Growth From Uncertainty

Understanding the Replication 'Crisis' in Infant Cognition

Jane Suilin Lavelle

School of Philosophy, Psychology and Language Sciences

University of Edinburgh

j.s.lavelle@ed.ac.uk

Abstract

Psychology is a discipline that has a high number of failed replications, which has been characterised as a 'crisis' on the assumption that failed replications are indicative of untrustworthy research. This paper uses Chang's concept of *epistemic iteration* to show how a research programme can advance epistemic goals despite many failed replications. It illustrates this through analysing an on-going large-scale replication attempt of Southgate's 2007 work exploring infants' understanding of false beliefs. It concludes that epistemic iteration offers a way of understanding the value of replications — both failed and successful — that contradicts the narrative centred around distrust.

Keywords: replication crisis; epistemic iteration; false belief; reproducibility; scientific progress

1 The crisis

“Don’t trust everything you read in the psychology literature. In fact, two thirds of it should probably be distrusted.” Thus opens a report in the journal ‘Nature’, commenting on the findings of the Open Science Foundation project which conducted replication attempts of 100 psychology experiments and reported that only “39% of effects were subjectively rated to have replicated the original result” (Baker, 2015, 1; Nosek et al. 2015, 943). Such claims lie at the foundation of a crisis in confidence in the field, whereby the failure of findings to replicate is often taken to imply (tacitly or otherwise) that they are false. The characterisation of this mass failure of reproducibility of psychological findings as a ‘crisis’ rests on the assumption that “Replication is one of the most important tools for the verification of facts within the empirical sciences” (Schmidt 2009, p. 90). Under such a characterisation, those findings which can be repeated by different researchers in different laboratories can be considered verified facts, and those which cannot are dismissed as coincidental, or the result of bad scientific practice (Loscalzo 2012; McNutt 2014; Nosek et al. 2022; Simons 2014). Subsequently, those fields which have a higher rate of failed replications are considered less trustworthy than those which have lower rates. Thus the high rate of replication failure in psychology constitutes, on this diagnosis, a crisis, in that the work produced by its researchers is considered to be unreliable.

The assumption that the successful replication of experiments distinguishes ‘trusted’ from ‘untrusted’ science has not gone unchallenged by philosophers, many of whom have argued that a high rate of replication failure can be perfectly compatible with responsibly conducted, high-quality science (Bird 2021; Feest 2019; Fletcher 2021; Irvine

2021; Lavelle 2022; Leonelli 2018; Schickore 2011). This paper offers a new addition to this counter-offensive. It argues that when researchers are working in fields which we don't yet know very much about, failed replications are not only to be expected but are necessary to furthering our understanding. I demonstrate this by a novel application of Hasok Chang's framework of 'Epistemic Iteration' (2004, 2012) to a very live and controversial puzzle in infant cognition, namely, whether babies can attribute false beliefs to others. Chang's aim is to show how progress can be made even when our starting point is shrouded in uncertainty, and I argue that the unfolding of the infant false belief research programme exemplifies this. Furthermore, Chang's notion of scientific progress places, front and centre, the idea that there are always multiple epistemic goals in play. While this is not a new idea , its emphasis helps us to see how, even though failed replications may not be informative about the hypothesis under consideration they nevertheless contribute to other epistemic aims, such as the validation or calibration of measurements or the refinement of concepts (section 3; see also van Dongen et al *unpub*). Finally, the paper uses the case study to illustrate one of Chang's most important contributions: that our scientific enquiries have to start somewhere. With hindsight, that starting point may look terribly bad. But in order for hindsight to occur, the starting point needs to be there. This is why failed replications are a necessary and expected part of good science: they are needed in order for the epistemic gains to be made that move us forward. A narrative of failed replications centred around 'distrust' not only masks these gains, but runs the risk of losing them altogether by casting dismissive doubt on the value of those fields currently experiencing high rates of failed replication.

2 Anticipatory looking: a case study

2.1 Children, babies, and the false belief task

The field of infant psychology is one which, I believe, is currently experiencing a large amount of uncertainty in some of its methods of measurements, while also grappling with conceptual questions about how to characterise the phenomena they intend to measure. Nowhere is this more manifest than in research examining infants' abilities to attribute psychological states to other agents. On the one side there are high-stakes[1] debates about the nature of the psychological states that infants attribute to others, and in particular whether they can attribute false beliefs to them. On the other side, there is a growing awareness that the methods of measurement, in particular those that rely on infants' spontaneous looking behaviours, are not as well understood as previously thought. Much of the key work in this field concerns pre-verbal infants[2] who have "limited attention spans, processing capacities and fine and gross motor skills" (Kominsky et al. 2022, p. 1). Consequently, most experimental paradigms rely on indirect measures to explore infants' cognitive capacities, e.g. by measuring how long a baby looks at a particular event, or where she looks. Some of the established causes for low rates of replication in the psychological sciences are attributed to small sample sizes leading to low statistical power, the specialised nature of the equipment required and

[1] Why these debates are considered 'high stakes' will be explained further in section 2.1.

[2] I will use the terms 'infants' and 'babies' to refer to children aged 24-months and younger, unless otherwise specified. This captures the age-range of most of the participants in this case-study.

lack of standardisation across measurement (Asendorpf et al. 2013; Collins 1985; Nosek et al. 2022). Infant psychology is a field afflicted by all these factors, plus the additional problem of incredibly sensitive and temperamental participants (Byers-Heinlein et al. 2020; Frank et al. 2017; Lavelle 2022; Peterson 2016). It is therefore unsurprising that there have been multiple studies in the field which researchers have had trouble replicating. This case-study focuses on one such replication project, concerning infants' understanding of other people's psychological states.

For decades it was widely accepted that children could not successfully attribute false beliefs to other people until around their fourth birthday. This was due to their performance on 'elicited false belief tasks'. In the original elicited false belief task (Wimmer and Perner 1983), 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, is an 'Elicited response' task as it requires the child to respond to a question asked by the experimenter: 'Where will Maxi look for his chocolate'.

This result for the elicited response false belief task has been replicated hundreds, if not thousands of times. It was therefore groundbreaking when Kristine Onishi and

Renée Baillargeon published a paper in 2005 arguing that 15-month olds showed evidence of attributing false beliefs to others. As 15-month olds cannot participate in elicted response tasks, they used a spontaneous response paradigm, which measured how long an infant looked at an event where an agent acts in a way that matches with their (the agent's) belief, in contrast to when the agent acts in a way that does not match with their belief. This is the *violation of expectation* paradigm, which works on the premise that infants look longer at events which surprise them (that is, which violate their expectations of what they predict will happen) than they do at events which match their expectations. They reported that infants would look longer at those test trials where the experimenter did not act in accordance with her belief about a toy's location, regardless of whether that belief was true or false, making the following claim:

> '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)

Naturally this paper caused quite a stir, disrupting the 'developmental dogma' of the previous twenty years that children below the age of 4-years could not attribute false beliefs to others (Rakoczy 2017). Until this point, the dominant conceptual frameworks had been designed to explain the developmental dogma; now these theories were hastily reconfigured to explain the new 'developmental gap' in performance between infants' responses on spontaneous response tasks, and children's performance on elicited response tasks. Onishi and Baillargeon's work was succeeded by a slew of research using a variety of spontaneous response methods to test infants' understanding of false beliefs, with a recent statement from Rose Scott and colleagues that "over thirty reports, using eleven different behavioral and neural methods, have yielded positive evidence of early false-belief understanding in non-traditional [i.e. spontaneous] tasks" (Scott, Roby, and Baillargeon 2022, p. 258). This paper follows the replication attempts of a spontaneous response task originally created by Victoria Southgate and colleagues (2007).[3] This task uses the 'Anticipatory Looking' (AL) paradigm, which is based on the premise that babies will look to where they expect an agent to go, before they see that agent's movements. Therefore if babies expect agents to behave in ways that are congruent with their (the agent's) beliefs, they should look to where an agent will look for an object, based on where that agent believes the object to be. The AL paradigm forms the basis of my case-study as there are multiple documented replication attempts, many of which use Southgate's stimuli.

At this point an important disclaimer is in order. Onishi and Baillargeon, Southgate and many others take the results of spontaneous response false belief tasks to support the hypothesis that infants can attribute false beliefs to others. This is a controversial explanation of the data. Other hypotheses abound: that infants' looking behaviour evidences the ability to track behavioural patterns in other agents, but that they do not attribute psychological states to them (Heyes 2014a; Heyes 2014b; Santiesteban et al. 2014); or that infants attribute psychological states to others that are similar to beliefs,

[3] This work was collaborative with Atsushi Senju and Gergely Csibra, but for brevity will be referred to as 'Southgate 2007'.

but which differ by being non-representational (Apperly and Butterfill 2009; Butterfill and Apperly 2013; Low et al. 2016). This paper will not evaluate these hypotheses.[4] Instead, it focuses on the existence of a phenomenon: whether infants anticipate that an actor will behave in a way that accords with her (the actor's) psychological states. I will refer to this as the 'Anticipation phenomenon'. The anticipation phenomenon describes a certain pattern of infant looking behaviour, but remains neutral on its causes, i.e. it makes no claims about whether the infant looks this way because she is attributing psychological states to the agent, or because she is tracking some behavioural pattern, or anything else. As the anticipation phenomenon is distinct from the diverse hypotheses evoked to explain it, should it turn out not to exist then each of the hypotheses just mentioned would require significant revision. Whether the anticipation phenomenon exists is the central question of this replication debate.

2.2 The anticipatory looking false belief task

In 2007 Victoria Southgate and colleagues published a study that used the anticipatory looking paradigm to examine 2-year olds' understanding of false beliefs. Participants watch a video showing a puppet, two boxes, each with a window above it and a human actor. First the baby watches the familiarisation trials: the puppet puts a ball in a box, while the actor watches; a chime sounds and two windows above the boxes flash, then the actor reaches through the window above the box with the ball in, placing her hand in the box. The baby watches this sequence twice (once for each box). The aims of the familiarisation trials are to show the baby that the actor wants the ball, and for the

[4] See Lavelle (2019) or Rakoczy (2022) for evaluations..

experimenters to check that the baby's looking behaviour demonstrates that they expect the actor to reach for where the ball is, i.e. that when the chime sounds the baby looks to the box where the ball is (more on this below). Next the babies watch one of two test conditions. In the first false-belief condition (*False-belief 1*), the actor watches as the puppet puts the ball in the left-side box, then moves it to the right-side box and closes the lid of the left. The actor then turns away, distracted by a phone ringing. The puppet takes the ball out of the right box and leaves the scene, taking the ball with it. The actor turns back to the scene and the chime sounds and the windows above the boxes flash. In this trial, babies should expect the actor to reach through the right window, with this expectation manifesting through (a) the babies looking first to the right window as soon as they perceive the chime and flashing cues (*first-look* measurement), and (b) by their looking longer at the right window than the left. The puppet's behaviour in the other test condition – *False-belief 2* – is the same as *False-belief 1*. But the actor is distracted as soon as the puppet places the ball in the left box and does not turn back to the scene until the puppet has left, meaning that she should reach through the left window when she turns back to the scene.

Southgate and colleagues reported that 9/10 infants in *False-belief 1* looked to the correct window when they perceived the cues, and 8/10 did so in *False-belief 2*. Regarding how long infants looked at the correct window, they write that "As the infants were familiarized to a delay of 1750ms between the onset of illumination and the opening of a window, we coded only the first 1750ms after onset of illumination on the test trial. The infants spent almost twice as long[5] focusing on the correct window as the incorrect window" (Southgate, Senju, and Csibra 2007, p. 590).

[5] An average of 956ms looking at the correct window, and 496ms at the incorrect window

As mentioned above, one of the roles of the familiarisation trials is to ascertain that infants show the right looking behaviours. Infants who did not look to where the actor should reach for the ball by the end of the second familiarisation trial were excluded from the study. This is because of two assumptions in the methodology:

1. That the baby's gaze direction indicates that they anticipate something to happen at that location.

2. That the baby's anticipation is caused by some kind of cognitive mechanism which tracks the actor's movements and predicts what she will do next.

These assumptions should be uncontroversial.[6] If infants did not show the right pattern of gaze in the familiarisation trials, this suggests either that they are not able to track simple goal-directed actions, or that their ability to do so is not revealed by the methodology. As both of these explanations for their behaviour mean that the AL methodology is not appropriate for examining that infant's understanding of false-belief, those who showed this behaviour were excluded from the study. An additional 11 babies were excluded from the study for failing to meet this criterion.

2.3 Replicating the anticipatory looking false belief task

This anticipatory looking false belief task has faced mixed replication success. Sebastian Dörrenberg and colleagues tested 66 two-year olds with Southgate's stimuli and found

[6] The exact nature of the cognitive mechanisms cited in (2) *are* subject to controversy (is the anticipation caused by attributing psychological states to the agent? or by tracking some behavioural cue?), but as explained in section 2.1 this is not a question for this paper.

that participants looked longer at the correct window only in *False-belief 1*. Similarly, infants' first look upon perceiving the cues were to the correct window in *False-belief 1*, but went more often to the incorrect window in *False-belief 2*. Tobias Schuwerk and colleagues also used Southgate's stimuli, but had to exclude 58% of participants (28 out of 48 children) for failing to look correctly at the end of the familiarisation period. Of the 20 participants who remained, only 7 looked first to the correct window, and there was no difference at all in how long they looked at the correct and incorrect windows across both trials. In the same year, Louisa Kulke and Hannes Rakoczy (2018) collected data on both published and unpublished attempts to replicate Southgate's experiment, showing that, of the 20 researchers who responded to their call for data, only 5 managed to successfully replicate Southgate's data (see *Table 1* for their criteria for evaluating replications).

[Insert Table 1 here]

What can be gleaned from this collection of replication data? Taking the more upbeat news first, it appears that more participants succeed in *False-belief 1* than in *False-belief 2* (Baillargeon, Buttelmann, and Southgate 2018). If robust, this pattern is something which theories of mindreading could reasonably accommodate. For example, infants need to hold in mind the actor's false-belief for longer in *False-belief 2* in contrast to *False-belief 1*, requiring a greater demand on their limited processing capacity and resulting in their forgetting the actor's belief and defaulting to reality. This would be in keeping with prominent accounts of why three-year olds fail elicited response tasks (Carruthers 2013, 2018, 2020; Scott and Baillargeon 2009, 2017) .

More worrying, however, is the lack of pattern in infants failing the familiarisation trials, ranging from over 50% of participants being excluded at this stage (Schuwerk, Priewasser, et al. 2018) to just 4% in other studies (Dörrenberg, Rakoczy, and Liszkowski 2018). On the basis of these data alone, one might question the anticipatory looking paradigm's suitability for measuring infants' anticipation of another's goal-directed movement, and this problem is made all the more pressing seeing as we do not understand why it works for some babies and not others. These data serve to highlight lacunae in our understanding of this methodology.

In their response to this, and other replication work concerning different false-belief tasks, Baillargeon, Buttelmann and Southgate wrote the following:

> "We do not agree with claims in some of the special-issue papers that these negative findings cast doubt on the conclusion that some capacity for belief understanding is already present in infants and toddlers. [...] [T]he non-replications stand in contrast to a large body of positive and convergent findings: as was mentioned earlier, over 30 published reports, using 11 different methods, have now provided evidence of false belief understanding in children under 3-years of age."(Baillargeon, Buttelmann, and Southgate 2018, p. 123)

Notably, these authors each support theories of mindreading which predict that infants should be able to attribute false-beliefs, and other psychological states, to other people. Yet researchers whose theoretical commitments lead them to be less confident that infants' understanding of psychological states stretches to false belief take quite a different interpretation of the replication data, claiming that we are not yet in a position to know whether infants attribute false beliefs to others (Poulin-Dubois et al. 2018).[7]

[7] These differences in opinion about whether the failed replications cast doubt on the hy-

Allow me to reiterate that the focus of this paper is the anticipation phenomenon (2.1), and not whether infants can attribute false beliefs to others. One can reasonably re-frame the debate just discussed to reflect this: one side believes that the data supports the existence of the anticipation phenomenon while the other does not; one side believes that a particular effect — infants looking towards where an agent will act — has been replicated while the other does not. What makes the debate more intractable are new doubts, revealed by this replication work, about how the AL paradigm works. This yields a double uncertainty. First, there is uncertainty about the phenomenon: we do not know whether infants expect another agent to act in accordance with her (the agent's) psychological states, which is why we are conducting the experiments in the first place. But additionally there is also uncertainty about our methods of measurement: we do not know if the AL paradigm is a reliable method, so that when infants' looking behaviour suggests they have not correctly anticipated the agent's behaviour we don't know if this is because they have not done so, or if they have but it somehow has not been captured by the constraints of the AL paradigm. These uncertainties about the measurement and the phenomenon in turn fuel interpretation of the replication data in different ways, dependant on one's prior theoretical leanings. Those who think infants can attribute psychological states to others will suggest there is something amiss with how the AL paradigm has been implemented, whereas those on the other side of the debate are more likely to accept the suitability of AL paradigm but question the existence of the phenomenon. This comes out particularly fiercely in an exchange about the suitability of

pothesis that infants and toddlers can understand false beliefs exemplify another problem running through replication debates, namely the 'experimenter's regress' (Collins 1985). For further discussion of this particular problem see Lavelle (2022).

the violation of expectation method for measuring infants' understanding of false belief, with Renée Baillargeon (2018) suggesting that small differences in how the paradigm was implemented were responsible for the failure to replicate her work. By contrast, Paula Rubio-Fernandez (2019) has expressed concerns that researchers are adjusting how they implement the paradigm until it yields results supportive of the view that infants can attribute false-beliefs to others (see also Peterson 2016) . And yet, if the phenomenon does exist (as many researchers believe it does), then calibrating our methods of measurement such that they can detect it could be a perfectly reasonable thing to do. The problems arise when, as here, there are doubts about the existence of the phenomenon.

This section has reviewed an on-going debate about how to interpret attempts to replicate Southgate's experiment, using the anticipatory looking paradigm to ascertain if infants can discriminate between belief-congruent and belief-incongruent behaviours. Thanks to these replication endeavours an important gap in our knowledge about the AL methodology has become apparent: we do not understand why a significant number of babies fail the familiarisation trial. This leads to more pressing questions in our application of the paradigm: what needs to be in place for us to be confident that it is suited to tracking infants' anticipations about events? And, when infants' looking behaviour fails to support the Anticipation hypothesis (2.1), is this because they have not made this discrimination, or because it has not been detected by the AL method?

The next section turns to work by Hasok Chang (2004, 2012) which argues that even when a field faces a conundrum such as the one outlined here, it is still able to yield epistemic goods. This is due to the process of 'Epistemic iteration', wherein by repeating experiments and keeping a variety of different theoretical options open, researchers are able to meet their epistemic goals and in so doing make progress with their discoveries. I will argue that replication is an essential part of the epistemic iterative process, and that therefore fields which experience high rates of failed replications can nevertheless be seen as producing important knowledge.

3 Epistemic Iteration

3.1 Imperfect ingredients and the ‘Principle of respect’

The structure of the puzzle outlined in section 2.3 is by no means unique to infant psychology. Every scientific field will, at various points in its history, have faced a problem where the current standards of measurement were inadequate for examining the phenomena researchers were interested in. Yet despite these uncertain foundations the scientists involved were able to progress towards their epistemic goals: calibrating a widely agreed new standard, improving theoretical unity or explanatory power, quality and quantity of evidence, or some other epistemic virtue (Chang 2004, p. 227). This movement, argues Chang, occurs thanks to the process he calls ‘Epistemic iteration’:

> “Epistemic iteration is a process in which successive stages of knowledge, each building on the preceding one, are created in order to enhance the achievement of certain epistemic goals. In each step, the later stage is based on the earlier stage, but cannot be deduced from it in any straightforward sense. Each link is based on the principle of respect and the imperative of progress, and the whole chain exhibits innovative progress within a continuous tradition. Iteration provides a key to understanding how

> knowledge can improve without the aid of an indubitable foundation. What we have is a process in which we throw very imperfect ingredients together and manufacture something just a bit less imperfect." (Chang 2004, p. 46)

Progress begins when a community acknowledges that their current system of knowledge is imperfect. In Chang's example, scientists realised that our sensations of hot and cold were insufficient to permit the investigation of the phenomena they were interested in. In our case, we could say that, prior to Onishi and Baillargeon's pioneering work, we lacked a method to investigate infants' understanding of false beliefs, as the only methods available were designed for children over 36-months. Moving forward to the debate as it stands today: replications of Southgate's work has served to spotlight 'imperfections' in our understanding of the AL paradigm, for example our lack of knowledge for why performance in the familiarisation trials is so variable. This is one of the most valuable functions of replications: highlighting gaps in our knowledge of which we were previously unaware (see 3.3).

How do we move on from this state of uncertainty? Here Chang argues that we develop a new standard, whose relation to the old one is captured by the 'principle of respect'. Our first iteration of thermoscopes needed to respect our folk sensations of temperature, showing that the things we reliably perceive as hot show a higher temperature than those that we reliably perceive as cold. But while guided by our sensations, the thermoscopes were not constrained by them. For, in being more accurate than our sensations they could later be used to correct judgments of temperature based on sensation alone: while a hand that has been in the snow will feel a bucket of tepid water as warm, and one that has been snug in a mitten will feel it as cold, the thermoscope will reveal that the water is a uniform temperature (Chang 2004, p. 43).

We see the principle of respect in action in the on-going multi-laboratory *Many Babies 2* collaboration, which is conducting a large-scale replication project concerning whether babies expect an agent to look for something based on her knowledge of where that thing is (Schuwerk, Kampis, et al. 2022). The study uses the AL paradigm. One of the 'imperfect' foundations upon which we set the AL paradigm is our acceptance that babies can attribute goals to other agents, and expect them to act on these goals. There are multiple lines of support for this acceptance. First, we know that adults cannot help but see certain movements as goal-directed, as was shown most famously by Heider and Simmel's work (1944) . Second, it is a feature widely observed in the non-human animal kingdom, from a pride of lions hunting an impala to Sarah the chimpanzee recognising the various outcomes her trainer's behaviour was aimed towards (Woodruff and Premack 1978). Third, there are strong evolutionary arguments for the ability to recognise goal-directed movements early in development as a critical means of enhancing survival. Fourth, there are a number of experiments using a range of different methods (e.g. the visual habituation paradigm), yielding evidence to support the claim that by 8-months infants reliably distinguish goal-directed from non-goal directed movements.[8] And finally, but by no means least, caregivers through the ages have treated their babies as though they can recognise goal-directed actions. Taken as a whole, this collection of reasons from a range of disciplinary perspectives — while imperfect — nevertheless give a foundation against which to calibrate an instance of the AL paradigm: if babies do not respond to a particular set of stimuli in ways that indicate that they have attributed a goal to the protagonist, then those stimuli need to be reconfigured until such a response

[8] See Luo (2010) for a review.

is reliably procured. The Epistemic Iteration framework explains why this kind of calibration is acceptable: we are calibrating to an imperfect starting point, but provided we keep an open mind about how the next iteration of measurement might change this (see below), it will be good enough. From their pilot work, the Many Babies 2 team are confident that their implementation of the AL paradigm is able to track babies' expectations of other's goal directed movements, with 68% of toddlers (65 x (18 - 25m)) and 69% of adults (42) looking to where a chaser (a bear) would go in order to catch a chasee (a mouse) (Schuwerk, Kampis, et al. 2022, p. 19).[9]

3.2 Enrichment, correction and contradiction

In the previous section I loosely used the phrase 'keep an open mind' about how iteration could change our imperfect starting point. I now draw on three more concepts from Chang to explain what this entails.

First, our new measurements may contradict our previous ones in some ways (see the example of the tepid bucket of water above). Some contradiction can be tolerated: after all, the whole point of developing a new system of measurement is because the previous one is in some way inadequate, so we should expect some differences in their outputs. But if every instantiation of the new system leads to a contradiction with the old, then this gives us good reason to abandon the new system. For example, if we could not generate any stimuli that caused babies to look to where a goal-directed agent should go,

[9] While the authors do not comment on why 30% of participants did not show the anticipatory looking, this can be explained by appeal to individual differences in attention span or motivation.

then this would raise questions of the suitability of the AL method for this age-group. Such doubt would be compounded if other methods did show that babies anticipate other people's goal-directed actions. But there is also a more subtle manifestation of this problem peculiar to infant cognition. Babies have very limited cognitive and motor abilities, and in adjusting the stimuli until participants show anticipatory looking behaviours one can end up with images and situations that are very far removed from the everyday reality that babies typically encounter. For example, the Many Babies 2 stimuli are a simple cartoon bear and mouse, an upside-down Y-shaped tunnel through which the bear chases the mouse, and a box at each end of the Y where the mouse hides. But generalisability is inherent to the nature of the cognitive capacity we are studying: if babies only show looking behaviours consistent with goal-attribution in a very specific circumstance and no other, then this is insufficient to support the claim that they anticipate other's goal directed behaviours, because this ability is meant to underpin all (or most) perceptions of goal-directed actions and not just a tiny sub-set of them.[10] If babies' looking behaviour were specific just to one set of stimuli, this would contradict the hypothesis at the centre of our imperfect foundation and lend support to abandoning the AL paradigm.

The second virtue of the iterative process is 'enrichment' wherein "the initially affirmed system is not necessarily negated but refined, resulting in the enhancement of some of its epistemic virtues" (Chang 2004, p. 228). The Many Babies 2 team are confident that their stimuli reliably cause babies to look where they expect the bear to

[10] Addressing this issue more substantively requires a closer analysis of the fragility of experimental effects, which remains a topic for another paper (although see Feest (2022), Kominsky et. al (2022) and van Bavel (2016) for contributions in this line).

chase the mouse. This places them in a position to extend their method from collecting data about a phenomenon about which we are reasonably confident (babies' ability to anticipate goal-directed action) to one about which we are less certain: babies' ability to anticipate what someone will do based on their epistemic states (knowledge versus ignorance). This work is currently underway, using the same stimuli as described for the study above with a minimal adjustment: whether the bear sees which box the mouse enters upon leaving the tunnel. If the babies' looking patterns do not show that they expect another to act on their knowledge states, the researchers can be reasonably confident that this is due to the babies' cognitive limitations rather than quirks of the stimuli or measurement window, as these remain the same as the pilot. This process instantiates the Principle of Respect, and also illustrates how the iterative process can lead to progress in allowing methods of experimentation to extend to new domains.

The last virtue of Epistemic Iteration that Chang discusses is 'self-correction'. This occurs when a new standard gives us reason to adjust our hypotheses that were based on data from the old standard. In this case, one could call the Many Babies 2 stimuli a step towards a new standard. However, the stimuli themselves cannot be the standard, for the reasons explained at the start of this section. Instead, we need to develop our understanding of *why* these stimuli are more successful at eliciting goal-directed anticipatory looking. Once this has been done, the principles can be applied to the creation of new stimuli which give more uniform data concerning false belief than Southgate's. Whether a self-correction is required depends on how these data turn out. Another form of self-correction is evident in the calibration process described above, as the Many Babies' 2 team developed their stimuli. The adjustment made to the stimuli to get the effect of anticipatory looking is itself a process of self-correction, and can only occur through repeatedly testing different participants.

3.3 Multiple epistemic goals

Central to Chang's framework is the idea that there are always multiple goals at work in scientific research, and his emphasis on this aspect is helpful for understanding the epistemic gains made in our case study, and through replication work more generally. More often than not, the stated goal of an experiment is to provide data for or against a specific hypothesis. If this is one's only goal, then failed replications are certainly problematic. Popper famously argued that replicating results is necessary for distinguishing data that supports a hypothesis from "mere isolated coincidence" (Popper 1959, p. 45). Later, Collins (1985) articulated the problem of the 'Experimenters' regress', namely, how different research teams decide which experimental outcome is the 'correct' one: that of the original or of the failed replication (see Feest 2016 for further discussion). Returning to our case study: the data from the replications are insufficient to allow us to evaluate the Anticipation hypothesis, thus they fail to meet this epistemic goal. Yet despite failure on this front, the analysis above shows how progress has been made towards achieving other epistemic goals: improving our understanding of how the anticipatory looking paradigm works, and in so doing making it a more reliable measure of infants' expectations. This view of progress seems to capture the epistemic gains that come from replication work better than a single-minded focus on whether the results support the hypothesis under consideration.

One worry with this characterisation of progress is that it does not match with how experimenters view their own work. Southgate and colleagues' aim was to test their false-belief hypothesis; the aim of those conducting the replication work was to test the Anticipation hypothesis; none of these parties succeeded in attaining these ends. Is it fair to argue for progress on the grounds that different epistemic goals have been achieved, when it is not at all clear that anyone involved in the work has these goals in mind?[11] I think this question can be addressed by revisiting part of the quotation cited in section 3.1: "Epistemic iteration is a process in which successive stages of knowledge, each building on the preceding one, are created in order to enhance the achievement of certain epistemic goals. *In each step, the later stage is based on the earlier stage, but cannot be deduced from it in any straightforward sense.*" This investigation of infants' understanding of false beliefs began with an imperfect foundation: the assumption that the AL methodology would be able to provide evidence for or against the false-belief hypothesis. From this beginning, it could not be deduced in a straightforward sense that the next step would be to dissemble the methodology. That this would be a productive step only became apparent later in the research journey, when the failed replications came in. It seems uncontroversial to say that improving our understanding of the AL methodology is an epistemic gain. But it is not one that could have been foreseen from the starting point, and thus could not have been a goal. Crucially, without the imperfect starting point, these gains would not have been possible. This is a liberal view of scientific progress but I do not think it is too liberal. It gives boundary limits for when more experiments are unhelpful: when they fail to meet any of the epistemic goals mentioned above. But it is nevertheless healthy, for science and philosophy, to consider the crossing-off of wrong answers to be a form of progress.

[11] I am grateful to an anonymous reviewer for raising this question

3.4 Uncertainty revisited

This section has argued that Epistemic Iteration offers a way of understanding how infant psychology can make epistemic gains despite the dual doubts — about the reliability of the AL method and the existence of the anticipation phenomenon — at its foundation. By using the Principle of Respect and building out from our initial assumption that infants can attribute goals to others, we can begin to calibrate the AL methodology, which in turn increases our confidence in its reliability when applied to phenomena we are less certain of, such as anticipating another's actions based on their knowledge states. Crucially, this iterative process can be applied to the other spontaneous methodologies which face the same double uncertainties about measurement and the existence of a phenomenon (e.g. Buttelman's spontaneous helping paradigm (2009) or the violation of expectation paradigm). Calibrating and standardising spontaneous methodologies is a key epistemic goal for infant psychology, and the above discussion outlines how this is possible even when we are uncertain about the phenomena in question.

One worry about this application of Epistemic Iteration is that the cases of infant cognition and temperature are disanalogous.[12] Those developing the first instruments to measure temperature knew that there was a phenomenon 'out there' to be measured, they were just unsure how to go about measuring it. By contrast, the central question of the infant psychology debate is whether babies expect people to act in ways that are congruent with their psychological states, and if so what the limits of this ability are (goal states,[13] knowledge states, belief states etc., as well as the content of these states).

[12] Thanks to Fan Yichu for pushing me on this point.

[13] Some authors deny that goal states are psychological (Roessler and Perner 2015), but the nuances of this particular debate are not relevant here.

In other words, it's not clear that a phenomenon exists to be measured, unlike the case of temperature. As observed by Kenneth Kendler, one cannot iterate "towards a target that isn't there" (Kendler 2012, p. 308).

I think this concern can be mitigated from two different angles. First, Chang himself is clear that Epistemic Iteration is valuable in helping us achieve our epistemic goals, even when we are unsure about whether our inquiries are targeting the phenomena we are after (see also Schaffner 2012):

> "It [Epistemic iteration] differs crucially from mathematical iteration in that the latter is used to approach the correct answer that is known, or at least in principle knowable, by other means. In epistemic iteration that is not so clearly the case."(Chang 2004, p. 45)

A null result is nevertheless an epistemic gain. If, after numerous iterative attempts at calibration and standardisation across all spontaneous methodologies, babies do not show looking behaviours consistent with the hypothesis that they anticipate the actions of other agents, then we accept that the research programme has contradicted its core hypothesis and that babies do not have this cognitive ability. As mentioned above, being able to cross-off a wrong answer can be useful.

A more likely scenario is that after several iterative processes aimed at improving calibration and standardisation for each spontaneous methodology there is no consensus about the nature of infants' anticipation of other agents' actions. This brings me to the second angle from which to address the worry. For, just as results from the AL methodology alone are insufficient to support the claim that babies anticipate other agents' actions, neither would the results from all spontaneous methodologies be sufficient to support this claim. Spontaneous methodologies are but one way of exploring and investigating infant cognition. Babies and their carers have, quite literally, always been a part of human history, and there is a vast, messy and contradictory body of folk knowledge about their abilities. I am reminded here of a passage from Jennifer Nansubuga Makumbi's novel *The First Woman* where a Ugandan trainee nurse writes home with news about her first days at medical school:

> "We have two orphan babies. I am not lying. Real breathing human babies, donated to the school by Ssanyu Babies' Home, to learn how to look after babies – winding and bathing them, tying nappies and diet. I said, but these Europeans know how to waste time. Who taught our mothers to bring up children?"

We are not yet in a position to know what infants know about other people's actions. But what we do know is that infants grow into pre-schoolers who can track false beliefs in others and recognise when someone is hiding their true emotions (Wellman 2014), and eventually into adults who can track three or four levels of deceit in Shakespearean-style plots. Care-givers do not notice a seismic change in their children when they go from failing to passing false-belief tasks, nor when they pass any other purportedly significant mindreading milestone[14] in tracking psychological states. We assume that infants know *something* about the actions of others, and as such there is a phenomenon there to be explored, no matter how crudely outlined.[15] Folk knowledge and evidence from other

[14] For examples of such milestones see Wellman (1990, 2001, 2014)

[15] An evaluation of how psychologists begin their initial 'ball-park' descriptions of phenomena is a topic for another paper (see Adetlua 2002, Haig 2013, Muthukrishnan and Henrich 2019 and Rozin 2001 for thoughtful contributions in this line.)

sources (see section 3.1) combined with the principle of respect are sufficient to ensure we start our investigations in broadly the right ball-park; and even if the phenomenon under investigation is even less well-understood than temperature was prior to the first thermometers, this does not foreclose the prospect of epistemic iteration leading to the fulfilment of our epistemic goals.

4 Imperfection not falsehood

This paper opened with a quote from an editorial in the journal *Nature* stating that two-thirds of what we read in psychology journals should not be trusted. This section reviews this sentiment in the light of the discussion in section 3.

The aim of Chang's framework is to show how we can make epistemic inroads in a scientific investigation, be our starting point ever so bad. From an imperfect starting point and with imperfect methodologies we can nevertheless end up with a better understanding of a phenomenon than that which we started with. Critically, the knowledge we gain would not have been possible had we not started *somewhere*: the imperfect starting point is necessary to attaining the goods that follow. This position stands in contrast to those who perceive a large number of failed replications to indicate untrustworthy science. A large number of failed replications should be expected when the starting point is bad, because there is so much uncertainty about the concepts under investigation and the methods used to find out about them. The problems arise when researchers fail to acknowledge their work for what it is: a process of building outwards from an uncertain foundation. An important lesson being learned from the replication crisis is that this starting point needs to be made more explicit (Bringmann, Elmer, and

Eronen 2022; Feest 2022; Sikorski and Andreoletti 2023).

Second, one cannot build the foundation of a scientific research programme on distrust.[16] But Epistemic Iteration shows that one can build such a foundation upon imperfection. This is not a simply an issue of petty word-play. Inherent in the distrust narrative is the sense that one would be irrational to continue in a field where so many findings fail to be replicated. Indeed, this is expressed with some force by Tal Yarkoni, who exhorts psychology graduates to go do something else with their lives (Yarkoni 2020). Epistemic Iteration, on the contrary, shows such a starting point to be acceptable as it implies that there is considerable scope for improvement, and plenty of work for scientists to do.

One may object that this is an overly Pollyanna-ish interpretation of a field with many failed replications. Sometimes, so the criticism goes, we *should* take a slew of failed replications to indicate that a hypothesis or research programme ought to be abandoned. How can we distinguish between a foundation that is imperfect but has scope for improvement, and one that is hopeless? We distinguish it through the system's ability to achieve the epistemic goals it sets, and those which consistently face self-contradiction in the pursuit of these goals can be abandoned (3.2). Getting the same data from the same methods is one epistemic goal, but it is not the only one; subsequently a large number of failed replications should not be the only reason to abandon a research programme.

[16] A reviewer cites Merton's 'Organised Scepticism' as a counter to this claim (Merton 1973). I agree. But the scepticism urged by Morton seems a more respectful kind than the dismissive tone often given to findings that fail to replicate in the current climate.

5 Conclusions

Infant psychology is a field with a high number of failed replications. Yet it is also, as argued in this paper, a field where significant epistemic gains are being made in our understanding of the methods used to investigate infant cognition. This is the case despite the high degree of uncertainty in the field regarding both the phenomena under investigation and the reliability of the methods used to examine them. This paper offers an explanation for how this can be in the form of Epistemic Iteration. Epistemic Iteration offers the tools to see how we can progress towards our epistemic goals even when our starting point, both in terms of the phenomena under examination and the methods used to examine it, is imperfect. When a field is in this stage of having relatively few affirmed foundations, it is unsurprising that it also has many instances of failed replications as there is so little to build on (Irvine 2021). Importantly, we need to start somewhere, and without the messy data generated by these imperfect concepts we would not be in a position to work out how we might advance our epistemic goals. It is as we start building on these data that we come closer to creating experiments that can be replicated.

There are several big issues that have been skirted in this piece which I defer to later papers. The biggest is how we should view progress within infant psychology, or even psychology as a more general field. The paper accepts without much defence Chang's proposal of progress as characterised by meeting epistemic goals, which gives a very localised view of progress as the goals of most epistemic import will vary from field to field and from time to time within a field. Future work could offer further defence of this view of progress, and of the coherentist approach more broadly endorsed by Chang, as appropriate for psychology. Another question is that raised in section 3.2 regarding the balance between making stimuli that are appropriate for infants and concerns about ecological validity and generalisability. The concern from ecological validity is that the stimuli are so different from life as encountered in the real world that one needs to carefully justify the claim that they are tapping into the same cognitive abilities that babies use 'in the wild'. The concern from generalisability is that the stimuli may be testing a very specific cognitive ability, e.g. an infant's theory of cartoon bears and mice, rather than the indefinitely flexible ability to track goals which is the real target of investigation (Feest 2022; Packer and Moreno-Dulcey 2013).

Through this survey of replications of Southgate's anticipatory looking false belief task and the Many Babies 2 project we see research that, far from being untrustworthy, exemplifies progress through the iterative processes of self-correction and enrichment. Research into infants' abilities to attribute psychological states to others has very few certain foundations, and I have shown how the progress made to date is based on the most stable, but still imperfect, of these, namely infants' ability to anticipate goal-directed actions. Thus, as failed replications are compatible with flourishing, progressive science, it is time to sever the connection between 'does not replicate' and 'untrustworthy, and to instead recognise the necessity of this work for the epistemic iterative cycle of accumulating knowledge.

Acknowledgements

This research was supported by a Humboldt Experienced Researcher fellowship, and by a British Academy grant 'Replication: crisis or opportunity' (SRG2000688) which funded a series of workshops where these ideas were developed. The author gratefully acknowledges these funders. Thanks also to participants at the BA workshops, Enno Fischer, Barry Maguire, the ‘Consciousness and cognition’ research group at Ruhr Universität, Bochum, and two anonymous reviewers for their time and invaluable feedback.

References

Adetula, Adeyemi et al. (2022). “Psychology should generalize from — not just to — Africa”. In: *Nature Reviews Psychology* 1.7, pp. 370–371. DOI: 10.1038/s44159-022-00070-y.

Apperly, Ian A and Stephen A Butterfill (2009). “Do humans have two systems to track beliefs and belief-like states?” In: *Psychological review* 116.4, p. 953.

Asendorpf, Jens B. et al. (2013). “Recommendations for increasing replicability in psychology”. In: *European Journal of Personality* 27.2, pp. 108–119. DOI: 10.1002/per.1919.

Baillargeon, Renée, David Buttelmann, and Victoria Southgate (2018). “Invited Commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations”. In: *Cognitive Development* 46.May, pp. 112–124. DOI: 10.1016/j.cogdev.2018.06.001.

Baker, Monya (2015). “Over half of psychology studies fail reproducibility test”. In: *Nature* 27, pp. 1–3. DOI: 10.1038/NATURE.2015.18248.

Bird, Alexander (2021). “Understanding the Replication Crisis as a Base Rate Fallacy”. In: *The British journal for the philosophy of science* 72.4, pp. 965–993.

Bringmann, Laura F., Timon Elmer, and Markus I. Eronen (2022). “Back to Basics: The Importance of Conceptual Clarification in Psychological Science”. In: *Current Directions in Psychological Science* 31.4, pp. 340–346. DOI: 10.1177/09637214221096485.

Buttelmann, David, Malinda Carpenter, and Michael Tomasello (2009). “Eighteen-month-old infants show false belief understanding in an active helping paradigm”. In: *Cognition* 112.2, pp. 337–342. DOI: 10.1016/j.cognition.2009.05.006.

Butterfill, Stephen A and Ian A Apperly (2013). “How to construct a minimal theory of mind”. In: *Mind & Language* 28.5, pp. 606–637.

Byers-Heinlein, Krista et al. (2020). “Building a collaborative psychological science: Lessons learned from ManyBabies 1”. In: *Canadian Psychology* 61, pp. 349–363.

Carruthers, Peter (2013). “Mindreading in infancy”. In: *Mind and Language* 28.2, pp. 141–172. DOI: 10.1111/mila.12014.

— (2018). “Young children flexibly attribute mental states to others”. In: *Proceedings of the National Academy of Sciences of the United States of America* 115.45, pp. 11351–11353. DOI: 10.1073/pnas.1816255115.

— (2020). “Representing the Mind as Such in Infancy”. In: *Review of Philosophy and Psychology*. DOI: 10.1007/s13164-020-00491-9.

Chang, Hasok (2004). *Inventing Temperature: Measurement and Scientific Progress.* Oxford University Press, pp. 1–304. DOI: 10.1093/0195171276.001.0001.

— (2012). *Is Water H2O? Evidence, Realism and Pluralism.* Springer.

Collaboration, Open Science (2015). “Estimating the reproducibility of psychological science”. In: *Science (New York, N.Y.)* 349.6251, aac4716. DOI: 10.1126/science.aac4716.

Collins, H.M. (1985). *Changing Order: replication and induction in scientific practice.* Sage.

Dörrenberg, Sebastian, Hannes Rakoczy, and Ulf Liszkowski (2018). “How (not) to measure infant Theory of Mind: Testing the replicability and validity of four non-verbal measures”. In: *Cognitive Development* 46.February 2017, pp. 12–30. DOI: 10.1016/j.cogdev.2018.01.001.

Feest, Uljana (2016). “The experimenters’ regress reconsidered: Replication, tacit knowledge, and the dynamics of knowledge generation”. In: *Studies in History and Philosophy of Science Part A* 58, pp. 34–45. DOI: 10.1016/j.shpsa.2016.04.003.

— (2019). “Why replication is overrated”. In: *Philosophy of Science* 86.5, pp. 895–905. DOI: 10.1086/705451.

— (2022). “Data quality, experimental artifacts, and the reactivity of the psychological subject matter”. In: *European Journal for Philosophy of Science* 12.1, pp. 1–25. DOI: 10.1007/S13194-021-00443-9/FIGURES/1.

Fletcher, Samuel C. (2021). “The role of replication in psychological science”. In: *European Journal for Philosophy of Science* 11.1. DOI: 10.1007/s13194-020-00329-2.

Frank, Michael C. et al. (2017). “A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-Building”. In: *Infancy* 22.4, pp. 421–435. DOI: 10.1111/INFA.12182.

Haig, Brian D (2013). “Detecting Psychological Phenomena: Taking Bottom-Up Research Seriously”. In: *American Journal of Psychology* 126.2, pp. 135–153.

Heider, Fritz and Marianne Simmel (1944). “An Experimental Study of Apparent Behavior”. In: *The American journal of psychology* 57.2, pp. 243–259.

Heyes, Cecilia (2014a). “False belief in infancy: a fresh look”. In: *Developmental science* 17.5, pp. 647–659.

— (2014b). “Submentalizing: I am not really reading your mind”. In: *Perspectives on Psychological Science* 9.2, pp. 131–143.

Irvine, Elizabeth (2021). “The Role of Replication Studies in Theory Building”. In: *Perspectives on psychological science* 16.4, pp. 844–853. DOI: 10.1177/1745691620970558.

Kendler, K.S. (2012). “Epistemic iteration as a historical model for psychiatric nosology: promises and limitations”. In: *Philosophical issues in psychiatry II: Nosology*. Ed. by K.S. Kendler and J. Parnas. Oxford: Oxford University Press, pp. 305–322.

Kominsky, Jonathan F et al. (2022). “Simplicity and validity in infant research”. In: *Cognitive Development* 63, pp. 1–13. DOI: https://doi.org/10.31234/osf.io/6j9p3.

Kulke, Louisa and Hannes Rakoczy (2018). “Implicit Theory of Mind – An overview of current replications and non-replications”. In: *Data in Brief* 16, pp. 101–104. DOI: 10.1016/j.dib.2017.11.016.

Lavelle, Jane Suilin (2019). *The Social Mind: A philosophical introduction*. Routledge.

— (2022). “When a crisis becomes an opportunity: the role of replications in making better theories”. In: *British Journal for the Philosophy of Science* 73.4, pp. 965–986.

Leonelli, Sabina (2018). “Rethinking reproducibility as a criterion for research quality”. In: *Research in the History of Economic Thought and Methodology* 36B, pp. 129–146. DOI: 10.1108/S0743-41542018000036B009.

Loscalzo, Joseph (2012). *Irreproducible experimental results: causes,(mis) interpretations, and consequences.*

Low, Jason et al. (2016). “Cognitive architecture of belief reasoning in children and adults: A primer on the two-systems account”. In: *Child Development Perspectives* 10.3, pp. 184–189.

Luo, Yuyan (2011). “Three-month-old infants attribute goals to a non-human agent”. In: *Developmental Science* 14.2, pp. 453–460. DOI: 10.1111/J.1467-7687.2010.00995.X.

McNutt, Marcia (2014). “Reproducibility”. In: *Science* 343, p. 229.

Merton, Robert (1973). “The normative structure of science (1942)”. In: *The sociology of science : theoretical and empirical investigations.* Ed. by Norman W. Storer. University of Chicago Press, pp. 267–278.

Muthukrishna, Michael and Joseph Henrich (2019). “A problem in theory”. In: *Nature Human Behaviour* 3.3, pp. 221–229. DOI: 10.1038/s41562-018-0522-1.

Nosek, Brian A. et al. (2022). “Replicability, Robustness, and Reproducibility in Psychological Science”. In: *Annual Review of Psychology* 73.1. PMID: 34665669, pp. 719–748. DOI: 10.1146/annurev-psych-020821-114157.

Packer, Martin J. and Fernando Moreno-Dulcey (2013). “This puppet will play a game with you: is its time to take child psychology out of the laboratory”. In: *Journal of Chemical Information and Modeling* 53.9, pp. 1689–1699.

Peterson, David (2016). “The baby factory: Difficult research objects, disciplinary standards, and the production of statistical significance”. In: *Socius* 2, pp. 1–10.

Popper, Karl (1959). *The logic of scientific discovery.* Hutchison.

Poulin-Dubois, Diane et al. (2018). “Do infants understand false beliefs? We don’t know yet – A commentary on Baillargeon, Buttelmann and Southgate’s commentary”. In: *Cognitive Development* 48.November, pp. 302–315. DOI: 10.1016/j.cogdev.2018.09.005.

Rakoczy, Hannes (2017). “In defense of a developmental dogma: children acquire propositional attitude folk psychology around age 4”. In: *Synthese* 194.3, pp. 689–707. DOI: 10.1007/s11229-015-0860-8.

— (2022). “The development of implicit theory of mind”. In: *The Routledge Handbook of the Philosophy of Implicit Cognition.* Ed. by J. Robert Thompson. Routledge. Chap. 26, pp. 336–350.

Roessler, Johannes and Josef Perner (2015). “Pro-social cognition: helping, practical reasons, and ‘theory of mind’”. In: *Phenomenology and the cognitive sciences* 14.4, pp. 755–767.

Rozin, Paul (2001). “Social Psychology and Science: Some Lessons From Solomon Asch”. In: *Personality and Social Psychology Review* 5.1, pp. 2–14.

Rubio-Fernández, Paula (2019). “Publication standards in infancy research: Three ways to make Violation-of-Expectation studies more reliable”. In: *Infant Behavior and Development* 54, pp. 177–188. DOI: 10.1016/j.infbeh.2018.09.009.

Santiesteban, Idalmis et al. (2014). “Avatars and arrows: Implicit mentalizing or domain-general processing?” In: *Journal of Experimental Psychology: Human Perception and Performance* 40.3, p. 929.

Schaffner, Kenneth (2012). “Coherentist approaches to scientific progress in psychiatry: comments on Kendler”. In: *Philosophical issues in psychiatry II: Nosology.* Ed. by K.S. Kendler and J. Parnas. Oxford: Oxford University Press, pp. 323–330.

Schickore, Jutta (2011). “The Significance of Re-Doing Experiments: A Contribution to Historically Informed Methodology”. In: *Erkenntnis* 75.3, pp. 325–347. DOI: 10.1007/s10670-011-9332-9.

Schmidt, Stefan (2009). “Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences”. In: *Review of General Psychology* 13.2, pp. 90–100. DOI: 10.1037/a0015108.

Schuwerk, Tobias, Dora Kampis, et al. (2022). “In-principle acceptance of Registered Report: Action Anticipation Based on an Agent’s Epistemic State in Toddlers and Adults”. In: *Child Development.* DOI: 10.31234/osf.io/x4jbm.

Schuwerk, Tobias, Beate Priewasser, et al. (2018). “The robustness and generalizability of findings on spontaneous false belief sensitivity: A replication attempt”. In: *Royal Society Open Science* 5.5. DOI: 10.1098/rsos.172273.

Scott, Rose M. and Renée Baillargeon (2009). “Which penguin is this? attributing false beliefs about object identity at 18 months”. In: *Child Development* 80.4, pp. 1172–1196. DOI: 10.1111/j.1467-8624.2009.01324.x.

— (2017). “Early False-Belief Understanding”. In: *Trends in Cognitive Sciences* 21.4, pp. 237–249. DOI: 10.1016/j.tics.2017.01.012.

Scott, Rose M., Erin Roby, and Renée Baillargeon (2022). “How Sophisticated Is Infants’ Theory of Mind?” In: *The Cambridge Handbook of Cognitive Development.* Ed. by Olivier Houdé and Grégoire Borst. Cambridge Handbooks in Psychology. Cambridge University Press, pp. 242–268. DOI: 10.1017/9781108399838.015.

Sikorski, Michał and Mattia Andreoletti (2023). "Epistemic Functions of Replicability in Experimental Sciences: Defending the Orthodox View". In: *Foundations of Science*. DOI: 10.1007/s10699-023-09901-4.

Simons, Daniel J (2014). "The value of direct replication". In: *Perspectives on psychological science* 9.1, pp. 76–80.

Southgate, Victoria, Atsushi. Senju, and Gergely. Csibra (2007). "Action anticipation through attribution of false belief by 2-year-olds". In: *Psychological Science* 18.7, pp. 587–592. DOI: 10.1111/j.1467-9280.2007.01944.x.

Van Bavel, Jay J. et al. (2016). "Contextual sensitivity in scientific reproducibility". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.23, pp. 6454–6459. DOI: 10.1073/pnas.1521897113.

Van Dongen, Noah et al. (n.d.). "Productive Explanation: A Framework for Evaluating Explanations in Psychological Science". In: *url = https://psyarxiv.com/qd69g/,* ().

Wellman, Henry M (1990). Cambridge, Massachusetts: The MIT Press.

— (2014). *Making minds : how theory of mind develops*. Oxford series in cognitive development. New York: Oxford University Press.

Wellman, Henry M, David Cross, and Julanne Watson (2001). "Meta-analysis of theory-of-mind development: The truth about false belief". In: *Child Development* 72.3, pp. 655–684. DOI: 10.1111/1467-8624.00304.

Wimmer, Heinz and Josef Perner (1983). "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception". In: *Cognition* 13.1, pp. 103–128. DOI: 10.1016/0010-0277(83)90004-5.

Woodruff, Guy and David Premack (1978). "Does the chimpanzee have a theory of mind?" In: *Brain and behavior sciences* 1, pp. 515–526.

Yarkoni, Tal (2020). “The generalizability crisis”. In: *Behavioral and Brain Sciences* 45, pp. 1–37. DOI: 10.1017/S0140525X20001685.