Getting to know you: Accuracy and error in judgments of character

Forthcoming in *Mind & Language*

Evan Westra
Department of Philosophy
University of Toronto

Correspondence:
Department of Philosophy, University of Toronto, 170 St George Street, Toronto, Canada. Email: ewestra@utoronto.ca

Funding Source:
Social Sciences and Humanities Research Council Postdoctoral Fellowship 756-2018-0012

Abstract: Character judgments play an important role in our everyday lives. However, decades of research on trait attribution have shown that these judgments are prone to a number of biases and cognitive distortions. This raises a skeptical worry about the epistemic foundations of everyday characterological beliefs that has some deeply disturbing implications. I argue that this skeptical worry is misplaced: under the appropriate informational conditions, our everyday character-trait judgments are in fact quite trustworthy. I then propose a mindreading-based account of trait attribution that explains both why these judgments are initially unreliable, and how they eventually become more accurate.

Keywords: character-trait attribution; mindreading; accuracy; character traits; situationism.

1. Introduction

A young woman returns home from a first date. “How did it go?” Her roommate asks. “I didn’t like him,” she replies. “He was arrogant and closed-minded.” “Sounds like a real jerk,” the roommate says sympathetically.

A search committee discusses their impressions of a job candidate. “I thought she was quite brilliant,” one committee member says. “She seemed very sociable, like she would get along well with the faculty,” another adds. They conclude that the candidate would be a great fit.

Over lunch, two friends argue over whether they should support a political candidate. “She seems dishonest and cold,” one friend complains. “That’s just how she’s been portrayed by her opponents,” the other disagrees. “If you read her book or listen to her interviews, it’s clear that she genuinely cares about people.”

Judgments of character like these pervade our everyday social interactions. In our friendships, in our work, in our politics, we play close attention to the character traits and dispositions of those around us. Many of these judgments are about specifically moral traits like honesty and compassion, others are about epistemic traits like wisdom, while still others are about more evaluatively neutral traits such as extraversion. Often, these judgments arise quite quickly during a first impression, and are based on subtle non-verbal cues from a person’s face and body language. But as we get to know a person, we continue to build upon and refine our knowledge about her character. Some of our most confidently held beliefs are about the traits of those with whom we are most intimately acquainted, such as our spouses, immediate family members, and close friends.

But can these judgments be trusted? A survey of some of the most prominent findings in the empirical literature on trait attribution and person perception would seem to suggest otherwise.

According to this body of research, our ordinary character judgments are riddled with a host of biases that lead us to attribute character traits on the basis of superficial or erroneous evidence. Most prominent among this list of biases is the "fundamental attribution error" (FAE), which leads us to leap to characterological or "dispositional" interpretations of observed behavior, even in cases where obvious contextual or "situational" factors would provide an equally good or better explanation (Gilbert et al., 1995; Jones & Harris, 1967; Ross, 1977). To the FAE, we may also add a suite of other social biases and heuristics that lead us to infer character traits on the basis of stereotypes and physical appearance (Fiske, Cuddy, & Glick, 2002; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). Collectively, these biases cause us to rely upon a naïve dispositionalist theory that predisposes us to see behavior as the product of global character traits, regardless of whether this inference is warranted (Ross & Nisbett, 2011). Let us call the idea that these biases render our ordinary judgments about global character unreliable *attributional skepticism* (henceforth, AS).

AS is distinct from *situationism*, another skeptical thesis about character traits. Famously, the leading proponents of situationism have argued that the stable and cross-situationally consistent "global" character traits posited in both Aristotelian virtue ethicists and in our ordinary folk discourse do not actually exist (or are at least very rare); rather, they claim that most of us possess "local" character traits that vary widely as a function of the situations we find ourselves in (Doris, 2002; Harman, 1999). AS and situationism both imply a similar epistemic conclusion: ordinary folk beliefs about character rarely rise to the level of knowledge. Indeed, many situationists have also invoked AS as part of an error theory about why it is that we believe that people possess global traits, even though almost none of us really do (Alfano, 2011; Doris, 2002, pp. 92–106). But while situationism is a thesis about the *metaphysics* of character traits, AS is a thesis about the *epistemology* of character-trait *judgments*. These two theses come apart: although situationism seems to entail some version of AS (how else to explain the persistent belief in global traits?), AS does not imply situationism, and is even compatible with a realist view of global traits. That is, a proponent of AS could agree that global traits really exist, and that they can be measured with the right psychometric instruments, such as the Big Five personality questionnaires. However, the proponent of AS could still insist that our ordinary folk practices of character-trait attribution are ill-suited to track these real psychological properties. We are, in short, poor judges of character.

Although it is compatible with realism about character, the skeptical challenge posed by AS is every bit as troubling as the one posed by situationism, and has many disturbing implications for both ordinary and philosophical thinking about character. For example, a common way to think of moral responsibility is to view character-inconsistent actions as less blameworthy than character-consistent ones (Frankfurt, 1971; Watson, 1975) – hence the exculpatory force of declarations like, "He was not himself!" But if our intuitive judgments about character are unreliable, then so are these judgments of moral responsibility. AS would also be problematic for moral theories that instruct us to look to the examples set by virtuous individuals for normative guidance, such as Linda Zagzebski's exemplarist virtue ethics (Zagzebski, 2017): if we take our cues from perceived moral exemplars, but our ability to identify these exemplars based on their characters is untrustworthy, how can we know that we are following the correct moral example? More generally, any normative theory that requires us to accurately judge a person's traits will be threatened by AS.

But more viscerally, AS implies that we usually do not truly *know* other people – even those whom we love deeply. If AS is right, then we cannot trust even our most confidently held beliefs about other people's character traits – not about our romantic partners, siblings, parents, best friends, co-workers, or anybody else. This is a profoundly alienating conclusion. In the news, we often hear the friends and family members of violent criminals express deep confusion and bewilderment upon learning that someone they thought they knew could do something so monstrous. "I guess you can never really know a person," they sometimes say. AS suggests that this is true across the board. For these reasons, it is important to set AS apart from situationism, and evaluate it on its own merits.

In this paper, I will argue that in spite of the evidence for AS, we should not lose faith in our folk attributions of character. Although our ordinary character judgments are sometimes based on certain heuristics and biases, they are nevertheless generally reliable when judged against appropriate standards. I will begin by reviewing the empirical case for AS in a bit more detail, and argue that it fails to support broad skepticism about our character judgments. Next, I present evidence from the literature on accuracy in personality judgment that our ordinary character judgments are in fact quite reliable under the right informational conditions. I then turn to the task of explaining how this evidence for accuracy can be reconciled with the evidence for error in trait attributions that motivates AS. To do this, I offer a theory of the underlying cognitive processes that ground both reliable and unreliable character judgment, which, I will suggest, are a part of our capacity for reasoning about mental states, or "mindreading."

1.1. The nature of character traits

First, however, I must say a few words about the scope of this paper, and its starting assumptions about the nature of character traits. To isolate the specific challenge posed by AS, it is necessary to clearly distinguish it from the kind of skepticism about character judgment that follows from situationism. If situationism were true, then AS becomes trivially true as well. If realism about global traits were true, however, then the truth of AS would remain an open question. Thus, discussion of AS is much more dialectically interesting given some form of realism about global traits. There are, moreover, many compelling reasons to doubt the empirical merits of situationism (Jayawickreme, Meindl, Helzer, Furr, & Fleeson, 2014; Paris, 2016; Sabini & Silver, 2005). Briefly, much of the empirical evidence cited in favor of situationism ultimately fails to support it. For example, situationists have often pointed to the Good Samaritan and Milgram experiments as demonstrating that compassion is determined by situational factors (Darley & Batson, 1973; Milgram, 1963). Yet these studies only report effects averaged across populations of individuals, and are not capable of measuring stable individual differences in behavior across contexts, which is the core of the global theory of traits (Jayawickreme et al., 2014). Situationists have also cited findings showing that individual behavior varies significantly between one situation and the next (Hartshorne & May, 1928); however, these findings fail to reflect the strong cross-situational patterns of behavior that emerge when we aggregate over a large number of observations (Epstein, 1979; Fleeson & Jayawickreme, 2017). Perhaps the greatest challenge to situationism comes from the vast body of evidence within personality psychology supporting the existence of the Big Five[1] trait dimensions, which are significant predictors of major life outcomes such as academic achievement, mental and physical health, job performance, and length of life (Noftle & Robins, 2007; Ozer & Benet-Martínez, 2006; Paris, 2016). In short, there is very good evidence for the existence of global traits, and very poor evidence for situationism. Therefore, a starting assumption of this paper will be that global traits do indeed exist, and that situationist challenges to the global character thesis fail.

Going forward, I will be taking it for granted that the Five Factor model of personality (or perhaps a related model such as HEXACO (Ashton & Lee, 2007)) provides a descriptively accurate approximation of the characterological dimensions along which people exhibit stable differences. More specifically, I will be adopting a theoretical approach to personality traits known as "Whole Trait Theory" (Fleeson & Jayawickreme, 2015). According to this view, trait terms refer to stable intraindividual distributions of cognitive, affective, motivational states, which tend to produce

[1] The Big Five trait dimensions are Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (or Emotional Stability); they are sometimes abbreviated to the acronym OCEAN. Each trait dimension is comprised of a number of trait "facets" (e.g. the facets of Agreeableness are trust, straightforwardness, altruism, compliance, modesty, and tender-mindedness), which in turn break down into various sub-facets, giving the Big Five a hierarchical structure (John & Srivastava, 1999).

characteristic patterns of behavior in trait-relevant situations (c.f. Mischel, 2004; Mischel & Shoda, 1995; Snow, 2010). For example, the trait of extraversion has been shown to partly consist in the habitual pursuit of certain types of goals, such as the having fun, making new friends, and getting others to do what you want; the trait of conscientiousness consists partly in the habitual pursuit of goals like completing tasks and using one's time efficiently (McCabe & Fleeson, 2012, 2016). In other words, the Whole Trait Theory postulates that the descriptive and predictive adequacy of the Big Five are explained by regularities at the level of her intentional mental states. Thus, in what follows, a belief of the form "S has character trait T" will be true just in case S is disposed to manifest a T-like distribution of psychological states over time. Reliable trait judgment consists in the capacity to engage in true trait attributions in a consistent manner.

On this model, the possession of a particular trait is consistent with the observation that people exhibit a great deal of intraindividual variability. For example, it is not unlikely that a moderately extraverted person will, over the course of a week, think, feel, and behave in ways that are consistent with both extreme extraversion and extreme introversion (Fleeson, 2001). But from week to week, or month to month, this same person will reliably manifest the same mean levels of extraverted states, and the same general distribution of extraversion-relevant behaviors (Fleeson, 2001; Fleeson and Gallagher, 2009; Heller *et al.*, 2007).[2] Less technically, one can think of the relation between traits and manifestations of trait-relevant psychological states as analogous to the relation between climate and weather: the latter describes noisy, highly variable phenomena on shorter time scales, while the former describes broad, stable patterns that only emerge on longer time scales.

Finally, note that the ensuing discussion concerns "character traits" in a broad sense. It is meant to include both normatively relevant traits such as moral and epistemic virtues and vices, and

[2] As reported using experience-sampling measures. Notably, mean levels of these self-reported "personality states" strongly correlate with global self-reported trait ratings (Finnigan & Vazire, 2018; Fleeson & Gallagher, 2009).

personality traits such as extraversion and open-mindedness. Other authors draw more fine-grained distinctions between these categories (e.g. Miller, 2014), but for present purposes a broad brush will suffice.

2. Attributional skepticism

By far the strongest evidence in support for AS comes from the study of the fundamental attribution error (FAE), which is the tendency to over-attribute the causes of behavior to intrinsic, dispositional factors while ignoring relevant situational factors. In a classic demonstration of this phenomenon, Ross and colleagues (1977) had their subjects participate in a quiz show game in which some played the role of "questioner," while others played the role of "contestant." The questioners were instructed to make up difficult questions that would be difficult for the contestants to answer. Afterwards, observers were asked to rate the intelligence of the questioners and contestants. Even though the contestant and questioner roles were arbitrarily assigned, and the questioners had a much easier task than the contestants, observers still generally rated the former as more intelligent than the latter. Observers, in other words, seemed to either not notice or discount the strong situational constraints on the behavior that they had just witnessed, and instead inferred that it transparently reflected the individual's stable traits. This basic phenomenon has been replicated dozens of times, and has been referred to as one of social psychology's most robust findings (Ross & Nisbett, 2011).[3]

Numerous other findings suggest that we are very quick to attribute global traits to other people on the basis of very slim evidence. For example, we tend to make very rapid inferences about traits on the basis of a person's facial appearance (Todorov, Said, Engell, & Oosterhof, 2008). Faces that appear older and more masculine tend to be seen as more dominant and powerful than faces that

[3] Notably, certain East Asian populations appear to be less prone to the FAE, although dispositional thinking is still common in these groups (Choi & Nisbett, 1998; Choi, Nisbett, & Norenzayan, 1999).

appear younger and more feminine. Faces with large, wide-set eyes tend to be seen as more trustworthy, while faces with small, narrow-set eyes are seen as less trustworthy. We also tend to view more attractive faces as more extraverted. These trait attributions are also made extraordinarily quickly, even when faces are only presented for 100 milliseconds (Todorov, 2013).

Another pathway to rapid trait attribution comes via stereotypes (Westra, 2017). Intuitively, many stereotypes about social groups prominently feature characterological information, and one can easily call to mind examples of stereotypes about groups that are viewed as lazy, greedy, deceitful, aggressive, ambitious, intelligent, and so on. Researchers interested in the contents of stereotypes have discovered that these trait attributions frequently follow a common pattern. When analyzing stereotypes using two basic trait dimensions – the "warmth" dimension and the "competence" dimension – we find that most common stereotypes fall into one of four types, each picking out a distinct cluster of correlated traits (Fiske et al., 2002). High-warmth, low-competence stereotypes, for example, tend to include traits like compassion, friendliness, and trustworthiness, but also traits like low intelligence, frivolity, and clumsiness; low-warmth, high-competence stereotypes, in contrast, include traits like high intelligence and ambition, and also deceitfulness and greed. These clusters of trait attributions get associated with particular groups and become rapidly activated when we encounter cues that indicate a person's group membership and influence our subsequent expectations about their intentions and behavior.

There is also evidence that we spontaneously encode behavioral information in terms of corresponding traits, and bind this trait information to particular actors (Uleman, Adil Saribay, & Gonzalez, 2008; Uleman, Hon, Roman, & Moskowitz, 1996). Todorov and Uleman, for example, showed participants pictures of different actors paired with various trait-implying behavioral descriptions (e.g. "Alice solved the murder halfway through the book," implying the trait "clever")

(Todorov & Uleman, 2003). Participants were then shown the images again, paired with different trait words, and were instructed to say whether the word had appeared in the behavioral description paired with the face in question. They found that false recognition rates were highest when the actors' faces were paired with trait words that fit the behavioral descriptions, even when exposures to the actors' faces were brief and when participants were under cognitive load, suggesting that these "spontaneous trait inferences" or STIs are both rapid and cognitively efficient.

In short, there is ample evidence that we are extremely quick to make inferences about traits on the basis of very superficial evidence, and that we also ignore other, relevant forms of counter-evidence that would otherwise undermine these inferences. Some situationists have argued that this is enough to cast doubt on the reliability of all of our ordinary judgments about traits. Alfano (2011) puts the point succinctly:

> The existence of these biases does not prove that no one has traits, nor does it demonstrate that no arguments could warrant the conclusion that people have traits. What it instead shows it that *regardless of whether people have traits, folk intuitions would lead us to attribute traits to them* [italics in original]. (Alfano, 2011, p. 133)[4]

The crux of this argument is that our ordinary beliefs about people's traits simply do not track their real psychological dispositions. Though these dispositions may well exist, we are not in an epistemic position to have any knowledge of them. While this view might allow that appropriate psychometric

[4] In addition to the FAE, Alfano (2011) argues that the false consensus effect specifically reinforces the FAE by leading us to believe that our intuitive trait attributions are widely shared. However, the false-consensus effect is a domain-general bias, and there is no evidence that its effects on trait attribution are any greater than on any other belief that we falsely presume to be widely shared. Along similar lines, Alfano argues that various other domain-general biases (such as the confirmation bias) further degrade the reliability of our faulty trait attributions. But because these biases affect all sorts of reasoning, and not just inferences about traits, they cannot explain why our ordinary character judgments *in particular* should be viewed as unreliable.

instruments could justify certain global trait attributions, our everyday judgments are woefully inadequate for the task.

3. In the long run, trait attribution is accurate

While the empirical support for AS may seem compelling, the fact that we display systematic patterns of error in our commonsense trait judgments does not entail that they are generally unreliable. After all, many highly reliable psychological mechanisms are prone to systematic error under certain conditions. Our eyes fail us all the time when it's dark out. Our ears don't work particularly well when we're in a loud concert. It's hard to taste accurately immediately after eating something spicy. When informational conditions are poor, our senses fail us. But this fact does not lead us to treat the deliverances of our senses as inherently untrustworthy. Rather, we recognize that these mechanisms operate reliably within a certain range of informational conditions, and unreliably outside that range.

I argue that the cases of faulty trait attribution invoked in support of AS are drawn from outside the range in which our trait judgments are normally reliable. In FAE scenarios, for example, participants are asked to make judgments about a stranger on the basis of a very limited set of observations. In spontaneous trait-attribution scenarios, participants rely on even less information: a one-sentence behavioral description. In face-based trait attributions, participants are literally asked to make judgments on the basis on facial appearance alone. But these are hardly adequate conditions for making an inference about a person's traits. If by hypothesis traits are *cross-situationally stable* properties of individuals, then almost any inference about the presence of a trait *based on single behavioral observations* will be unreliable. To expect otherwise is just to have unrealistic expectations about the conditions under which trait attributions can be apt. A well-supported global trait attribution should require numerous observations of the target in many different contexts.[5] In short, measuring the reliability of trait attributions on the basis of single observations of behavior is like giving a person an eye exam in the dark, or a hearing test during a metal concert. Yes, our judgments in these contexts are unreliable. But this is neither surprising nor evidence for skepticism about these judgments more generally.

3.1. Accuracy in personality judgments

But when *can* we expect our trait attributions to be reliable? One very natural suggestion is that our folk judgments about traits are most likely to be accurate when we make them about people with whom we are well acquainted. While we may not be so reliable in our trait attributions about strangers, perhaps we are much better when it comes to judging the traits of our family members, romantic partners, and close personal friends. My beliefs about the traits of my brother, for instance, are based on a lifetime of observations and interactions with him. Indeed, sometimes our family members know us even better than we know ourselves. Or take the experience of dating: on a first date, we may form various first impressions about a person's character that we subsequently revise after being romantically involved with them for several months or years; these later beliefs are almost always more confidently held than the earlier ones. This more advanced characterological knowledge plays a huge role in our everyday lives, and a crucial part in the maintenance of our most important relationships. Surely these beliefs about character are much more reliable – and more significant – than the superficial first impressions that form the basis for the skeptical worry outlined above.

[5] A possible exception to this would be single observations of behaviors that are both unusual and highly diagnostic for an underlying trait. Witnessing someone torture a puppy, for example, is very good evidence that they are cruel, both because such behavior is quite rare, and because it is a strong indicator of cruelty (Uhlmann, Zhu, & Diermeier, 2014).

This proposal is borne out by a substantial body of research on accuracy in ordinary people's personality judgments (Connelly & Ones, 2010; Funder, 1995; Kenny, 1991; Vazire & Carlson, 2011). Whereas many psychologists researching "person perception" have focused on the various errors and biases that affect our beliefs about other people, accuracy researchers are most interested in the conditions under which we get these judgments right. When we shift our attention from error to accuracy, it becomes apparent that, for all our heuristics and biases, we're actually pretty *good* judges of character.

Different researchers operationalize "accuracy" in different ways, but in general, it is measured by looking at the degree of consensus between different third-party personality judgments ("other-other agreement"), and the degree of correspondence between third-party and self-reported personality judgments ("self-other agreement").[6] In practice, these studies involve a target individual who fills out a standard personality test about herself, and a number of other "judges" who fill out the same test about the target. If these individuals are tracking the same underlying psychological properties, the reasoning goes, we should expect their judgments to correlate with one another; stronger correlations between the different test results are thus interpreted as reflecting greater accuracy. This approach should be familiar to philosophers of science: convergence across relatively independent observations represents what William Whewell called "a consilience of inductions," and speaks to the reality of the underlying phenomenon being measured (Helgeson, 2013; Whewell, 1989) – in this case, global traits.

[6] A few studies have used third-party personality judgments to predict specific behaviors, both in laboratory and naturalistic contexts, and obtained positive results (Fast & Funder, 2008; Funder, 1995; Vazire et al., 2010). However, these correlations observed tend to be somewhat weaker than those observed in self-other agreement and third-party consensus measures (Connelly & Ones, 2010). This is unsurprising: ordinary people display a wide range of variability in their trait-relevant behaviors throughout their day-to-day lives; trait-level regularities only appear when behavior is measured in an aggregate form (Epstein, 1979; Fleeson, 2001). Thus, even weak correlations between third-party trait-attributions and specific behavioral outcomes are still notable. Still, for the purposes of argument, I focus here on consensus-based methods.

Of course, mere convergence is not enough to establish accuracy. If these personality judgments are really tracking actual traits, then the degree of convergence between them should vary as a function of evidential factors that ought to influence accuracy (Funder, 1995). For example, traits like extraversion have more transparent behavioral correlates, which should make them easier to observe, and thus elicit higher degrees of intersubjective agreement; in contrast, more "internal" traits, such as introspectiveness, lack obvious behavioral correlates, and therefore ought to elicit lower levels of agreement. Similarly, judges who have had more opportunities to observe a target (e.g. close acquaintances) ought to show greater agreement than those who have had less (e.g. strangers). The trait-relevance of the evidence a judge has should also matter: whether or not I know that you have a clean desk at work tells me a lot about how conscientious you are, but not a lot about whether you are "open to experience." In other words, accuracy in personality judgment should lead to systematic variability in the correlations between different third-party judgments. In contrast, AS predicts that this variability should be insensitive to the appropriateness of different evidential factors.

In a meta-analysis spanning 263 independent samples and 44,178 targets, Connelly and Ones found strong evidence that many of the above factors do indeed moderate the accuracy of personality judgments (Connelly & Ones, 2010). In particular, both self-other agreement and other-other consensus were strongest when the judges were *intimate acquaintances* of the target. This was most pronounced in dating partners and spouses, followed by other family members and friends, followed by work colleagues and casual acquaintances. These correlations were particularly strong for less "observable" traits, such as Agreeableness, Openness, and Emotional Stability. In contrast, more "observable" ones such as Extraversion exhibited stronger agreement and consensus among strangers and in the sample overall. In other words, the intimacy of a relationship between a judge and a target often translates into greater accuracy in global trait judgments, particularly for traits that are hardest to infer. Thus, unsurprisingly, the better you know someone, the better you know her character.

3.2. Objections

One potential objection to this kind of methodology is that consensus and agreement might be tracking something other than real character traits. Widely shared stereotypes, for instance, would also generate consensus across judges, but seem unlikely to reflect judges' sensitivity to targets' unique personalities.

But this cannot explain the pattern of results that we see in accuracy research. How, for instance, could stereotypes account for greater consensus among close personal acquaintances than strangers, since stereotypes are *ex hypothesi* widely shared? Accuracy researchers have also developed techniques for quantifying unique contributions of widely shared beliefs about personality versus personal knowledge of the target, which they label as "normative accuracy" and "distinctive accuracy," respectively (Biesanz, West, & Millevoi, 2007; Cronbach, 1955). Roughly, normative accuracy measures the extent to which a judge's personality ratings track the average personality ratings within the sample, while distinctive accuracy measures correlations between judges after controlling for normative accuracy. Consistent with the hypothesis that agreement and consensus are driven by judges' actual observations of the targets, Biesanz and colleagues have found that distinctive accuracy increases the longer a judge knows a target, while normative accuracy does not. Thus, while widely shared beliefs about personality do account for some level of baseline consensus among strangers, additional gains in consensus reflect unique experiences of the target (Biesanz et al., 2007).

A critic might also argue that ordinary personality judgments track outward *performances* of character that do not necessarily reflect deeper psychological realities. For example, a person may project an outward image of extraversion, while secretly despising the company of others. Unlike the stereotype hypothesis, which explains consensus as the result of widely shared information, such performances of character would generate greater levels of agreement only among judges who have directly observed the target, such as more intimate acquaintances. This would explain why such individuals display more "accurate" character judgments, even though these judgments would not correspond to the target's real character, which remains concealed.

There are a few reasons to doubt this hypothesis. First, it implies that most people are constantly engaged in elaborate and lifelong acts of deception across all of their relationships. This is implausible. Second, there are telling asymmetries between self- and other-reported personality ratings that speak in favor of genuine trait tracking. For example, third parties' judgments about highly evaluative traits (such as creativity and intelligence) tend to be more accurate predictors of highly desirable or undesirable life outcomes (e.g. work and educational outcomes, depression, and relationship satisfaction) (Luan et al., 2018; Vazire et al., 2010; Vazire & Carlson, 2011). In other words, people have blind spots about their own personalities (particularly when it comes to personally important traits) that third parties are better able to detect. Thus, even if our personalities were simply mere performances, these performances still reveal more about the performers than they themselves realize.

Another potential objection to this argument comes from Jonathan Webber, who anticipates how a situationist might respond to the suggestion that trait judgments become more accurate with acquaintance (Webber, 2007). He suggests that accurate beliefs about the character traits of one's loved ones could simply be due to familiarity with that person's behavior *in particular situations*. On this view, increasing accuracy in personality judgments reflects a tacit capacity for tracking situational influences on behavior, rather than broad-based underlying dispositions.

However, accuracy in personality judgments appears to be robust across many different situations, insofar as it is robust across different kinds of relationship. Both siblings and romantic partners, for instance, show high levels of self-other agreement and other-other consensus. Both of these types of relationships are likely to reflect a fairly broad range of situational experiences with a person, compared with, say, a college dorm-mate or co-worker, whose personality judgments are generally less accurate. Moreover, the situations in which these high-acquaintance relationships take place do not completely overlap: your knowledge of your sibling's character is derived from very different kinds of situations than your knowledge of your spouse's character. Yet such judgments tend to show reliable correlations both with each other and with the self-ratings of the targets themselves. If judges were tracking situational factors rather than characterological ones, these correlations would be quite surprising.

Overall, the evidence for accuracy in personality judgments provides an impressive counterpoint to the case for unreliability and error. While it is true that we are prone to many trait-attribution biases, the scope of this evidence is limited to a narrow and unrepresentative subset of the interactions in which we make character judgments. Under the right informational conditions, our character judgments are quite reliable. Arguably, these are also the conditions in which it is most important to us that we have accurate character judgments – namely, in our close personal relationships.

4. To learn about character, we read minds

One shortcoming of the preceding discussion is that the accuracy literature is silent about the cognitive bases for our trait judgments. When it comes to reconciling the evidence for error with the evidence for accuracy, this becomes a problem. Why, for instance, are we so prone to make hasty trait attributions? Why do we not simply refrain from making these judgments until we have sufficient evidence? And how is it that we transition from these rough and unreliable initial inferences to fine-grained and accurate judgments as we get to know a person better? Though the threat of AS has been held at bay, many questions remain.

4.1. Trait attribution and theory of mind

The place to start, I suggest, is the nature of character traits themselves. As I stated in the introduction, traits are constituted by stable distributions of psychological states, which in turn produce trait-relevant behavior. Given this account of traits, it follows that reliable, accurate trait judgments will be those that track regularities in our underlying mental states. Accurate trait attribution should thus depend upon our ability to accurately reason about patterns in people's mental states, a capacity known as "mindreading" or "theory of mind."

There is a large body of evidence suggesting that theory of mind and trait attribution are functionally related (for a review, see Westra (2018)). For example, there is both neuroimaging and behavioral evidence that we are able to dynamically update representations of people's traits in response to new mental-state information, especially when this involves mental states that are inconsistent with a prior trait attribution (Cloutier, Gabrieli, O'Young, & Ambady, 2011; Kestemont, Vandekerckhove, Ma, Van Hoeck, & Van Overwalle, 2013; Ma et al., 2011). Inconsistent mental-state information also appears to moderate the tendency to commit the FAE: when we observe someone act in a trait consistent way, but for a trait inconsistent motivation, we do not attribute a corresponding trait to that person (Krull, Seger, & Silvera, 2008; Reeder, 2009; Reeder, Kumar, Hesson-McInnis, & Trafimow, 2002; Reeder, Vonk, Ronk, Ham, & Lawrence, 2004). Conversely, inferences about a person's stable traits also appear to shape our beliefs about their mental states, such as whether or not we interpret a foreseeable side-effect of an action was intended or unintended (Sripada, 2012; Sripada & Konrath, 2011). The connection between trait attribution and psychological knowledge is also evident in childhood: by 15 months of age, children are able to use regularities in a person's emotional expressions combined with perceptual cues to make trait-like inferences about her expected behavior (Repacholi, Meltzoff, Hennings, & Ruba, 2016; Repacholi, Meltzoff, Toub, & Ruba, 2016). Around the same age that they start passing the false-belief task (Wimmer & Perner, 1983), children start to be able to use linguistic trait labels to attribute desires and predict behavior (Heyman & Gelman, 2000; Liu, Gelman, & Wellman, 2007).[7] In short, from childhood onwards representations of traits seem to shape our expectations about people's psychological states and behavior, and information about mental states moderates inferences about their underlying traits.

There is also tight functional integration between trait attribution and mindreading at the neural level, where both processes are supported by the same "social brain network" (Van Overwalle, 2009). It also appears that mental state and trait representations share a common neural code. In a recent study employing multivariate pattern analysis or "neural decoding" techniques, Thornton and Mitchell found that a classifier trained on neuroimaging data of participants thinking about different traits (e.g. agreeableness) was then able to successfully predict when participants were making inferences about corresponding psychological states (e.g. happiness), suggesting that the brain uses a common representational space and population codes to individuate both stable traits and the transitory mental states that constitute them (Thornton & Mitchell, 2018). In other words, we seem to think about traits and mental states in a very similar way.

Thus, it seems as though our social cognition is structured in a way that reflects the real metaphysical relationship between traits and mental states. One challenge in interpreting these results, however, is that most well-known theories of mindreading have not specifically considered the role of trait

[7] Interestingly, the tendency to commit the FAE in explanations of behavior emerges relatively late, around age 6 (Seiver, Gopnik, & Goodman, 2013). Discussing these findings, Meltzoff and Gopnik suggest that it reflects the emergence a higher order belief or "framework theory" about the probability that a given action reflects a person's underlying traits (Meltzoff & Gopnik, 2013). Because the tendency to commit the FAE varies from culture to culture, they suggest that the acquisition of this higher order expectation may be moderated by environmental factors, such as adults' use of trait-based explanations of behavior.

representations (e.g. Butterfill and Apperly, 2013; Goldman, 2006; Nichols and Stich, 2003; Wellman, 2014). Recently, however, a number of authors working within the framework of predictive coding have begun to address this oversight (Bach & Schenke, 2017; Koster-Hale & Saxe, 2013; Tamir & Thornton, 2018; Theriault & Young, 2018; Westra, 2018). The key insight of this approach has been to think of trait attribution and mindreading as two levels of representation within a broader, hierarchically structured system for action prediction. Tamir and Thornton (2018), for example, propose a top-down model of social cognition wherein character information represented in a low-dimensional trait space is used to generate predictions about a person's likely mental states, which are represented in a parallel state space. This mental state information – along with priors about transitional probabilities between states – is then used to compute the person's likely actions. Adopting a more Bayesian formulation of a similar idea, Westra (2018) has proposed that trait representations function as over-hypotheses (Kemp, Perfors, & Tenenbaum, 2007) that help us to assign prior probability distributions over the different possible mental states that a particular person might exhibit.[8] For example, the belief that someone is ambitious might lead us think it is more antecedently likely that that person will have self-serving motivations rather than altruistic ones. These prior probabilities can then inform how we interpret particular actions that would otherwise be ambiguous between multiple intentional interpretations (e.g. when a self-serving, ambitious person makes a large charitable donation). In both of these models, traits play a similar role: narrowing down the space of possible mental-state hypotheses in order to help us generate person-specific behavioral predictions.

As Westra (2018) notes, the superordinate position of traits within the hierarchy of psychological representations can help to explain our habit of making hasty trait attributions based on minimal

[8] See also Meltzoff and Gopnik (2013), footnote 9.

information. Since traits have a top-down effect upon mental-state attributions and action predictions, and help us to generate an initial hypothesis about what person will do, it makes sense to prioritize trait attribution in the social prediction process. This is not unlike how "gist" representations function in visual processing: initial predictions about a scene are the product of rapid, coarse-grained representations, which are then updated and filled in as more fine-grained information is processed (Bar, 2007; Chaumon, Kveraga, Barrett, & Bar, 2014; Clark, 2015). Likewise, spontaneous trait inferences provide our social predictions with an initial gist-like representation of a person's mind, which can then be updated and fine-tuned as we acquire more information about their actual mentalistic dispositions. Thus, rapid trait attributions can serve as an early scaffold of the social prediction process.

One might object that this analogy to visual processing is misleading. While gist representations are imprecise, they do contain real information about visual scenes; in contrast, spontaneous trait inferences based on faces and stereotypes carry no real information about a person's character. Thus, while gist representations provide a reliable baseline for scene processing, the same cannot be said for early trait attributions. However, early trait attributions do not need to accurately reflect a target's character in order to help us learn about it. All they need to do is generate predictions. The extent to which these predictions match or fail to match their targets creates new information in the form of prediction errors, which can be used as feedback to revise the model and generate new predictions (Ma et al., 2012; Mende-Siedlecki, Cai, & Todorov, 2013). Much like learning in other domains (Gopnik & Wellman, 2012; Perfors, Tenenbaum, Griffiths, & Xu, 2011), learning about a person's character can be characterized an iterative process of prediction, feedback, and updating (c.f. Cunningham *et al.*, 2007). In this manner, even early, inaccurate trait attributions can help us acquire knowledge of character in the long run.[9]

A consequence of this model is that we should develop increasingly nuanced models of a person's psychology as we come to learn more about them. Consistent with this picture, there is evidence that as we get to know a person more intimately, we think about their behavior in a much more detailed, mentalistic fashion. Serena Chen observed the emergence of this tendency in a study that investigated how people reason about their significant others (Chen, 2003). Chen asked participants to each describe the reaction that their significant other, a generic stereotypical individual, and a nonsignifcant other would have to a range of situations using "IF…THEN…" prompts (e.g. "IF at a party, THEN [Amy talks to everyone]"). Participants were additionally asked to provide explanations of these responses, which were coded based on their references to psychological states. Chen found that participants tended to provide a significantly higher proportion of psychological state explanations for their romantic partners than for the other two social targets, and that people were significantly more confident about these explanations. Participants also found it easier to generate more descriptions based on psychological prompts (e.g. "IF Amy feels threatened…") for their significant others than for other social targets. These findings are especially significant in light of the evidence that romantic partners are also highly accurate in their personality judgments, suggesting that these kinds of inferences are supported by more rich and detailed mindreading (Connelly & Ones, 2010).

[9] Of course, many factors might interfere with this iterative updating process. If one lacks the opportunity to gather more information about a target, or the motivation to do so, then one's early, inaccurate trait attributions are likely to persist unchanged. By the same token, the extent to which one is able to learn about a person's character will be partly a function of the evidence one is able to gather, and partly a function of one's motivations (Westra, 2017).

Thus, while our initial impressions of a person are prone to a range of biases, as the attributional skeptic points out, these biases simply reflect the early stages of a long-term, mentalistic learning process. Early characterological judgments provide an initial guess as to a person's likely distribution of states, but we gather more information and receive feedback on our predictions, this distribution is iteratively updated. With increasing experience, this model of a person's likely mental states should provide an increasingly accurate approximation of her actual mental-state distributions – that is, her character.[10]

4.2. Mindreading or mindshaping?

There is one element that is absent from this account: *the influence of the judge herself upon the target*. The mindreading framework can sometimes give the impression that our relation to other minds is essentially detached and spectatorial (Hutto, 2004). But as judges of character, we do not merely observe the behavior of other people in a dispassionate manner: we actively engage with them, and are ourselves a part of their environment. Indeed, a consequence of my account is that the best judges of character will also be exceptionally important elements of a target's social circle (think, for instance, of the incomparable impact that a parent has upon a child's life). This opens up an opportunity for judges to causally influence the character traits of the targets through their own actions. As some philosophers have suggested, many different social practices provide avenues for this "mindshaping" to occur: teaching and pedagogy, knowledge transmission through testimony,

[10] A reviewer observed that this account seems to tread a fine line between moving from a coarse-grained representation of a person's character traits to a more nuanced one, and *overcoming* a prepotent tendency to represent a person in terms of traits in favor of a more accurate, mentalistic model. This distinction turns on one's conception of traits. According to the account of traits that I start with in section 1, traits are constituted by distributions of mental states; on this view, knowledge of character *just is* knowledge of a person's likely mental states. If, however, one holds a different theory of traits – say, the behavioristic account of traits adopted by situationists (e.g. Doris, 2002) – then this objection would have some purchase, as the knowledge acquired through this mentalistic learning process would not then count as knowledge *of character traits*. But insofar as the view of traits that I've adopted is empirically well-supported, and also aligns well with our folk psychological understanding of traits, this objection does not pose a threat to the current account.

punishment and reward, moral praise and blame, and even simple kindness and cruelty all create experiences that could influence a person's attitudes and behavior (McGeer, 2007; Zawidzki, 2013) – and, ultimately, her character.[11]

The mindshaping hypothesis suggests a final objection to the account that I have been defending. Rather than simply developing greater empirical knowledge of a person's character through careful observation, it could be argued that we are in effect making our character trait judgments come true through our own interventions, as a kind of self-fulfilling prophecy.[12] We are not good judges of character because we are good mind*readers*, but because we are good mind*shapers*.

The idea that our relationships and the people in our lives influence our personalities is obviously correct. It would thus be a mistake to discount the idea that judges' attributions are true in part because of their own actions. However, mindshaping cannot explain the broad patterns of accuracy in character judgment that we see in the data. For one thing, targets interact with many different people in their lives, and tend to have multiple significant relationships. Each of these intimate acquaintances will most likely have a distinct impact on the target's personality. Thus, a judge's task is not simply to predict the impact of her own actions upon the target, but the effects of the actions of all the target's other intimate acquaintances as well. Any advantage in trait attribution brought about through mindshaping would thus be negated by the mindshaping activities of other people (unless these judges somehow managed to collude to bring about the same characterological end). Moreover, one of the signatures of accuracy is convergence among multiple judges' trait ratings

[11] Notably, Alfano has invoked a similar notion to support his fictionalist approach to virtue ethics: by labeling people as virtuous, we can actually encourage them to act as though they really do possess those virtues (Alfano, 2013).

[12] This proposal is similar to the oft-cited Pygmalion effect, which purported to show that teachers' expectations about students' academic performance caused students' academic performance to improve (Rosenthal & Jacobson, 1968). For a recent critique of this finding, see Jussim (2012).

across the target's different relationships. This suggests that what explains judges' accuracy is something common across all of their experiences (namely, the character of the target), rather than facts about their individual interactions with the target. Finally, while personality is certainly influenced by experience, it also has a heritable genetic component, which limits the extent to which it could be influenced by mindshaping *qua* environmental factor (McCrae et al., 2000). Thus, while judges are sometimes mindshapers when it comes to targets' character traits, accurate trait judgment will still depend upon reliable mindreading.

5. Conclusion

This paper began with a skeptical challenge to our everyday judgments of character and the threat attributional skepticism seems to pose to our experiences of knowing other people. By drawing on the literatures on accuracy in personality judgment and mindreading, I then sketched out an account of why our character judgments can be initially unreliable, as well as the inferential processes that ultimately render them more trustworthy. I do not claim that these judgments are reliable across the board, or that biases like the fundamental attribution error are morally and epistemically unproblematic. But I maintain that broad skeptical worries about character-trait attributions are empirically unwarranted.

In effect, this has been a defense of common sense. Prominent social psychologists and philosophers have encouraged us to mistrust our ordinary character judgments, but this has never been consistent with our everyday experiences. We are not, for instance, constantly shocked by the behavior of our friends and family; cases where people turn out to be very different people than their friends and family believed them to be are the exception, not the rule. Such a fact is so mundane that we become habituated to it, and eventually forget about it altogether. When this happens, it is normal for errors in character judgment to appear highly salient. But skeptical arguments based on such errors must be evaluated against a backdrop of overall reliability. Accurate character judgments are common, and they form a baseline for our most meaningful social interactions and life plans.

Acknowledgements

Many people provided feedback on earlier drafts of this paper. Special thanks are due to Maria Altepeter, Michael Barkasi, James Beebe, Peter Carruthers, John Michael, Christian Miller, Jennifer Nagel, Jeremy Reid, Shannon Spaulding, and Haijo Westra for their comments and suggestions. I am also grateful to the participants and organizers of the "Character and the Morally Exceptional: Empirical Discoveries and Moral Improvement" Interdisciplinary Summer Seminar at Wake Forest University, reading groups at the University of Rochester and the University of Toronto, and audience members at the 2018 Central States Philosophical Association conference in Buffalo and the 2019 APA-Central Division conference in Denver.

References

Alfano, M. (2011). Explaining Away Intuitions About Traits: Why Virtue Ethics Seems Plausible (Even if it Isn't). *Review of Philosophy and Psychology*, *2*(1), 121–136. http://doi.org/10.1007/s13164-010-0045-9

Alfano, M. (2013). *Character as Moral Fiction*. Cambridge, UK: Cambridge University Press.

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*(2), 150–166.

Bach, P., & Schenke, K. C. (2017). Predictive social perception: Towards a unifying framework from action observation to person knowledge. *Social and Personality Psychology Compass*, *11*(7), 1–17. http://doi.org/10.1111/spc3.12312

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*(7), 280–289. http://doi.org/10.1016/j.tics.2007.05.005

Biesanz, J. C., West, S. G., & Millevoi, A. (2007). What do you learn about someone over time? The relationship between length of acquaintance and consensus and self-other agreement in judgments of personality. *Journal of Personality and Social Psychology*, *92*(1), 119–135. http://doi.org/10.1037/0022-3514.92.1.119

Butterfill, S., & Apperly, I. (2013). How to Construct a Minimal Theory of Mind. *Mind and Language*, *28*(5), 606–637.

Chaumon, M., Kveraga, K., Barrett, L. F., & Bar, M. (2014). Visual predictions in the orbitofrontal cortex rely on associative content. *Cerebral Cortex*, *24*(11), 2899–907. http://doi.org/10.1093/cercor/bht146

Chen, S. (2003). Psychological-State Theories About Significant Others: Implications for the Content and Structure of Significant-Other Representations. *Personality and Social Psychology Bulletin*, *29*(10), 1285–1302. http://doi.org/10.1177/0146167203255226

Choi, I., & Nisbett, R. E. (1998). Situational salience and cultural differences in the correspondence bias and actor-observer bias. *Personality and Social Psychology Bulletin*, *24*(9), 949–960. http://doi.org/10.1177/0146167298249003

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin*, *125*(1), 47–63. http://doi.org/10.1037/0033-2909.125.1.47

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.

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. http://doi.org/10.1016/j.neuroimage.2011.04.051

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. http://doi.org/10.1037/a0021212

Cronbach, L. J. (1955). Processes affecting scores on" understanding of others" and" assumed similarity.". *Psychological Bulletin*, *52*(3), 177.

Cunningham, W. A., Zelazo, P. D., Packer, D. J., & Van Bavel, J. J. (2007). The iterative reprocessing model: A multilevel framework for attitudes and evaluation. *Social Cognition*, *25*(5), 736–760. http://doi.org/10.1521/soco.2007.25.5.736

Darley, J. M., & Batson, C. D. (1973). " From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, *27*(1), 100–108.

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

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, *37*(7), 1097.

Fast, L. A., & Funder, D. C. (2008). Personality as Manifest in Word Use: Correlations With Self-Report, Acquaintance Report, and Behavior. *Journal of Personality and Social Psychology*, *94*(2), 334–346. http://doi.org/10.1037/0022-3514.94.2.334

Finnigan, K. M., & Vazire, S. (2018). The incremental validity of average state self-reports over global self-reports of personality. *Journal of Personality and Social Psychology*, *115*(2), 321–337. http://doi.org/10.1037/pspp0000136

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. http://doi.org/10.1037//0022-3514.82.6.878

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*(6), 1011–1027. http://doi.org/10.1037/0022-3514.80.6.1011

Fleeson, W., & Gallagher, P. (2009). The Implications of Big Five Standing for the Distribution of Trait Manifestation in Behavior: Fifteen Experience-Sampling Studies and a Meta-Analysis. *Journal of Personality and Social Psychology*, *97*(6), 1097–1114. http://doi.org/10.1037/a0016786

Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, *56*, 82–92.

Fleeson, W., & Jayawickreme, E. (2017). Challenging Doris' Attack on Aggregation: Why We are Not Left "Completely in the Dark" about Global Virtues. *Ethical Theory and Moral Practice*, *20*(3), 519–536. http://doi.org/10.1007/s10677-017-9810-5

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, *68*(1), 5. http://doi.org/10.2307/2024717

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1521).

Funder, D. C. (1995). On the Accuracy of Personality Judgement: A Realistic Approach. *Psychological Review*, *102*(4), 652–670.

Gilbert, D. T., Malone, P. S., Aronson, J., Giesler, B., Higgins, T., Ross, L., … Trope, Y. (1995). The Correspondence Bias. *Psychological Bulletin*, *117*(1), 21–38. http://doi.org/10.1037/0033-2909.117.1.21

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading.* Oxford University Press.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–108. http://doi.org/10.1037/a0028044.Reconstructing

Harman, G. (1999). Moral Philosophy Meets Social Psychology : Virtue Ethics and the Fundamental Attribution Error. *Proceedings of the Aristotelian Society, New Series*, *99*, 315–331.

Hartshorne, H., & May, M. A. (1928). *Studies in the nature of character: Vol. 1. Studies in deceit*. New York, NY: Macmillan Publishers.

Helgeson, C. (2013). The Confirmational Significance of Agreeing Measurements. *Philosophy of Science*, *80*(5), 721–732. http://doi.org/10.1086/673924

Heller, D., Komar, J., & Lee, W. B. (2007). The dynamics of personality states, goals, and well-being. *Personality and Social Psychology Bulletin*, *33*(6), 898–910. http://doi.org/10.1177/0146167207301010

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. http://doi.org/10.1006/jecp.1999.2555

Hohwy, J. (2013). *The predictive mind.* Oxford University Press.

Hutto, D. D. (2004). The limits of spectatorial folk psychology. *Mind and Language.* http://doi.org/10.1111/j.0268-1064.2004.00272.x

Jayawickreme, E., Meindl, P., Helzer, E. G., Furr, R. M., & Fleeson, W. (2014). Virtuous states and virtuous traits: How the empirical evidence regarding the existence of broad traits saves virtue ethics from the situationist critique. *Theory and Research in Education*, *12*(3), 283–308. http://doi.org/10.1177/1477878514545206

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., Vol. 2, pp. 102–138). New York, NY: Guilford.

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

Jussim, L. (2012). *Social Perception and Social Reality: Why Accuracy Dominates Bias and Self-Fulfilling Prophecy*. New York, NY: Oxford University Press.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321. http://doi.org/10.1111/j.1467-7687.2007.00585.x

Kenny, D. A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review*, *98*(2), 155–163. http://doi.org/10.1037/0033-295X.98.2.155

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. http://doi.org/10.1093/scan/nss022

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. *Neuron*, *79*(5), 836–848. http://doi.org/10.1016/j.neuron.2013.08.020

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. http://doi.org/10.1016/j.jesp.2007.05.004

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*. http://doi.org/10.1111/jopy.12413

Ma, N., Vandekerckhove, M., Baetens, K., Overwalle, F. Van, Seurinck, R., & Fias, W. (2011). Inconsistencies in spontaneous and intentional trait inferences. *Social Cognitive and Affective Neuroscience*, *7*(8), 937–950. http://doi.org/10.1093/scan/nsr064

Ma, N., Vandekerckhove, M., Baetens, K., Overwalle, F. Van, Seurinck, R., & Fias, W. (2012). Inconsistencies in spontaneous and intentional trait inferences. *Social Cognitive and Affective Neuroscience*, *7*(8), 937–950. http://doi.org/10.1093/scan/nsr064

McCabe, K. O., & Fleeson, W. (2012). What Is Extraversion For? Integrating Trait and Motivational Perspectives and Identifying the Purpose of Extraversion. *Psychological Science*, *23*(12), 1498–1505. http://doi.org/10.1177/0956797612444904

McCabe, K. O., & Fleeson, W. (2016). Are traits useful? Explaining trait manifestations as tools in the pursuit of goals. *Journal of Personality and Social Psychology*, *110*(2), 287–301. http://doi.org/10.1037/a0039490

McCrae, R. R., Costa, P. T., Ostendorf, F., Angleitner, A., Hrebickova, M., Avia, M. D., …

HrebfEkova, M. (2000). Nature over nurture: temperament, personality, and life span development. *Journal of Personality and Social Psychology*, *78*(1), 173–186.

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

Meltzoff, A. N., & Gopnik, A. (2013). Learning about the mind from evidence: Children's development of intuitive theories of perception and personality. *Understanding Other Minds*, 19–34. http://doi.org/10.1103/PhysRevLett.115.197401

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, *8*(6), 623–631. http://doi.org/10.1093/scan/nss040

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, *67*(4), 371.

Miller, C. B. (2014). *Character and moral psychology*. Oxford: Oxford University Press.

Mischel, W. (2004). Toward an Integrative Science of the Person. *Annual Review of Psychology*, *55*(1), 1–22. http://doi.org/10.1146/annurev.psych.55.042902.130709

Mischel, W., & Shoda, Y. (1995). A cognitive-active system theory of personality: Reconceptualising situations, dispositions, dynamics and invariance in personality structure. *Psychological Review*, *12*(1), 246–268.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds.* Oxford: Oxford University Press.

Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: big five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology*, *93*(1), 116.

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the Prediction of Consequential Outcomes. *Annual Review of Psychology*, *57*(1), 401–421. http://doi.org/10.1146/annurev.psych.57.102904.190127

Paris, P. (2016). Scepticism about Virtue and the Five-Factor Model of Personality. *Utilitas*, *29*(4), 1–30. http://doi.org/10.1017/S0953820816000327

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302–21. http://doi.org/10.1016/j.cognition.2010.11.015

Reeder, G. D. (2009). Mindreading: Judgments About Intentionality and Motives in Dispositional Inference. *Psychological Inquiry*, *20*(1), 1–18. http://doi.org/10.1080/10478400802615744

Reeder, G. D., Kumar, S., Hesson-McInnis, M. S., & Trafimow, D. (2002). Inferences about the morality of an aggressor: the role of perceived motive. *Journal of Personality and Social Psychology*, *83*(4), 789–803. http://doi.org/10.1037/0022-3514.83.4.789

Reeder, G. D., Vonk, R., Ronk, M. J., Ham, J., & Lawrence, M. (2004). Dispositional Attribution: Multiple Inferences About Motive-Related Traits. *Journal of Personality and Social Psychology*, *86*(4), 530–544. http://doi.org/10.1037/0022-3514.86.4.530

Repacholi, B. M., Meltzoff, A. N., Hennings, T. M., & Ruba, A. L. (2016). Transfer of Social Learning Across Contexts: Exploring Infants' Attribution of Trait-Like Emotions to Adults. *Infancy*, *21*(6), 785–806. http://doi.org/10.1111/infa.12136

Repacholi, B. M., Meltzoff, A. N., Toub, T. S., & Ruba, A. L. (2016). Infants' Generalizations About Other People' s Emotions : Foundations for Trait-Like Attributions. *Developmental Psychology*, *52*, 364–378. http://doi.org/dx.doi.org/10.1037/dev0000097

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. New York, NY: Holt, Rinehart and Winston Inc.

Ross, L. (1977). The Intuitive Psychologist And His Shortcomings: Distortions in the Attribution Process. *Advances in Experimental Social Psychology*. http://doi.org/10.1016/S0065-2601(08)60357-3

Ross, L., & Nisbett, R. E. (2011). *The person and the situation: Perspectives of social psychology*. Pinter & Martin Publishers.

Sabini, J., & Silver, M. (2005). Lack of Character? Situationism Critiqued. *Ethics*, *115*(3), 535–562. http://doi.org/10.1086/428459

Seiver, E., Gopnik, A., & Goodman, N. D. (2013). Did She Jump Because She Was the Big Sister or Because the Trampoline Was Safe? Causal Inference and the Development of Social Attribution. *Child Development*, *84*(2), 443–454. http://doi.org/10.1111/j.1467-8624.2012.01865.x

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

Sripada, C. S. (2012). Mental state attributions and the side-effect effect. *Journal of Experimental Social Psychology*, *48*(1), 232–238. http://doi.org/10.1016/j.jesp.2011.07.008

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

Tamir, D. I., & Thornton, M. A. (2018). Modeling the Predictive Social Mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. http://doi.org/10.1016/j.tics.2017.12.005

Theriault, J., & Young, L. (2018). *Social Prediction in the Theory of Mind Network*. PsyArXiv.

Thornton, M. A., & Mitchell, J. P. (2018). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex*, *28*(10), 3505–3520. http://doi.org/10.1093/cercor/bhx216

Todorov, A. (2013). Making up your mind after 100-ms exposure to face. *Psychological Science*, *17*(7), 592–598.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology*, *66*, 519–545. http://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. http://doi.org/10.1016/j.tics.2008.10.001

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors faces. *Journal of Experimental Social Psychology*, *39*(6), 549–562. http://doi.org/10.1016/S0022-1031(03)00059-3

Uhlmann, E. L., Zhu, L. L., & Diermeier, D. (2014). When actions speak volumes: The role of inferences about moral character in outrage over racial bigotry. *European Journal of Social Psychology*, *44*(1), 23–29. http://doi.org/10.1002/ejsp.1987

Uleman, J. S., Adil Saribay, S., & Gonzalez, C. M. (2008). Spontaneous Inferences, Implicit Impressions, and Implicit Theories. *Annual Review of Psychology*, *59*(1), 329–360. http://doi.org/10.1146/annurev.psych.59.103006.093707

Uleman, J. S., Hon, A., Roman, R. J., & Moskowitz, G. B. (1996). On-line evidence for spontaneous trait inferences at encoding. *Personality and Social Psychology Bulletin*, *22*(4), 377–394. http://doi.org/10.1177/0146167296224005

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*(3), 829–858. http://doi.org/10.1002/hbm.20547

Vazire, S., & Carlson, E. N. (2011). Others sometimes know us better than we know ourselves. *Current Directions in Psychological Science*, *20*(2), 104–108. http://doi.org/10.1177/0963721411402478

Vazire, S., Chung, C., Freeman, H., Mehta, P., Baquero, C., Harrison, H., … Beard, S. (2010). Who Knows What About a Person ? The Self – Other Knowledge Asymmetry (SOKA) Model. *Journal of Personality and Social Psychology*, *98*(2), 281–300. http://doi.org/10.1037/a0017908

Watson, G. (1975). Free agency. *Journal of Philosophy*, *72*, 205–220.

Webber, J. (2007). Character, common-sense, and expertise. *Ethical Theory and Moral Practice*, *10*(1), 89–104.

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

Westra, E. (2017, September 26). Stereotypes, theory of mind, and the action–prediction hierarchy. *Synthese*, pp. 1–26. Springer Netherlands. http://doi.org/10.1007/s11229-017-1575-9

Westra, E. (2018). Character and theory of mind: an integrative approach. *Philosophical Studies*, *175*(5), 1217–1241. http://doi.org/10.1007/s11098-017-0908-3

Whewell, W. (1989). Novum organon renovatum. In R. E. Butts (Ed.), *In William Whewell: Theory of*

Scientific Method,. Indianapolis: Hackett.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128. http://doi.org/10.1016/0010-0277(83)90004-5

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

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