Synthese (2021) 198:8213–8232
https://doi.org/10.1007/s11229-020-02566-7

FOLK PSYCHOLOGY: PLURALISTIC APPROACHES

Folk personality psychology: mindreading and mindshaping in trait attribution

Evan Westra[1]

Received: 6 May 2019 / Accepted: 31 January 2020 / Published online: 3 February 2020

Abstract
Character-trait attribution is an important component of everyday social cognition that has until recently received insufficient attention in traditional accounts of folk psychology. In this paper, I consider how the case of character-trait attribution fits into the debate between mindreading-based and broadly 'pluralistic' approaches to folk psychology. Contrary to the arguments of some pluralists, I argue that the evidence on trait understanding does not show that it is a distinct, non-mentalistic mode of folk-psychological reasoning, but rather suggests that traits are ordinarily understood as mentalistic dispositions. I also examine several ways in which trait attribution might also serve regulative, 'mindshaping' functions by promoting predictable norm-governed behavior, and argue that mindreading plays several important roles in these cases as well. I conclude that an appreciation of the relationship between trait attribution and mindreading is crucial to understanding the role it plays in our folk psychology.

Keywords Folk psychology · Mindreading · Theory of mind · Character traits · Mindshaping · Pluralism · Social cognition

1 Introduction

The term 'folk psychology' refers to "our everyday capacity to make sense of the behavior of other agents" (Spaulding 2018a, p. 8). Among mainstream philosophers of cognitive science, folk psychology has been construed as the prediction and explanation of behavior in terms of mental-state concepts, such as belief, desire, and intention—a capacity also referred to as *mindreading* or *theory of mind* (Carruthers and Smith 1996). Much of the debate surrounding mindreading has focused on the procedure we use to attribute mental states [e.g. via quasi-scientific theorizing (Gopnik and Wellman 1992), first-person simulation (Gordon 1986), or via a perception-like

✉ Evan Westra
evan.westra@utoronto.ca

[1] Department of Philosophy, University of Toronto, 170 St George Street, Toronto, Canada

module (Scholl and Leslie 1999)], and how and when this ability develops in childhood [e.g. whether the capacity to attribute mental states is innately channeled or acquired via experience (Carruthers 2013; Wellman 2014)].Throughout all these different debates and the different positions within them, the major focus of mindreading theorists has been to explain the capacity for belief-desire reasoning, especially how we attribute beliefs to one another (Spaulding 2018b).

To some, this framework has seemed somewhat limited. In particular, *pluralists* about folk psychology have argued that predicting and explaining behavior in terms of belief-desire attribution does not fully capture what we do when we come to understand one another (Andrews 2012; Fiebich et al. 2016; Spaulding 2018b). Instead, they propose that ordinary folk-psychological reasoning involves a wide range of concepts, inferential strategies, and cognitive processes besides belief-desire attribution, and that the scope of the latter is in fact quite restricted. Certain philosophers in the pluralist camp have also claimed that we often employ folk psychology in a *regulative* fashion—also called 'mindshaping'—in order to manipulate those around us to behave in certain predictable, norm-governed ways (McGeer 2007; Zawidzki 2013). The upshot of these proposals is that mainstream, mentalistic approaches to the cognitive underpinnings of social cognition are deeply misguided and require radical revision.

One folk-psychological strategy that has largely escaped the attention of mindreading theorists is the attribution of character or personality traits, such as 'mean', 'nice', 'intelligent', 'generous', 'extraverted', etc. (Andrews 2008; Fiebich and Coltheart 2015; Westra 2018). This form of folk-psychological reasoning is highly consequential for our everyday social interactions: it plays a role in how children learn from others (Lane et al. 2013), stereotyping (Fiske et al. 2002), moral judgment (Uhlmann et al. 2015), beliefs about personal identity (Strohminger and Nichols 2014), and even pragmatic inference (Pexman et al. 2005). There is also evidence that over time, third-party personality judgments (especially about close friends) reliably predict real-life outcomes (Connelly and Ones 2010; Luan et al. 2018; Vazire and Carlson 2011). All this suggests that trait attribution is not only an important part of our folk psychology, but a reliable and adaptive one as well. Thus, the absence of trait attribution from traditional debates about folk psychology is quite glaring.

In recent years, several proponents of folk-psychological pluralism have used the case of trait attribution to support their case, arguing that it is a form of folk-psychological reasoning that need not involve any kind of mental-state attribution at all (Andrews 2008; Fiebich and Coltheart 2015). Instead, they suggest that trait concepts enable us to predict and explain behavior by referring to behavioral dispositions alone. Trait attribution has also been cited as a form of regulative folk psychology that is used to shape people's behaviors so that they behave in normatively desirable ways (Mameli 2001). If these claims are right, then proponents of the mindreading-based approach to folk psychology might have cause to worry. Given its significance for everyday social cognition, the idea that trait attribution might be largely separate from mindreading would indicate that that approach leaves quite a bit out. Thus, trait attribution could turn out to be grist for the pluralist mill.

In response to these pluralistic proposals, I have argued elsewhere that trait attribution should actually be understood as *a part of mindreading* (Westra 2018). Far from being distinct from mindreading, character-trait attribution fits into the same folk-psychological theories that guide our attributions of beliefs and goals, and is part of the broader neural network responsible for mentalizing (for similar views, see also Meltzoff and Gopnik 2013; Reeder 2009; Tamir and Thornton 2018). According to this approach, traditional mindreading-based theories of folk psychology (in particular, the theory–theory) would only require some minor modifications in order to accommodate trait attribution.

This paper takes up this debate and maps out to what degree trait attribution in fact fits into a broadly pluralistic approach to folk psychology. In the first part of the paper, I defend this mentalistic theory of trait attribution, and argue that the evidence concerning the prediction and interpretation of behavior in terms of traits—especially in infancy and early childhood—does not provide support for the pluralist approach. In the second part of the paper, I consider the idea that trait attributions serve regulative, mindshaping functions above and beyond the prediction and explanation of behavior. After reviewing several plausible cases of trait-based mindshaping, I argue that reasoning about character traits does sometimes function as a form of regulative folk psychology, but that mindreading quite often plays a number of different roles in this process as well. Overall, I conclude that in order for us to understand how character-trait attribution fits into our overall folk psychology, we must appreciate its close connections with mental-state attribution.

2 Trait attribution and mindreading

Folk-psychological pluralists have generally used trait attribution as an illustration of how one might engage in folk-psychological reasoning without mentalizing. Most prominently, Andrews (2008, 2012) has proposed that trait concepts are used to refer to behavioral dispositions. On this view, believing that a person is extraverted involves the expectation that she will engage in a range of extraversion-relevant behaviors, such as approaching and spending time around large groups of people while smiling and talking; likewise, believing a person is shy means that she is disposed to avoid large groups of people and to speak very little when around such groups. Construed as behavioral dispositions, traits can support agent-specific behavioral predictions about what an agent will do in particular situations, even if the predictor has not made any inferences about the agent's underlying mental states. Personality traits can also be invoked to explain behavior when a person's reasons for acting are unknown or inscrutable, as when we attribute a person's erratic behavior to their nervousness (Malle 2006).[1] In short, trait reasoning offers a route to behavioral understanding that need not involve any mindreading at all.

According to the mentalistic approach to traits defended in this paper, personality traits are not merely construed as dispositions to *behave* in certain ways, but rather as dispositions to *form certain types of mental states*. To attribute to someone the trait of compassion, for instance, is to view that person as predisposed to form compassion-relevant mental states and emotions (e.g. desires to help people in need, to believe that

[1] See Korman and Malle (2016) for a recent study that calls this idea into question.

helping people in need is a moral duty, to feel sadness when confronted with suffering, etc.). Inferring that a person has a trait like compassion thus allows an observer to form expectations about their likely mental states in a given situation, and thereby better predict her behavior. Conversely, forming inductive generalizations about a person's mental states enables us to make deeper inferences about a person's stable character. In other words, traits are treated as an underlying psychological variable or factor that helps us to account for statistical regularities at the level of a person's mental states; attributing a particular trait to an individual helps us to generate prior probabilities for hypotheses about their likely mental states. Thus, trait representations help us to generate probabilistic models of how specific individuals' minds work, which enable us to make person-specific predictions about how they will think and act in different situations (see also Tamir and Thornton 2018).

In this model, trait-based reasoning is not a folk-psychological alternative to theory of mind: it is *a part of* our theory of mind, operating in conjunction with attributions of beliefs, desires, intentions, emotions, and other psychological states. Just as people posit unobservable mental states to explain and predict regularities at the level of behavior (Gopnik and Wellman 1992, 2012), so people also posit unobservable character traits to predict and explain higher order regularities at the level of mental states. Thus, trait reasoning is not a folk-psychological alternative to mindreading, but rather an extension of a person's more basic mindreading abilities.

A large body of neuroimaging data in neurotypical adult populations supports this mentalistic picture of how we reason about traits. The brain network known to support mentalizing [comprising the temporal-parietal junction, posterior superior temporal sulcus, medial prefrontal cortex, precuneus, and temporal poles (Van Overwalle 2009)] is heavily involved in trait reasoning, and in the updating of trait inferences in response to information about a person's beliefs and actions (Cloutier et al. 2011; Ferrari et al. 2016; Hassabis et al. 2014; Kestemont et al. 2013; Ma et al. 2011). For example, the tendency to explain behavior in terms of traits rather than situational factors [known as the "fundamental attribution error" or "correspondence bias" (Gawronski 2004; Gilbert et al. 1995; Jones and Harris 1967)] is reliably predicted by neural activity in regions associated with spontaneous mentalizing, suggesting that these judgments involve tacit mental-state attributions (Moran et al. 2014). Recent studies using multivoxel pattern analysis (or "neural decoding") techniques also suggest that the brain uses a common neural code to represent both stable character traits and transient mental-states (Thornton and Mitchell 2018). All of this suggests that, at the neural level, trait attribution is a part of a broader mentalizing system.

The intuitive relationship between trait and mental-state attribution is also evident in our explicit trait and mental-state attributions. For example, there is some evidence that the traits we attribute to a person are shaped by information about that person's mental states (Reeder 2009). Ames and colleagues found that participants were less likely to attribute the trait of helpfulness to an individual performing a helpful action (i.e. falling prey to the correspondence bias) if they also had information that that helpful action was performed *unwillingly* (Ames et al. 2004); a number of other authors have found that evidence of ulterior motives also attenuates the correspondence bias (Fein 1996; Krull et al. 2008). Encouraging people to reason deliberately about a target's psychological perspective has a similar effect (Hooper et al. 2015). People thus seem to intuitively understand that facts about a person's mental states are relevant when trying to infer their traits.[2]

Conversely, there is some evidence that we use inferred trait information when making judgments about a person's intentions. For instance, in their research on the psychological underpinnings of the "side-effect effect," Sripada and Konrath found that participants' judgments about whether or not a particular outcome was intentional was explained by a their inferences about the actor's character or "deep self", (Sripada 2012; Sripada and Konrath 2011). Knowing about a person's character, in other words, provides us with evidence about their likely intentions**.** Overall, in our elicited trait judgments, intention attributions, and in the neural underpinnings of trait and mental-state reasoning, there is compelling evidence that we understand traits to be mentalistic dispositions.

2.1 Non-mentalistic trait attribution in autism spectrum disorder?

In response to these data, a pluralist could concede that we *sometimes* think of traits as mentalistic dispositions, while still maintaining that trait attributions are also non-mentalistic in a significant range of cases. Along these lines, Andrews cites Carol Gray's "Social Stories Therapy" (Gray 2007), an intervention for children with autism spectrum disorder (ASD) that trains them to associate trait terms (e.g. "happy") with behavioral patterns (e.g. smiling and laughing). Since ASD populations are known to suffer from mindreading deficits (Baron-Cohen et al. 1985), and the training procedure does not involve any references to mental states, this is plausibly an instance of trait attribution in the absence of mindreading. Such cases seem to show that trait attribution without mindreading is at least possible, even if two processes sometimes co-occur in practice.

However, there are a few issues with the particulars of this case. First, it is not obvious that this is really an instance of trait attribution, as opposed to emotion recognition (Westra 2018). Second, even if we grant that it does involve trait attribution, several meta-analyses have shown that the effectiveness of Social Stories interventions are questionable at best, which casts doubt on the idea that people who undergo this therapy actually use it to predict and interpret behavior (Ali and Frederickson 2006; Reynhout and Carter 2011). Third, using ASD populations as a way to control for the role of mentalizing in some other socio-cognitive process is itself problematic: although mentalizing is difficult for people with ASD and follows an atypical developmental trajectory, many individuals with ASD are in fact able to reason about mental states with varying degrees of proficiency (Back et al. 2007; Mitchell 2013; Parsons and Mitchell 1999); thus, the presence of trait attribution in ASD would not itself constitute evidence for the pluralist account. There is therefore little reason to think that the Social Stories Therapy case reveals an instance of non-mentalistic trait attribution.

[2] In Westra (2018), I also suggested that the correspondence bias functions as a kind of mindreading heuristic: rapidly attributing traits to a person upon first encountering helps us derive an initial set of priors for the kinds of mental states they might have.

Finally, even if it turned out that individuals with ASD do in fact engage in non-mentalistic forms of trait attribution, this would only license pluralism about trait-understanding in the special case of ASD; it would tell us nothing about how trait attribution functions in neurotypical populations. But if pluralism is to offer a genuine alternative to the mindreading-based approach to trait reasoning, then it needs to offer a theory of everyday, neurotypical trait attribution. And if the case of trait attribution is to provide actual empirical support for the broader pluralist framework, then the pluralist must show that trait attribution and mindreading do in fact diverge in a substantial range of ordinary cases.

2.2 Trait-like reasoning *before* mindreading?

A more compelling strategy for the pluralists to establish a divergence between mindreading and character-trait attribution would be to show that children are able to reason about traits *before* they are able to reason about mental states. If trait attribution turned out to be developmentally prior to mindreading, it would show that trait attribution in no way depends upon a capacity for mindreading, even if the two folk-psychological strategies ultimately become integrated in adulthood. But does this evidence exist?

The case from the existing pluralist literature that comes closest to showing this comes from Fiebich and Coltheart's (2015) notion of "non-linguistic associations$_{\text{pi}}$". Non-linguistic associations$_{\text{pi}}$ are a non-mentalistic strategy for keeping track of person-specific properties like traits without explicitly representing them by relying upon prior associations between particular *agents*, *behaviors*, and *specific contexts* (e.g. implicitly tracking the fact that Greg is generous by associating *Greg* with *tipping* in *restaurants*). On Fiebich and Coltheart's account, this ability to form expectations based on person-specific dispositions emerges early in ontogeny, but persists into adulthood. Notably, it does require some proto-mentalistic abilities, such as the capacity to recognize intentional agents and "a few core teleological principles" of intentional action (Fiebich and Coltheart 2015, p. 242), albeit not the capacity to reason with genuine mental state concepts.[3]

Fiebich and Coltheart's (2015) primary example of associations$_{\text{pi}}$ early in development comes from studies about infants' expectations about agents' goal-directed reaching behaviors (Luo and Johnson 2009; Woodward 1998). In these studies, infants are first shown a series of familiarization trials in which an agent repeatedly reaches for one of two toys; then, in the test trial, they are shown the agent either reaching for the same toy (an expected outcome) or a different toy (an unexpected outcome). In general, these tasks show that infants are surprised (i.e. have greater looking times) when they see the unexpected outcome than when they see the expected one, suggesting that they have registered that this particular agent is stably disposed to pursue a particular goal object. [Notably, infants do not look longer at this unexpected outcome

[3] Non-linguistic associations$_{\text{pi}}$, are distinct from *linguistic* associations$_{\text{pi}}$, which emerge later in development and require the possession of an explicit, lexicalized trait concept (e.g. the word "generosity"). Unlike their non-linguistic counterparts, they are said to permit more flexible behavioral predictions, since they allow that behavioral and contextual information can be associated with trait words, as well as specific individuals.

when the agent's view of its goal object is obstructed, suggesting that behavioral predictions based on associations$_{\text{pi}}$ are sensitive to facts about perceptual access (Luo and Johnson 2009)].

Although Fiebich and Coltheart (2015) cite these cases as examples of associations$_{\text{pi}}$ and discuss them in the context of reasoning about character traits (Fiebich and Coltheart 2015, pp. 242–244), they are not really about trait attribution as we normally think of it. Though they do show that infants making person-specific inferences about dispositions of some kind or other, these dispositions are quite narrow, and are always directed towards a particular object in a particular kind of choice context. Personality or character traits, in contrast, are typically construed as stable dispositions that apply broadly, across different contexts (c.f. Doris 2002). At best, these studies show that infants are able to track person-specific *preferences* within a given situation, but they say nothing about their ability to track person-specific traits (mentalistic or otherwise).

A better set of cases for our purposes comes from a series of studies investigating trait-like understanding in 15-month-olds by Repacholi et al. (2008, 2016a, b). These studies all employed different versions of an "emotional eavesdropping paradigm" (Repacholi and Meltzoff 2007). In the original version of this paradigm, infants first observe a series of interactions between two agents, an "Experimenter" and an "Emoter." Across several trials, the Experimenter pulls out and manipulates a series of different toys. As the Experimenter does this, the Emoter enters and displays either a neutral reaction or an angry reaction to the Experimenter's demonstration. After three repeated trials, the Experimenter removes a fourth toy, demonstrates an action on it, then hands it to the infant to play with while the Emoter looks on with a neutral expression. The basic finding from this paradigm is that infants are slower to play with these objects and less likely to imitate the actions of the Experimenter in the Anger conditions than in the Neutral conditions, as if anticipating that the Emoter might become angry with them as well. Notably, this effect was sensitive to whether or not the Emoter could see the infant while she plays with the toy (Repacholi et al. 2008, 2016a, b). Repacholi and colleagues also found that in a subsequent trial in which the Emoter requests a toy from the infant, infants are faster to give it up if they have just witnessed the Anger conditions—a finding the experimenters interpret as a form of social appeasement (Repacholi et al. 2016a, b).

Like in the preference-understanding studies cited by Fiebich and Coltheart, infants in these experiments seem to be tracking agents' stable dispositions across trials—in this case, displaying a particular kind of emotional response rather than reaching—in a way that is sensitive to the agent's perceptual access. But unlike the preference-understanding studies, infants display the ability to form these expectations even when the context shifts from one in which the infant is observing two third-parties interacting to one in which she is being observed by two third parties, and to one in which she is directly engaging with the agent. Thus, these infants seem to have inferred not only "A gets mad at B when A sees B play with a toy," but also "A will also get mad at me when A sees me play with a toy (so I had better be careful)" and "A will get mad at me if I don't give her the toy when she asks for it (so I had better give it to her quickly)." Thus, infants' social understanding in these studies appears to be relatively broader and more trait-like than the cases of non-linguistic associations$_{\text{pi}}$ that Fiebich and Coltheart cite.[4]

Our central question, however, is whether or not these studies show that infants' trait understanding is mentalistic or non-mentalistic. This is not a simple question to answer: as with many studies involving preverbal infants and non-verbal animals, it is possible to give both cognitively rich and sparse interpretations of the findings (Carruthers 2013; Heyes 2014a; Penn and Povinelli 2007). Thus, from a mentalistic standpoint, one could argue that these studies show that infants are able to form trait-like generalizations over agents' goals and over their emotional states, that they are able to integrate those generalizations with their knowledge of perception to generate different predictions across contexts, and that they are able respond appropriately on the basis of these predictions. On the other hand, one could potentially generate any number of associative, behavior-reading alternative explanations that only appeal to infants' expectations about regularities in surface behavior. However, the fact that infants' expectations in these studies seem to be sensitive to facts about perception, generalizable to new contexts, and variable in their behavioral consequences (e.g. recoiling from the toy versus quickly giving it up) suggests that their reasoning about the Emoter is not closely tied to surface features of her behavior, but rather tracking multiple underlying variables. This would seem to rule out the simplest kinds of behavior-reading explanations of these results and lend support to a mentalistic interpretation of these findings, and thus cut against the claim that trait attribution are dissociated in infancy.[5]

These particular findings from Repacholi and her colleagues represent just one body of evidence that pluralists might use to make the case that trait attribution is developmentally prior to mentalizing; it is certainly possible that they might be able to make this case in another way. But they illustrate a broader problem for this argumentative strategy: it is very difficult to say whether these early socio-cognitive abilities represent genuine forms of trait understanding, and whether or not they involve genuine mindreading. This makes it difficult for a pluralist to convincingly show that the two abilities are developmentally dissociated, if only because it is not at all clear how to interpret the evidence in question.

[4] It is not clear from this experiment just how broad these representations are, or how many other situations they would generalize to. It may be that children's trait-like representations are at this stage still relatively "local" and limited to a small range of situations, and only become more "global" as children make more social observations and acquire trait language. Or it may be that these initial trait-like concepts actually pick out very broad evaluative categories (e.g. "good guy" and "bad guy"), and only become more refined over time.

[5] Another possible interpretation is that infants in studies like these are merely relying upon an "implicit" or "minimal" theory of mind, and that their performance does not require "full-blown" forms of mindreading (Apperly and Butterfill 2009; Butterfill and Apperly 2013). This appears to be what Fiebich and Coltheart (2015) think is going on in preference attribution studies. However, the two-systems theory is itself quite controversial, and the evidence and theoretical motivations for it has come under significant criticism from a number of authors (Carruthers 2015; Christensen and Michael 2015; Heyes 2014b; Michael and Christensen 2016; Santiesteban et al. 2014; Westra 2017). Given these concerns, I will leave a two-systems interpretation of these findings for another venue.

2.3 Trait understanding in early childhood is mentalistic

When we set the infancy data aside and consider instead the evidence on children's explicit, verbal understanding of traits, the case for dissociation becomes weaker still, and we actually find meaningful support for the mentalistic approach. Around the same time that children start to demonstrate an explicit understanding of traits and their effects on behavior, they also understand that traits are interconnected with people's psychological states (Heyman 2009).

Unfortunately, the developmental timeline for explicit trait attribution has been the source of some confusion. For some time, the consensus in the literature on trait attribution seemed to be that it emerges relatively late, well after children develop an explicit understanding of concepts like belief, desire, and knowledge. A number of early studies on the topic seemed to show that children do not display any understanding of traits in their explicit predictions until they are around 7 years old (Kalish 2002; Rholes and Ruble 1984), nor mention traits in their explanations of behavior until even later (Peevers and Secord 1973). If this timeline were right, it would actually show that the development of trait understanding and mental-state reasoning are somewhat discontinuous, which might be problematic for the mentalistic approach.

However, the methods employed in these studies set the bar for trait-understanding far too high. Peevers and Secord, for instance, employed a free description methodology, asking children to describe their friends in their own words and coding these descriptions for references to traits (Peevers and Secord 1973). But understanding traits and being disposed to spontaneously invoke them when describing people are two different things. Other studies employed a behavioral prediction methodology, first showing children vignettes describing characters engaging in a certain behavior one or more times, and then asking them whether the character would perform a similar action again (Kalish 2002; Rholes and Ruble 1984). As Liu and colleagues point out, however, these studies never directly ask children about traits; rather, they simply ask children to make explicit predictions about future behavior based on a limited sample of past behavior, and assume that predictions about behavioral consistency reflect trait attributions (Liu et al. 2007). The fact that younger children are unwilling to do this does not necessarily reflect a lack of trait understanding. It could mean that children are unsure whether a particular trait attribution is justified, or whether they have enough evidence to predict behavioral consistency; indeed, some studies using this methodology have had success with younger children simply by increasing the number of behavioral examples provided (Aloise 1993). Moreover, in studies that explicitly ask children to map behaviors onto trait labels (e.g. "mean," "nice", "shy"), or to predict the behavior of a character who they have been told possesses a particular trait, they are able to do so by at least 4 years of age (Heyman and Gelman 1998, 1999; Liu et al. 2007), with a few studies showing competence by age 3 (Heyman and Gelman 2000; Yuill and Pearson 1998). Thus, the capacity to reason about traits in an explicit manner emerges right around the same time that children are also developing the ability to reason explicitly about desires, knowledge, and belief (Wellman and Liu 2004).

Even more importantly, it is clear that children's understanding of traits at this age is distinctly psychological in character. For example, Heyman and Gelman found that three- and four-year-olds are both more likely to attribute a similar preference to a new character if she shared a trait label with a familiar one than if she had a similar outward appearance (Heyman and Gelman 2000). In another study, the same authors also found that four-year-olds could successfully predict the emotions and desires of a character labeled as "nice", "mean", and "shy" or "not shy" in different situations (Heyman and Gelman 1999). Yuill and Pearson (1998) also found that for children around this age, correctly identifying the desires and emotions of a person with a particular trait was correlated with their understanding of the subjectivity of desires. Thus, between the ages of three and five, children not only display an explicit understanding of character traits but also an understanding of how stable traits translate into emotions and desires in different contexts (for a review, see Heyman 2009).

2.4 Taking stock

In this section, I first surveyed some of the evidence supporting a mentalistic approach to trait-understanding in adults, and then considered two strategies for showing that mindreading and trait attribution are dissociable: arguments from ASD and arguments from development. But neither the Social Stories Therapy case nor the case of trait-like reasoning in infancy provide clear evidence for a dissociation between trait attribution and mindreading; the latter might even support opposite conclusion. The developmental case grows weaker still when we consider slightly older children's understanding of traits, which is clearly mentalistic. As it stands, the weight of the evidence favors the mentalistic approach, and offers little support for pluralism.

In response to this conclusion, a proponent of the pluralistic approach to trait attribution might argue that none of the evidence discussed in this section ought to be viewed as undermining their core thesis. After all, none of the evidence in this section has shown that trait attribution is *necessarily* mentalistic—merely that trait attribution and mindreading happen to co-occur some of the time. For all this evidence, it might be the case that trait attribution and mentalizing are still dissociable in principle, even if this rarely occurs in practice. In other words, if we were to interpret the pluralist as defending the weaker claim that non-mentalistic trait attribution is *possible*, then the fact that trait attribution is sometimes or even often mentalistic would not cause them any embarrassment.

However, framing pluralism as a claim about which kinds of folk-psychological reasoning are possible would actually undermine its status as an empirical framework. It may well be that non-mentalistic forms of trait attribution are metaphysically or logically possible, but the truth of these claims would be of little interest to cognitive scientists, who are concerned with generalizations about how social cognition actually works. If we interpret it as a robust positive thesis about every folk psychology and ordinary trait attribution, then pluralism might well have something to add to this endeavor. If pluralism were instead interpreted as a weak modal thesis, however, then it would amount to little more than a conceptual claim about the nature of folk psychology, and offer little insight into the underpinnings of everyday social cognition. As such, defenders of pluralism should be wary of adopting this line of argument.

Moreover, even if the mentalistic theory of trait attribution were true, it would not show that the general thesis of folk-psychological pluralism is false, even when we interpret it as a theory of actual social cognition; that never rested on the case of trait attribution alone. For everything that has been argued about trait attribution here, the pluralists might be absolutely right that folk psychology involves many non-mentalistic processes, and that the global scope of mindreading in everyday social cognition has been greatly exaggerated. But when it comes to giving an adequate account of the role of character-trait attribution in our folk psychology, we had best pay close attention to its relationship with mentalizing.

3 Trait attribution, mindreading, and mindshaping

I now turn to a different pluralistic proposal about trait attribution that is more promising: the idea that trait attribution also has *regulative* or *mindshaping* functions (Mameli 2001; McGeer 2007; Zawidzki 2008, 2013). In contrast to mindreading approaches to folk psychology, which emphasize the activities of predicting, interpreting, and explaining behavior, the regulative view of folk psychology proposes that another way in which we come to understand others is by engaging in practices that cause both ourselves and others to behave in a norm-governed manner. When distributed throughout the population, mindshaping practices cause people to be intrinsically motivated to conform to social norms, and also deter people from deviating from those norms. And once one knows the local social norms, this makes behavior a lot easier to predict, because everyone around you is in the business of making themselves and each other maximally predictable and understandable. Some examples of third-personal mindshaping include moral praise, blame, pedagogy and teaching, and storytelling; first-personal examples of mindshaping include imitative learning and making promises and other forms of commitments.

While mindshaping is often presented as a departure from traditional mentalistic approaches to folk psychology (McGeer 2007; Zawidzki 2013), several commentators have noted that mindshaping often seems to involve (or perhaps even require) substantial amounts of mindreading (Michael 2014; Peters 2019).[6] In order to engage in effective teaching, for example, it helps to monitor the *understanding* and *knowledge* of one's pupil. When we blame a person for their actions, it's often because we take them to hold blameworthy *intentions* or *ill will*.[7] More generally, when seeking to effect a change in a person's mental states that will cause them to behave in a predictable way,

[6] Notably, Zawidzki's account of mindshaping is committed to a version of the two-systems theory of mindreading (Apperly and Butterfill 2009; Butterfill and Apperly 2013), and thus distinguishes between "implicit" and "explicit" forms of mental-state attribution. His view thus acknowledges the role of "implicit" forms of mindreading in mindshaping, which involve relational, non-propositional, proto-mentalistic concepts. However, Zawidzki would deny that mindshaping requires explicit mindreading (i.e. predicting and explaining behavior in terms of "full-blown" propositional attitudes with linguistically specifiable contents) (Zawidzki 2011, 2013).

[7] This is by no means a universal feature of moral judgment: the relevance of mental states like intentions in judgments of blameworthiness has been shown to vary across cultures (McNamara et al. 2019).

it helps to know what that person's mental states are, and to be able to predict how their minds might change in response to different actions. In terms of one's theory of social cognition, this means that focusing on the primacy of mindshaping at the expense of mindreading would be misguided. The two processes are best understood as complementary, working in tandem with one another (Peters 2019).

In this section, I show how this mindreading-friendly approach to mindshaping can be fruitfully applied to character-trait attribution. In particular, I focus on the way that mindshaping and mindreading interact in our reasoning about *moral* character in three types of cases: virtue-labeling, gossip about moral character, and moral pedagogy through virtuous exemplars. Moral character judgment is a very obvious candidate for regulative trait attribution, since it has very clear normative functions (Goodwin et al. 2014; Uhlmann et al. 2015). But even when character-trait attribution plays this highly normative role, it often involves a variety of different forms of mentalizing as well. Highlighting this interaction further supports the claim that the folk-psychological role of character-trait attribution is deeply intertwined with mindreading, even when its function appears to be regulative rather than predictive.

3.1 Virtue-labeling

One classic example of how moral trait attributions can normatively regulate behavior (and indeed, a classic example of mindshaping more generally) comes from a study by Richard Miller and colleagues (Mameli 2001; Miller et al. 1975; Peters 2019). This study aimed to determine whether students would be less likely to litter relative to a control group when they were exposed to reasoned arguments (the persuasion condition), or when they were exposed to false praise for being ecologically conscious (the attribution condition). In the persuasion condition, students were subject to a series of lectures and admonishments by the teacher, principal, and janitor about the importance of being clean. In the attribution condition, students were instead praised for how clean and ecologically conscious and conscientious they were. The study found that in the immediate post-test, students in the attribution condition were more likely to appropriately dispose of their litter than the students in the persuasion condition, but that both groups were tidier relative to a control group. However, in the later post-test, only students in the attribution condition remained tidier than the control group, suggesting that mere trait attributions were an effective way of motivating normatively desirable behavior.

This "virtue-labeling" phenomenon has since been documented in a number of studies since Miller's (for a review, see Alfano 2013, pp. 88–91). What makes it a particularly compelling case for the mindshaping theorist is that the attribution itself does not need to be grounded in any kind of belief about the actual traits of the target; indeed, children in the attribution condition were *not* initially more ecologically conscious than their peers in the other conditions. The attribution instead functioned as a kind of self-fulfilling prophecy[8] by imbuing the children with a sense of how they

[8] It is worth noting, however, that most virtue theorists would deny that the dispositions created by virtue-labeling count as genuine virtues on account of the fact that they lack the appropriate motivational structure; this is why Alfano calls the results of virtue labeling "factitious virtues" (Alfano 2013; Miller 2017).

ought to behave (Mameli 2001). In short, it was not the psychological content of the attribution that achieved the desired effect, but rather its normative force.

In line with the mindreading-friendly view of mindshaping, virtue-labeling also seems to require a fair amount of mentalizing in order for it to be effective. Consider: either the mindshaper believes that the trait attribution is true, or she believes that it is false. If she believes that it is true, then the trait attribution is a straightforward case of mindreading that also happens to have mindshaping effects (c.f. Peters 2019). If she believes that it is false, then she is lying. The basic capacity to lie is closely linked with theory-of-mind development, as is the ability to lie effectively (Ding et al. 2015; Talwar and Lee 2008; Williams et al. 2016), and it is widely believed that mentalizing is involved in deception in adults as well (for a review, see Spence et al. 2004). And so, in cases of both sincere and insincere virtue-labelling, some form of mindreading is likely involved on the part of the would-be mindshaper.[9] Something similar may be true of the mindshaping target: Peters (2019) has noted that virtue attributions are unlikely to have the desired behavioral effects unless the target attributes either belief or knowledge to the attributor as well, or else they would not believe that the trait attribution is genuine.[10]

In short, virtue-labelling may very well require *both* the attributor and the target to engage in mindreading in order to effectively promote normative behavior. However, as the Miller et al. case illustrates, the trait attribution *itself* need not be an act of descriptive mindreading. Instead, other forms of mindreading play an auxiliary role, scaffolding the mindshaping function of the trait attribution in order to bring about the desired regulative effect.

3.2 Gossip about traits

Another way that trait attribution can function in a regulative manner is through prosocial gossip. Though gossip is often thought of as a kind of vice, in social groups it allows for the sharing of reputational information about conspecifics' cooperative (or uncooperative) behavior (Dunbar 2004). This promotes prosocial behavior in two ways. First, access to reputational information helps individuals avoid cooperating with prospective social partners who might take advantage of them, and instead choose more

[9] One of the reviewers of this paper suggested that a mindshaper in such a situation might knowingly make a false virtue-attribution in order to bring about a particular behavior in the target without ever considering the fact that this behavior is caused by a false belief. In this case, the mindshaper would achieve their desired behavioral end without any sort of mindreading. This possibility is interesting, but also highly speculative. The broad consensus in the literature on lying and deception is that it involves mindreading.

[10] In reply to this point, a proponent of a strong, anti-mentalistic version of the mindshaping hypothesis could argue that participants in virtue-labelling tasks do not attribute any mental states to the attributor, and instead simply accept the assertions of certain informants by default, especially if the informant is an authority figure like the teacher in Miller et al. (1975). There is something to this objection: generally, the literature on trust in testimony in young children shows that they are strongly inclined to believe what they're told (Jaswal et al. 2010, 2014). However, children as young as three are also less trusting when informants seem less confident in their assertions (Jaswal and Malone 2007), and by four their trust in testimony is moderated by information about a speaker's ignorance and prior reliability (Kushnir and Koenig 2017). The fact that children's trust in testimony is moderated by evidence of ignorance suggests that their default disposition is to tacitly attribute *knowledge* to informants, and to accept their testimony on that basis unless given reason to think that they are not in fact knowledgeable.

reliable cooperative partners (Feinberg et al. 2014; Sommerfeld et al. 2007). Second, in a population where reputation-sharing is common, the payoffs of free-riding are significantly diminished, since a single defection can lead to ostracism, and foreclose access to future cooperative endeavors (Feinberg et al. 2012; Wu et al. 2016). Collectively, these two forces create a strong selection pressure for prosociality and fairness through the mechanism of partner choice (Baumard et al. 2013; Sperber and Baumard 2012).

Trait attributions are an efficient way to communicate past reputational information, because they support broad generalizations about the person in question, and thus serve as a useful guide for inferring their future dispositions towards cooperative behavior. And indeed, empirically documented cases of prosocial gossip do tend to involve a good number of trait attributions, particularly in behavioral economics studies that allow participants to "gossip" (i.e. send notes) to other players about conspecifics' behavior in trust games and public goods games. For example, Feinberg et al. (2012) report that when individuals chose to gossip after observing defections in a trust game, they typically wrote things like "Try not to give too much to Participant B. He/she's really *selfish*," and "Participant B is extremely *greedy*; send 0 points[emphasis added]" (Feinberg et al. 2012, p. 1021). Similarly, gossip notes sent in Wu et al. (2016)'s public goods game frequently featured trait attributions such as "generous," "stingy," "miser," "unfair," and so on (Wu et al. 2016, pp. SI 12–25).[11] In both cases, the presence of this kind of gossip increased overall levels of prosocial behavior.

In this case, the regulative effects of trait-based gossip arguably occur *as the result of* familiar, descriptive forms of mindreading. After all, trait attributions communicate descriptive psychological information about conspecifics that supports predictions and explanations of future behavior. At the population level, this creates selection pressures that ultimately shift behavioral norms towards increased prosociality and fairness. Thus, in this case the mindshaping effects of trait attribution are not merely scaffolded by mindreading processes, but rather an indirect effect of them.

However, it is also quite plausible that trait-based gossip also succeeds in shaping minds via a more direct, non-mentalistic route. Because gossiped-about traits are also highly moralized, they may also serve to promulgate normative standards for behavior. For example, if you hear your peers whisper in hushed tones that a person who tipped a server 15% on a meal is "stingy", this tells you something about that person's psychology, but also something about the local tipping norms. While the psychological aspect of this gossip might lead you to mistrust or avoid the allegedly stingy tipper, the normative aspect might lead you to adjust your own tipping behavior in the future (and perhaps even gossip about observed instances of sub-15% tipping yourself). Thus, trait-based gossip can also help us learn about social norms, and thereby promote norm-governed behavior in a manner that does not obviously rely upon mindreading.

[11] Participants in this task also often used more colorful and profane epithets to describe non-cooperators in their gossip notes. Arguably, this amounts to a form of trait attribution as well: calling someone a "jerk", for example, involves making a claim about the sort of person they are, rather than just a description of their behaviors (Schwitzgebel 2019).

3.3 Moral pedagogy

One final, more speculative way in which trait attribution might support mindshaping is through moral pedagogy.[12] Often, remarking upon a person's moral virtues or vices serves a didactic function, communicating to one's audience that a certain trait is praiseworthy or blameworthy. This is especially obvious when we are highlighting the attributes of real moral exemplars, such Ghandi's temperance, Mother Theresa's compassion, or the humility of Pope Francis (Zagzebski 2017). But perhaps the most recognizable forms of pedagogical virtue and vice attributions are to be found in fiction and storytelling. Take, for instance, a film like *The Wizard of Oz*. The Scarecrow, the Tin Man, and the Lion all go through character arcs in which they are initially portrayed as lacking certain virtues (wisdom, compassion, and courage). Then, throughout the film their true characters are revealed through their actions, and eventually they are told by the wizard that they possessed these virtues all along. Another familiar example is Charles Dickens' *A Christmas Carol,* in which the main character Scrooge undergoes a paranormal journey that transforms him from a selfish miser into a generous, compassionate person. For a young child, this kind of parable serves as a means for imparting virtue and vice concepts and modeling virtuous conduct through simple, highly familiar characters (see also Cain et al. 1997, Study 1).

Moral pedagogy through portrayals of character also emerges in more complex narratives, and can facilitate moral learning in mature adults as well as children. One example of this comes from a narrative structure that Noël Carroll refers to as a "virtue wheel" (Carroll 2002). A virtue wheel consists in a juxtaposition of several characters that vary along different moral trait dimensions, such that the deficiencies in virtue in one character serve to highlight the virtue of another. By way of example, Carroll discusses how in the film *Howard's End*, the true virtue of imaginativeness is highlighted through a contrast between three siblings: Margaret, Helen, and Tibby Schlegal. Although all three characters are imaginative, Helen is portrayed as imaginative to a pathological degree, obsessed as she is by her empathy for others. At the other extreme, Tibby is imaginative in a scholarly, aesthetic way, but so engrossed in his intellectual pursuits that he fails to direct his imagination towards the needs of others. Only Margaret's imaginativeness exemplifies real virtue, because it balances both the ability to see good in others and a concern for practicality. Carroll argues that virtue wheels like this one can facilitate moral learning by helping readers to recognize virtues and vices in others, and prompting them to engage in conceptual analysis about what sorts of dispositions virtue requires.

One possible way that such narratives could result in mindshaping would be if they caused people to imitate the behaviors of the virtuous exemplars in question. This form of mindshaping would not necessarily require mentalizing: simply recognizing that a certain kind of behavior is normatively desirable would be sufficient to achieve the relevant effects. But from a normative, virtue-theoretical perspective, this kind of learning would be quite superficial. This is due to the mentalistic nature of the virtues and vices: like other traits, they are at their core psychological dispositions to have the

[12] For discussion of the narrative, normative role of propositional attitude attributions, see Hutto (2009) and Zawidzki (2013).

right kinds of values, to feel certain kinds of sentiments, and to respond appropriately to reasons (Miller 2013; Snow 2010). This means that in order for a pupil to *actually* learn from exemplars or fictional characters about the true nature of virtue, she could not merely imitate the behavior of these individuals, as this would tell her nothing about the underlying attitudes that make their behavior virtuous (as opposed to merely continent). She must therefore be able to appreciate the psychological factors that motivate the exemplar's actions. Thus, in order for this kind of exemplarist moral pedagogy to convey actual moral knowledge, learners must use their theory of mind.

3.4 Taking stock

Collectively, these cases illustrate how the folk-psychological function of trait attribution extends beyond the prediction and interpretation of behavior by actively promoting conformity to normative standards. But they also show that this regulative form of trait attribution often involves a variety of different mindreading processes. In the case of virtue labelling, the normative effects of trait attributions are scaffolded by mindreading on the part of both the mindshaper and her target. In the case of prosocial gossip, trait attributions promote norm-governed behavior as an indirect effect of communicated psychological information, and as a direct effect of communicated normative information. In moral pedagogy, mindreading itself plays a normative role: while a certain amount of mindshaping can occur simply through the imitation of virtuous behavior, genuine moral learning requires a pupil to engage in more sophisticated forms of mental-state attribution.

4 Conclusion

In this paper, I have explored how and where character-trait attributions might fit into a broader debate about the nature of everyday folk psychology. I first considered several arguments for the claim that trait attribution might constitute a distinctly non-mentalistic mode of folk-psychological reasoning, and found them to be lacking. I then turned to the idea that trait attributions serve regulative rather than descriptive functions. I discussed several ways in which character-trait attributions might occur, and for each one noted the roles played by mindreading.

The overarching point of this discussion has been that in order to understand the role of character-trait attribution in our folk psychology, we need to recognize that reasoning about character is closely integrated with traditional forms of mindreading. This is not to deny that character-trait attribution in some ways goes beyond the traditional mindreading-based framework, or that it can have significant regulative effects. Indeed, the mindshaping approach can help to illuminate the distinctive ways that character-trait attributions promote norm-governed behavior. But even in the places where our understanding of trait attribution requires us to go beyond the traditional mentalistic framework, mindreading still ends up being a crucial part of the story.

Acknowledgements This study was funded by Social Sciences and Humanities Research Council Postdoctoral Fellowship 756-2018-0012.

References

Alfano, M. (2013). *Character as moral fiction*. Cambridge: Cambridge University Press.

Ali, S., & Frederickson, N. (2006). Investigating the evidence base of social stories. *Educational Psychology in Practice,22*(4), 355–377.

Aloise, P. A. (1993). Trait confirmation and disconfirmation: The development of attribution biases. *Journal of Experimental Child Psychology,55*(2), 177–193.

Ames, D. R., Flynn, F. J., & Weber, E. U. (2004). It's the thought that counts: on perceiving how helpers decide to lend a hand. *Personality and Social Psychology Bulletin,30*(4), 461–474.

Andrews, K. (2008). It's in your nature: A pluralistic folk psychology. *Synthese,165*(1), 13–29.

Andrews, K. (2012). *Do apes read minds?: Toward a new folk psychology*. Cambridge, MA: MIT Press.

Apperly, I., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review,116*(4), 953–970.

Back, E., Ropar, D., & Mitchell, P. (2007). Do the eyes have it? Inferring mental states from animated faces in autism. *Child Development,78*(2), 397–411.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition,21*(1), 37–46.

Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences,36*(1), 59–78.

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

Cain, K. M., Heyman, G. D., & Walker, M. E. (1997). Preschoolers' ability to make dispositional predictions within and across domains. *Social Development,6*(1), 53–75.

Carroll, N. (2002). The wheel of virtue: Art, literature, and moral knowledge. *The Journal of Aesthetics and Art Criticism,60*(1), 3–26.

Carruthers, P. (2013). Mindreading in infancy. *Mind and Language,28*(2), 141–172.

Carruthers, P. (2015). Mindreading in adults: Evaluating two-systems views. *Synthese,192*, 1–16.

Carruthers, P., & Smith, P. K. (1996). *Theories of theories of mind*. Cambridge: Cambridge University Press.

Christensen, W., & Michael, J. (2015). From two systems to a multi-systems architecture for mindreading. *New Ideas in Psychology,40*, 48–64.

Cloutier, J., Gabrieli, J. D. E., O'Young, D., & Ambady, N. (2011). An fMRI study of violations of social expectations: When people are not who we expect them to be. *NeuroImage,57*(2), 583–588.

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin,136*(6), 1092–1122.

Ding, X. P., Wellman, H. M., Wang, Y., Fu, G., & Lee, K. (2015). Theory-of-mind training causes honest young children to lie. *Psychological Science,26*(11), 1812–1821.

Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.

Dunbar, R. I. M. (2004). Gossip in evolutionary perspective. *Review of General Psychology,8*(2), 100–110.

Fein, S. (1996). Effects of suspicion on attributional thinking and the correspondence bias. *Journal of Personality and Social Psychology,70*(6), 1164.

Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological Science,25*(3), 656–664.

Feinberg, M., Willer, R., Stellar, J., & Keltner, D. (2012). The virtues of gossip: Reputational information sharing as prosocial behavior. *Journal of Personality and Social Psychology,102*(5), 1015–1030.

Ferrari, C., Vecchi, T., Todorov, A., & Cattaneo, Z. (2016). Interfering with activity in the dorsomedial prefrontal cortex via TMS affects social impressions updating. *Cognitive, Affective, and Behavioral Neuroscience,16*(4), 626–634.

Fiebich, A., & Coltheart, M. (2015). Various ways to understand other minds: Towards a pluralistic approach to the explanation of social understanding. *Mind and Language,30*(3), 235–258.

Fiebich, A., Gallagher, S., & Hutto, D. D. (2016). Pluralism, interaction, and the ontogeny of social cognition. In *Routledge handbook on the philosophy of the social mind* (pp. 1–21).

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2002). A model of (often mixed stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychologyersonality and Social Psychology,82*(6), 878–902.

Gawronski, B. (2004). Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. *European Review of Social Psychology,15*(1), 183–217.

Gilbert, D. T., Malone, P. S., Aronson, J., Giesler, B., Higgins, T., Ross, L., et al. (1995). The correspondence bias. *Psychological Bulletin,117*(1), 21–38.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology,106*(1), 148–168.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind and Language,7*(1–2), 145–171.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin,138*(6), 1085–1108.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind and Language,1*(2), 158–171.

Gray, C. (2007). *Writing social stories with Carol Gray*. Arlington, TX: Future Horizons.

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2014). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex,24*(8), 1979–1987.

Heyes, C. (2014a). False belief in infancy: A fresh look. *Developmental Science,17*(5), 647–659.

Heyes, C. (2014b). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science,9*(2), 131–143.

Heyman, G. D. (2009). Children's reasoning about traits. *Advances in Child Development and Behavior,37*, 105–143.

Heyman, G. D., & Gelman, S. A. (1998). Young children use motive information to make trait inferences. *Developmental Psychology,34*(2), 310–321.

Heyman, G. D., & Gelman, S. A. (1999). The use of trait labels in making psychological inferences. *Child Development,70*(3), 604–619.

Heyman, G. D., & Gelman, S. A. (2000). Preschool children's use of trait labels to make inductive inferences. *Journal of Experimental Child Psychology,77*(1), 1–19.

Hooper, N., Ergogan, A., Keen, G., Lawton, K., & McHugh, L. (2015). Perspective taking reduces the fundamental attribution error.pdf. *Journal of Contextual Behavioral Science,4*, 69–72.

Hutto, D. D. (2009). Folk psychology as narrative practice. *Journal of Consciousness Studies,16*(6–8), 9–39.

Jaswal, V. K., & Malone, L. S. (2007). Turning believers into skeptics: 3-Year-olds' sensitivity to cues to speaker credibility. *Journal of Cognition and Development,8*(3), 263–283.

Jaswal, V. K., Croft, A. C., Setia, A. R., & Cole, C. A. (2010). Young children have a specific, highly robust bias to trust testimony. *Psychological Science,21*(10), 1541–1547.

Jaswal, V. K., Pérez-Edgar, K., Kondrad, R. L., Palmquist, C. M., Cole, C. A., & Cole, C. E. (2014). Can't stop believing: Inhibitory control and resistance to misleading testimony. *Developmental Science,17*(6), 965–976.

Jones, E., & Harris, A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology,3*, 1–24.

Kalish, C. W. (2002). Children's predictions of consistency in people's actions. *Cognition,84*(3), 237–265.

Kestemont, J., Vandekerckhove, M., Ma, N., Van Hoeck, N., & Van Overwalle, F. (2013). Situation and person attributions under spontaneous and intentional instructions: An fMRI study. *Social Cognitive and Affective Neuroscience,8*(5), 481–493.

Korman, J., & Malle, B. F. (2016). Grasping for traits or reasons? How people grapple with puzzling social behaviors. *Personality and Social Psychology Bulletin,42*(11), 1451–1465.

Krull, D. S., Seger, C. R., & Silvera, D. H. (2008). Smile when you say that: Effects of willingness on dispositional inferences. *Journal of Experimental Social Psychology,44*(3), 735–742.

Kushnir, T., & Koenig, M. A. (2017). What I don't know won't hurt you: The relation between professed ignorance and later knowledge claims. *Developmental Psychology,53*(5), 826–835.

Lane, J. D., Wellman, H. M., & Gelman, S. A. (2013). Informants' traits weigh heavily in young children's trust in testimony and in their epistemic inferences. *Child Development,84*(4), 1253–1268.

Liu, D., Gelman, S. A., & Wellman, H. M. (2007). Components of young children' s trait understanding: Behavior-to-trait inferences and trait-to-behavior predictions. *Child Development,78*(5), 1543–1558.

Luan, Z., Poorthuis, A. M. G., Hutteman, R., Denissen, J. J. A., Asendorpf, J. B., & van Aken, M. A. G. (2018). Unique predictive power of other-rated personality: An 18-year longitudinal study. *Journal of Personality,87*(3), 532–545.

Luo, Y., & Johnson, S. C. (2009). Recognizing the role of perception in action at 6 months. *Developmental Science,12*(1), 142–149.

Ma, N., Vandekerckhove, M., Baetens, K., Overwalle, F. V., Seurinck, R., & Fias, W. (2011). Inconsistencies in spontaneous and intentional trait inferences. *Social Cognitive and Affective Neuroscience,7*(8), 937–950.

Malle, B. F. (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. New York: Mit Press.

Mameli, M. (2001). Mindreading, mindshaping, and evolution. *Biology and Philosophy,16*(5), 597–628.

McGeer, V. (2007). The regulative dimension of folk psychology. In D. D. Hutto & M. Ratcliffe (Eds.), *Folk psychology re-assessed* (pp. 137–156). New York: Springer.

McNamara, R. A., Willard, A. K., Norenzayan, A., & Henrich, J. (2019). Weighing outcome vs intent across societies: How cultural models of mind shape moral reasoning. *Cognition,182*, 95–108.

Meltzoff, A. N., & Gopnik, A. (2013). Learning about the mind from evidence: Children's development of intuitive theories of perception and personality. In *Understanding other minds* (pp. 19–34).

Michael, J. (2014). Mindshaping: a new framework for understanding human social cognition. *Journal of Consciousness Studies,21*(11–12), 170–177.

Michael, J., & Christensen, W. (2016). Flexible goal attribution in early mindreading. *Psychological Review,123*(2), 219.

Miller, C. B. (2013). *Moral character: An empirical theory*. Oxford: Oxford University Press.

Miller, C. B. (2017). *The character gap: How good are we?*. Oxford: Oxford University Press.

Miller, R. L., Brickman, P., & Bolen, D. (1975). Attribution versus persuasion as a means for modifying behavior. *Journal of Personality and Social Psychology,31*(3), 430–441.

Mitchell, P. (2013). Mentalizing in autism: Interpreting facial expressions, following gaze, reading body language and inferring traits. *Journal of Educational Sciences and Psychology,LXV*(1), 111–120.

Moran, J. M., Jolly, E., & Mitchell, J. P. (2014). Spontaneous mentalizing predicts the fundamental attribution error. *Journal of Cognitive Neuroscience,26*(3), 569–576.

Parsons, S., & Mitchell, P. (1999). What children with autism understand about thoughts and thought bubbles. *Autism,3*(1), 17–38.

Peevers, B. H., & Secord, P. F. (1973). Developmental changes in attribution of descriptive concepts to persons. *Journal of Personality and Social Psychology,27*(1), 120–128.

Penn, D. C., & Povinelli, D. J. (2007). *The comparative delusion: The 'behavioristic'/ 'mentalistic' dichotomy in comparative theory of mind research 1 introduction 2 flogging the behavioristic strawman* (pp. 1–25).

Peters, U. (2019). The complementarity of mindshaping and mindreading. *Phenomenology and the Cognitive Sciences,18*(3), 533–549.

Pexman, P. M., Glenwright, M., Hala, S., Kowbel, S. L., & Jungen, S. (2005). Children's use of trait information in understanding verbal irony. *Metaphor and Symbol,21*(1), 39–60.

Reeder, G. D. (2009). Mindreading: Judgments about intentionality and motives in dispositional inference. *Psychological Inquiry,20*(1), 1–18.

Repacholi, B. M., & Meltzoff, A. N. (2007). Emotional eavesdropping: Infants selectively respond to indirect emotional signals. *Child Development,78*(2), 503–521.

Repacholi, B. M., Meltzoff, A. N., Hennings, T. M., & Ruba, A. L. (2016a). Transfer of social learning across contexts: Exploring infants' attribution of trait-like emotions to adults. *Infancy,21*(6), 785–806.

Repacholi, B. M., Meltzoff, A. N., & Olsen, B. (2008). Infants' understanding of the link between visual perception and emotion: "If she can't see me doing it, she won't get angry". *Developmental Psychology,44*(2), 561–574.

Repacholi, B. M., Meltzoff, A. N., Toub, T. S., & Ruba, A. L. (2016b). Infants' generalizations about other people' s emotions: Foundations for trait-like attributions. *Developmental Psychology,52*, 364–378.

Reynhout, G., & Carter, M. (2011). Evaluation of the efficacy of Social Stories$^{\mathrm{TM}}$ using three single subject metrics. *Research in Autism Spectrum Disorders,5*(2), 885–900.

Rholes, W. S., & Ruble, D. N. (1984). Children's understanding of dispositional characteristics of others. *Child Development,55*(2), 550–560.

Santiesteban, I., Catmur, C., Hopkins, S. C., Bird, G., & Heyes, C. (2014). Avatars and arrows: implicit mentalizing or domain-general processing? *Journal of Experimental Psychology Human Perception and Performance,40*(3), 929–937.

Scholl, B. J., & Leslie, A. M. (1999). Modularity, development and 'Theory of Mind'. *Mind and Language,14*(1), 131–153.

Schwitzgebel, E. (2019). *A theory of jerks and other philosophical misadventures*. New York: MIT Press.

Snow, N. E. (2010). *Virtue as social intelligence: An empirically grounded theory*. New York, NY: Routledge.

Sommerfeld, R. D., Krambeck, H.-J., Semmann, D., & Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Sciences,104*(44), 17435–17440.

Spaulding, S. (2018a). *How we understand others: philosophy and social cognition*. London: Routledge.

Spaulding, S. (2018b). Mindreading beyond belief: A more comprehensive conception of how we understand others. *Philosophy Compass,13*, e12526.

Spence, S. A., Hunter, M. D., Farrow, T. F. D., Green, R. D., Leung, D. H., Hughes, C. J., et al. (2004). A cognitive neurobiological account of deception: Evidence from functional neuroimaging. *Philosophical Transactions of the Royal Society B: Biological Sciences,359*, 1755–1762.

Sperber, D., & Baumard, N. (2012). Moral reputation: An evolutionary and cognitive perspective. *Mind and Language,27*(5), 495–518.

Sripada, C. S. (2012). Mental state attributions and the side-effect effect. *Journal of Experimental Social Psychology,48*(1), 232–238.

Sripada, C. S., & Konrath, S. (2011). Telling more than we can know about intentional action. *Mind and Language,26*(3), 353–380.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition,131*(1), 159–171.

Talwar, V., & Lee, K. (2008). Social and cognitive correlates of children's lying behavior. *Child Development,79*(4), 866–881.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences,22*(3), 201–212.

Thornton, M. A., & Mitchell, J. P. (2018). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex,28*(10), 3505–3520.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science,10*(1), 72–81.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping,30*(3), 829–858.

Vazire, S., & Carlson, E. N. (2011). Others sometimes know us better than we know ourselves. *Current Directions in Psychological Science,20*(2), 104–108.

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

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development,75*(2), 523–541.

Westra, E. (2017). Spontaneous mindreading: A problem for the two-systems account. *Synthese,194*(11), 4559–4581.

Westra, E. (2018). Character and theory of mind: An integrative approach. *Philosophical Studies,175*(5), 1217–1241.

Williams, S., Moore, K., Crossman, A. M., & Talwar, V. (2016). The role of executive functions and theory of mind in children's prosocial lie-telling. *Journal of Experimental Child Psychology,141*, 256–266.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition,69*(1), 1–34.

Wu, J., Balliet, D., & Van Lange, P. A. M. (2016). Gossip versus punishment: The efficiency of reputation to promote and maintain cooperation. *Scientific Reports,6*(1), 23919.

Yuill, N., & Pearson, A. (1998). The development of bases for trait attribution: children's understanding of traits as causal mechanisms based on desire. *Developmental Psychology,34*(3), 574–586.

Zagzebski, L. T. (2017). *Exemplarist moral theory*. Oxford: Oxford University Press.

Zawidzki, T. W. (2008). The function of folk psychology: mind reading or mind shaping? *Philosophical Explorations,11*(3), 193–210.

Zawidzki, T. W. (2011). How to interpret infant socio-cognitive competence. *Review of Philosophy and Psychology,2*(3), 483–497.

Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. Cambridge, MA: MIT Press.