment.

5 Fragile Effects

This section ties together themes from sections 3 and 4 — how our theories shape the phenomena we are looking for, and the role of experimenter skill in eliciting responses from infants — to show how they give rise to tension between the concepts the IFB task

is intended to track, and the apparent fragility of those concepts.

Baillargeon's knowledge about having a small pause between the familiarisation trial and the test trial reveals a particular commitment that constrains and informs the design of her work: infant cognitive effects are fragile, and therefore great care must be taken in the experimental set-up to ensure that the conditions are optimal for those fragile effects to manifest. Replications which omit one or another detail may fail to reveal the effect under consideration. That infants' ability to attribute false-beliefs to others is fragile is not, to my knowledge, something that has been made explicit in discussions of the IFB experiments. Acknowledging that this could be the case leads to two important considerations: scoping the effect, and its ecological validity.

In her detailed examination of the 'Mozart effect', Uljana Feest shows how attempts to replicate the effect resulted in scoping its limits beyond the initial broad claim that 'listening to Mozart has positive effects on brain development' ([Feest, 2016, 38]). A similar interpretation can be offered of the IFB replications. It may be that instead of the broad claim that 'infants can attribute false beliefs to others' a more careful one is warranted, namely, that when the conditions are just-so, (with the specifications to be fleshed out through further replication work), babies look longer at belief-incongruent than belief-congruent behaviour. But when is an effect so fragile as to be non-existent? The answer will be that the effect is fragile if the theory can specify the conditions under which it will manifest (and why), and absent if no such scoping is forthcoming or supported by the theory. A different replication debate illustrates this well: in the literature on whether adults automatically process another's perspective, where replication attempts have been as messy as those discussed in this paper, Cathleen O'Grady and colleagues ([forthcoming]) used failed replications to develop a theory of

why this ability might not manifest robustly, and the conditions under which we might expect to see it.

Several authors (Feest [2016]; Leonelli [2018]; Strack [2017]) have argued that failed replications can play a crucial epistemic role in allowing researchers to scope the limits of the effects under study, and to better understand the conditions under which they manifest. This, in its turn, raises theoretical questions about the nature of the phenomenon in question: if the conditions under which infants can attribute false beliefs to others become very narrow, then how similar is the infant's concept to the pre-experimental theoretical concept BELIEF? Beliefs are representational states that can take any content. As such, they are indefinitely flexible, and while their flexibility will depend on the conceptual capacities of the individual,[7] it would be odd to say that an infant who could only understand false-beliefs about an object's location when the object is a toy watermelon slice and the hiding places are one green and one yellow box, has the BELIEF concept. Advocates of the hypothesis that infants can distinguish belief-congruent from belief-incongruent behaviour thus need to tread the fine line between using failed replications of their work to offer a clearer account of the conditions under which the VoE method can successfully track belief understanding, while also retaining an account of the concept of BELIEF[8] that fits with their broader theoretical commitments about cognitive architecture.

A further issue is what theories predict in terms of the ecological validity of the experimental data. If one maintains that attributing psychological states to others is

[7]For example, one couldn't expect an infant to have beliefs about global warming

[8]The same applies to hypotheses positing the concepts REGISTRATION or BEHAVIOURAL PATTERN

fundamental to social understanding, then a limited hypothesis which suggests infants only manifest this ability in carefully controlled laboratory situations is problematic. Should relatively small tweaks to a careful experimental set-up mean that infants are no longer sensitive to another's false beliefs, then there is a question about whether, in the messy real world, attributing beliefs to others is their primary way of understanding their behaviours (Klein et al. [2014]). On the other hand, if one maintains that attributing psychological states to others is not the primary way in which we interact (Bermúdez [2003, 2007]; De Jaegher et al. [2010]; Hutto [2017]; Ratcliffe [2006]), then evidence to suggest that this ability only occurs in very specialised laboratory conditions is taken to support the view that in more ecological contexts, social interactions do not depend on this ability.

6 Theory and the Replication Crisis

The previous sections showed how different researchers' understanding of the FALSE BELIEF concept and their practice in using the VoE methodology affects their evaluation of replications of the IFB experiment. This section explores how this corner of developmental psychology intersects with broader theoretical analyses of psychology's 'replication crisis', examining two explanations for the high rate of failed replications in the field: first, that the discipline lacks over-arching frameworks to guide theoretical intuitions; and second, that as a new discipline, there will be a higher number of false hypotheses under consideration.

6.1 A lack of 'over-arching' theories

In a recent paper, Michael Muthukrishna and Joe Henrich claimed that

> 'Many subfields within psychology (though not all!) lack any overarching, integrative general theoretical framework that would allow researchers to derive specific predictions from more general premises. Without a general theoretical framework, results are neither expected nor unexpected based on how they fit into the general theory and have no implications for what we expect in other domains.' ([Muthukrishna and Henrich, 2019, 221]).

Although not explicitly acknowledged by the authors, this observation has its roots deep in the philosophy of science and the Duhem-Quine thesis that hypotheses are always tested against a backdrop of theoretical commitments. When a prediction does not turn out as hypothesised, there are a number of places where it might have gone wrong. The more established the science, the less likely it is that the foundational theory is incorrect and the more likely it is that some aspect of the experiment is at fault. If you drop a packet of mints into a bottle of diet cola and no fountain of pop ensues, this is most likely to be because the cola is flat, or that the surface of the mints was too smooth to cause the reaction. As there are thousands of replications of the cola-mint fountain experiment conducted in gardens through the world (and documented online), it is less likely that your experiment has revealed a gap in the current principles of thermodynamics, surface science or eruption theory (Coffey [2008]). These 'overarching, integrative general theoretical frameworks' are well-established and give us good reason to expect particular experimental results; when the result is not forthcoming, the first place to look is at the methods and materials of the experiment itself rather than the

theoretical framework informing it.

As a relatively new science, many aspects of psychology lack the well-established frameworks that guide researchers' intuitions about what to expect. In no place is this more salient than the case of infancy research, where systematic studies with infants and toddlers have been conducted for around forty or so years, while many of the methods and equipment used to explore infants' understanding are even younger, originating in the mid- to late 1990's. This paper has shown that there is little in the way of consensus about how the VoE method should be carried out, what kinds of concept it can track, and how robust these concepts might be. Each of these disagreements reflects a lack of well-established framework informing the experiments and the concepts they are designed to track. Consequently, when a claim like '15-month old infants attribute false beliefs to others' is published, the research community does not have a strong sense of whether this is likely to be a false-positive or a real phenomenon, as it neither supports nor contradicts a foundational theory within the field. It is equally difficult to evaluate failed replications of the work, as it will not be clear whether something has gone wrong in the experimental set-up (because there is so little agreement about what the correct experimental set-up should be), or whether the replication's null result is the correct one, again because a failed replication neither supports nor contradicts a well-established existing theory.

6.2 Theoretical frameworks and the Base-Rate fallacy

Alexander Bird has argued that psychology's replication crisis can be partially explained by 'the fallacy of neglecting the base-rate' when it comes to generating successful

hypotheses (Bird [2018]):

> 'If there is both (i) a low background rate of truth among hypotheses proposed and tested and (ii) a significance level set at a value that although small is not negligible, then we would expect a high proportion of positive results of tests of those hypotheses to be erroneous. Hence we would expect many replication studies to fail to reveal what the original studies seemed to show.' ([2018], pp. 11-2)

I will set aside the point about significance, although at $p = 0.05$ developmental psychology experiments generally do have a significance value that 'although small is not negligible'. Instead, the focus is on how having relatively few over-arching background theories affects the base-rate fallacy.

Is developmental psychology a field where there is a low background rate of truth among the hypotheses proposed and tested? The ideas developed above suggest that it is. If there are few well-accepted over-arching theories to inform the hypotheses being tested then (a) the theories which are informing the hypotheses might not be correct; and (b) even if correct, they may not yet be sufficiently worked out to yield hypotheses with a high degree of specificity (e.g how much longer infants should look at incongruent events, section 3.2). In Bayesian terms, when there is uncertainty about the general theories informing specific hypotheses, we are not in a position to give those hypotheses a high prior probability. Furthermore, it may not be clear which hypothesis should be awarded a higher prior than another, due to the lack of consensus about the background theory informing each of them. The relative newness of psychology as a discipline means that researchers may 'lack the intuition' ([Muthukrishna and Henrich, 2019, p.221]) that

tells them when a hypothesis is likely to be false, e.g. it is not clear whether or not babies ought to be capable of attributing false-beliefs to other people.

In practice, developmental psychologists do not seem to lack such intuitions, conducting experiments only when they are confident of the result.[9] Babies are hard work: they require specialised laboratory facilities and equipment; experiments are time consuming with each participant requiring a settling-in play session before the experiment begins and a post-experiment debrief for caregivers, plus inevitable delays due to naps, distractions, nappy changes and snacks; as well as routine expenses of compensating caregivers for their time and travel, and a customary souvenir toy for the baby. One thus needs a reasonably high level of confidence in one's hypothesis before taking on these costs. The bigger question is whether this confidence is justified. Experimenters are confident in their hypotheses because they are confident about the background theories informing them, e.g. if one is committed to the claim that babies are capable of using complex, abstract concepts, then one is likely to be highly confident that they are capable of attributing false beliefs to others. Bird's claim is a descriptive, not a normative one: while the experimenters concerned may have high confidence in their hypotheses, the nature of the field is such that this confidence may be ill-founded. The remedy, in line with Muthukrishna and Henrich's work, is to step back and comb through these theories, making explicit the reasons for (a) why we should expect babies to be capable of using abstract concepts, and (b) more specifically, why we should expect them to be capable of tracking false beliefs. Exploring these claims requires making clear commitments about the cognitive architecture required for achieving these goals, and

[9]While I have no hard evidence for this claim, many conversations with developmental psychologists suggest that it is the case.

analysis of the onto- and phylogenetic developments such a cognitive architecture demands. An exemplar of this approach is Susan Carey ([2009]), whose theory sets out constraints on infants' cognitive capacities that are informed by research in both developmental psychology and cognitive ethology. Working within these constraints, her account predicts that infants cannot attribute false-beliefs to others, and offer an explanation of the data in term of tracking the agent's goals ([Carey, 2009, 209]) (making her position one of the 'psychologically sparse' type, see section 3.1).

A counter to this argument is that we need to know what babies can achieve in order to ensure our theories can accommodate these data. This tactic is in clear evidence in many philosophical papers discussing mindreading and social cognition (Carruthers [2011, 2013]; Lavelle [2019]; Michael and Christensen [2016]), where a list of experimental data is held as the touchstone against which to evaluate various theories. But the arguments presented in this section and throughout the paper suggest that we are simply not in a position to know which infant data are robust (in contrast to the 'developmental dogma' that is passing the Maxi task), precisely because there is, as yet, so much vagueness in the theories the data are intended to support (or refute). John Ioannidis argued that 'The greater the flexibility in designs, definitions, outcomes and analytic model in a scientific field, the less likely the research findings are to be true' ([2005], p.44). . Section three showed just some of the controversies over design and definitions that feature in the infant cognition literature, suggesting that a piecemeal analysis of each of the moving parts (e.g. what can be tracked using the VoE methodology? how much variation across participants do we expect and why?) is the best chance of getting the background theories into better shape, so as to allow them to make more specific predictions to be tested.

While this paper agrees with Bird and others (Fanelli [2009]; Ioannidis [2005]) who maintain that the base-rate fallacy accounts for why so many studies in psychology fail to replicate, it is wary of the diagnosis that more replication will always solve the problem: 'in due course, certain results will stand out as reliable (thanks to successful replication)' (Bird [2018], p.25; see also Moonsinghe *et al* [2007]) . The fallacy of the base-rate applies to the number of hypotheses researchers have to choose from; replication can only tell us if particular data points are robust and not whether these data points support one or another hypothesis. The result of the Maxi test is a developmental dogma, but there remains controversy about how to interpret it. While this reliable result serves to narrow the field of hypotheses by some degree (e.g. by ruling out all those that predict four-year olds should fail the test), it still 'rules in' a huge number. The problem becomes more pressing when applied to infant cognition, where there remains controversy about what even constitutes a replication of an experiment (section four). Imagine that agreement was reached about an experimental procedure replicating the original, and multiple repetitions of the experiment yielded data matching the original data. What we would end up with is that 'infants under these very specific conditions (e.g. with live actors and the pauses in the correct places between conditions) look longer when an actor acts in a way that does not match with her beliefs'. This has become a robust finding, but there are any number of reasons for why the data come out in this way.

Repeating a specific finding does not in and of itself reduce the space of false hypotheses; what is required is a worked through theory of why the specific finding comes out as it does, and which predicts other findings of equal specificity (Franklin and Howson [1984]). This is not to say that there is no value in repeating experiments: repeating experiments allows us to understand where variation might lie and how robust

or fragile we expect the result to be, both of which point to where the theory needs to be more specific and supported. And getting a stable data point, albeit a very specific one, is better than no data point at all. The moral is simply that successful replications alone cannot resolve the theory-crisis that underlies the current replication crisis. It is only once the theories are sufficiently worked through to make predictions about which effects should be robust, under which circumstances and why, that progress can be made on this front.

7 Conclusion

It is well-known through philosophy of science that theory infuses every aspect of experiment, from how instruments are calibrated to commitments about the phenomena those apparatus purport to track. Yet, with a few exceptions (Collins [1989]; Fiedler [2017]; Strack [2017]), this point has been under-appreciated in the recent debates about the replication 'crisis'. This paper has aimed to show some of the points at which theory intersects with experiment in the literature on infants' understanding of false belief, and in so doing revealed the complexity that afflicts our evaluations of replications as 'failed' or 'successful'. The current run of experiments serve a critical epistemic role in scoping the conditions under which infants look longer at belief-incongruent versus congruent events, which in turn yields a more nuanced understanding of the effect to inform the theories guiding the experimental work. However, we do not yet know (a) whether VoE experiments reveal that infants can attribute false beliefs to others, or (b) whether VoE studies of IFB replicate. Until some community consensus is reached on what a belief is, how robust or fragile that concept is, whether the VoE can track it, and how specific the

infant's 'expectations' are, these questions remain open.

There are many more facets of the replication debate to be explored, beyond the interaction of theory (or lack thereof) and experiment. For instance, this paper has not touched on how to determine whether an experiment is sufficiently similar to the original to count as a replication, or how many changes to the experiment can be made before it no longer counts as a replication. These questions fall under the debates concerning 'direct' and 'conceptual' replication. There is a positive agenda to be drawn from the current replication debates that isn't captured by the term 'crisis' that has become synonymous with them: it is time to rethink the replication 'opportunity'.

8 Acknowledgments

I am grateful to Jo Wolff, Alasdair Richmond and Thom Scott-Phillips for their comments on drafts of this paper, and to the hugely constructive comments of three reviewers for this journal. Thanks to Hugh Rabagliati and Josef Perner for their insights into the world of developmental psychology, and to audiences of the British Society for Philosophy of Science meeting in Durham, 2019, and the PPIG in Edinburgh, 2018, where nascent versions of these ideas were presented.

References

Apperly, I. [2010], *Mindreaders: the cognitive basis of theory of mind*, Psychology Press.

Baillargeon, R. [2004], 'Infants' reasoning about hidden objects: evidence for event-general and event-specific expectations', *Developmental science* **7**(4), 391–414.

Barone, P., Corradi, G. and Gomila, A. [2019], 'Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis', *Infant Behavior and Development* **57**, 101350.

Bergelson, E., Bergmann, C., Byers-Heinlein, K., Cristia, A., Cusack, R., Dyck, K., Frank, M. C., Gervain, J., Gonzalez, N., Hamlin, K. et al. [forthcoming], 'Quantifying sources of variability in infancy research using the infant-directed speech preference', *Advances in Methods and Practices in Psychological Science* .

Bermúdez, J. L. [2003], 'The domain of folk psychology', *Royal Institute of Philosophy Supplements* **53**, 25–48.

Bermúdez, J. L. [2007], *Thinking without words*, Oxford University Press.

Bird, A. [2018], 'Understanding the replication crisis as a base rate fallacy', *The British Journal for the Philosophy of Science* .

Brown, J. R., Donelan-McCall, N. and Dunn, J. [1996], 'Why talk about mental states? the significance of children's conversations with friends, siblings, and mothers', *Child Development* **67**(3), 836–849.

Buttelmann, D., Baillargeon, R. and Southgate, V. [2018], 'Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations', *Cognitive Development* .

Butterfill, S. A. and Apperly, I. A. [2013], 'How to construct a minimal theory of mind', *Mind & Language* **28**(5), 606–637.

Carey, S. [2009], *The origin of concepts*, Oxford University Press.

Carruthers, P. [2011], *The opacity of mind: An integrative theory of self-knowledge*, OUP.

Carruthers, P. [2013], 'Mindreading in infancy', *Mind & Language* **28**(2), 141–172.

Coffey, T. S. [2008], 'Diet coke and mentos: What is really behind this physical reaction?', *American Journal of Physics* **76**(6), 551–557.

Collaboration, O. S. et al. [2015], 'Estimating the reproducibility of psychological science', *Science* **349**(6251), 1–8.

Collins, H. M. [1984], '*When* do scientists prefer to vary their experiments?', *Studies in History and Philosophy of Science Part A* **15**(2), 169–174.

Collins, H. M. [1985], *Changing order*, Sage.

Collins, H. M. [1989], *The meaning of experiment: replication and reasonableness*, Weidenfeld, pp. 82–92.

Cooper, J. [2013], 'On fraud, deceit and ethics', *Journal of Experimental Social Psychology* **49**(2), 314.

De Jaegher, H., Di Paolo, E. and Gallagher, S. [2010], 'Can social interaction constitute social cognition?', *Trends in cognitive sciences* **14**(10), 441–447.

Dörrenberg, 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* .

Fanelli, D. [2009], 'How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data (how many falsify research?)', *PLoS ONE* **4**(5), e5738.

Fang, F. C., Steen, R. G. and Casadevall, A. [2012], 'Misconduct accounts for the majority of retracted scientific publications', *Proceedings of the National Academy of Sciences* **109**(42), 17028–17033.

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.

Ferguson, C. J. [2015], 'Everybody knows psychology is not a real science: Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public', *The American psychologist* **70**(6), 527.

Fiedler, K. [2017], 'What constitutes strong psychological science? the (neglected) role of diagnosticity and a priori theorizing', *Perspectives on Psychological Science* **12**(1), 46–61.

Franklin, A. D. [1981], 'What makes a'good'experiment?', *The British Journal for the Philosophy of Science* **32**(4), 367–374.

Franklin, A. and Howson, C. [1984], 'Why do scientists prefer to vary their experiments?', *Studies in History and Philosophy of Science Part A* **15**(1), 51–62.

Heyes, C. [2014*a*], 'False belief in infancy: a fresh look', *Developmental science* **17**(5), 647–659.

Heyes, C. [2014*b*], 'Submentalizing: I am not really reading your mind', *Perspectives on Psychological Science* **9**(2), 131–143.

Holmes, H. A., Black, C. and Miller, S. A. [1996], 'A cross-task comparison of false belief understanding in a head start population', *Journal of Experimental Child Psychology* **63**(2), 263–285.

Hood, B. M. [2004], 'Is looking good enough or does it beggar belief?', *Developmental Science* **7**(4), 415–417.

Hutto, D. D. [2017], 'Basic social cognition without mindreading: minding minds without attributing contents', *Synthese* **194**(3), 827–846.

Ioannidis, J. P. [2005], 'Why most published research findings are false', *PLoS medicine* **2**(8), e124.

Kaiser, M. K., Proffitt, D. R. and Anderson, K. [1985], 'Judgments of natural and anomalous trajectories in the presence and absence of motion.', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **11**(4), 795.

Kerr, N. L. [1998], 'Harking: Hypothesizing after the results are known', *Personality and Social Psychology Review* **2**(3), 196–217.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C. et al. [2014], 'Investigating variation in replicability', *Social psychology* .

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**(6), 888–900.

Lavelle, J. S. [2019], *The Social Mind: A Philosophical Introduction*, Routledge.

Leonelli, S. [2018], *Rethinking Reproducibility as a Criterion for Research Quality*, Emerald Publishing Limited, pp. 3–10.

Lilienfeld, S. O. and Waldman, I. D., eds [2017], *Psychological science under scrutiny: Recent challenges and proposed solutions*, John Wiley & Sons.

Low, J., Apperly, I. A., Butterfill, S. A. and Rakoczy, H. [2016], 'Cognitive architecture of belief reasoning in children and adults: A primer on the two-systems account', *Child Development Perspectives* **10**(3), 184–189.

Luo, Y. and Baillargeon, R. [2007], 'Do 12.5-month-old infants consider what objects others can see when interpreting their actions?', *Cognition* **105**(3), 489–512.

McNutt, M. [2014], 'Reproducibility', *Science* **343**, 229.

Michael, J. and Christensen, W. [2016], 'Flexible goal attribution in early mindreading', *Psychological Review* **123**(2), 219–227.

Miller, G. [2010], 'Investigation leaves field in the dark about a colleague's work', *Science* **329**(5994), 890–891.

Moonesinghe, R., Khoury, M. J. and Janssens, A. C. J. [2007], 'Most published research findings are false—but a little replication goes a long way', *PLoS medicine* **4**(2), e28.

Muthukrishna, M. and Henrich, J. [2019], 'A problem in theory', *Nature Human Behaviour* p. 1.

O'Grady, C., Scott-Phillips, T., Lavelle, J. S. and Smith, K. [forthcoming], 'Perspective-taking is spontaneous but not automatic', *Quarterly Journal of Experimental Psychology* .

Onishi, K. H. and Baillargeon, R. [2005], 'Do 15-month-old infants understand false beliefs?', *Science* **308**, 255–258.

Pashler, H. and 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.

Perner, J., Ruffman, T. and Leekam, S. R. [1994], 'Theory of mind is contagious: You catch it from your sibs', *Child Development* **65**(4), 1228–1238.

Peterson, D. [2016], 'The baby factory: Difficult research objects, disciplinary standards, and the production of statistical significance', *Socius* **2**, 1–10.

Polanyi, M. [1967], 'The tacit dimension.', *Garden City, NY* .

Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., Krist, H., Kulke, L., Liszkowski, U., Low, J. et al. [2018], 'Do infants understand false

beliefs? we don't know yet–a commentary on Baillargeon, Buttelmann and Southgate's commentary', *Cognitive Development* **48**, 302–315.

Powell, L. J., Hobbs, K., Bardis, A., Carey, S. and Saxe, R. [2018], 'Replications of implicit theory of mind tasks with varying representational demands', *Cognitive Development* **46**, 40–50.

Ratcliffe, M. [2006], *Rethinking commonsense psychology: A critique of folk psychology, theory of mind and simulation*, Springer.

Rizzo, M. T. and Killen, M. [2018], 'How social status influences our understanding of others' mental states', *Journal of experimental child psychology* **169**, 30–41.

Romero, F. [forthcoming], 'Philosophy of science and the replicability crisis', *Philosophy Compass* .

Santiesteban, I., Catmur, C., Hopkins, S. C., Bird, G. and Heyes, C. [2014], 'Avatars and arrows: Implicit mentalizing or domain-general processing?', *Journal of Experimental Psychology: Human Perception and Performance* **40**(3), 929.

Scott, R. M. and Baillargeon, R. [2017], 'Early false-belief understanding', *Trends in Cognitive Sciences* **21**(4), 237–249.

Southgate, V., Senju, A. and Csibra, G. [2007], 'Action anticipation through attribution of false belief by 2-year-olds', *Psychological Science* **18**(7), 587–592.

Southgate, V. and Vernetti, A. [2014], 'Belief-based action prediction in preverbal infants', *Cognition* **130**(1), 1–10.

Sprenger, J. [2018], 'The objectivity of subjective bayesianism', *European Journal for Philosophy of Science* **8**(3), 539–558.

Strack, F. [2017], 'From data to truth in psychological science. a personal perspective.', *Frontiers in Psychology* **8**, 702.
URL: *https://www.frontiersin.org/article/10.3389/fpsyg.2017.00702*

Stroebe, W. and Strack, F. [2014], 'The alleged crisis and the illusion of exact replication', *Perspectives on Psychological Science* **9**(1), 59–71.

Surian, L., Caldi, S. and Sperber, D. [2007], 'Attribution of beliefs by 13-month-old infants', *Psychological Science* **18**(7), 580–586.

Wellman, H. M. [2014], *Making minds: How theory of mind develops*, Oxford University Press.

Wellman, H. M., Cross, D. and Watson, J. [2001], 'Meta-analysis of theory-of-mind development: The truth about false belief', *Child Development* **72**(3), 655–684.

Wimmer, H. and 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.

Yott, J. and Poulin-Dubois, D. [2012], 'Breaking the rules: Do infants have a true understanding of false belief?', *British Journal of Developmental Psychology* **30**(1), 156–171.