Character and theory of mind: An integrative approach

Evan Westra
(Forthcoming in *Philosophical Studies*)

Abstract: Traditionally, theories of mindreading have focused on the representation of beliefs and desires. However, decades of social psychology and social neuroscience have shown that, in addition to reasoning about beliefs and desires, human beings also use representations of character traits to predict and interpret behavior. While a few recent accounts have attempted to accommodate these findings, they have not succeeded in explaining the relation between trait attribution and belief-desire reasoning. On the account I propose, character-trait attribution is part of a hierarchical system for action prediction, and serves to inform hypotheses about agents' beliefs and desires, which are in turn used to predict and interpret behavior.

1. Introduction

As highly social beings, we need to be able to rapidly predict and interpret the behavior of those around us in order to thrive. We do this, the usual explanation goes, by reasoning about the unobserved representational states that cause behavior – a process variously referred to as *theory of mind*, *mindreading*, and *folk psychology*. Standard models of mindreading, such as the theory-theory, the simulation theory, and various hybrid models, tend to focus especially on how we predict and interpret behaviors in terms of beliefs and desires. This focus is epitomized by the field's longstanding fascination with the false-belief task, which is used to measure children's understanding of the representational nature of belief (Onishi and Baillargeon 2005; Rakoczy 2015; Wellman et al. 2001; Wimmer and Perner 1983). As a result, questions about the developmental, cognitive, and evolutionary underpinnings of belief-reasoning tend to dominate social cognition research.

Due to this narrow focus on beliefs and desires, other conceptual tools that we use to interpret behavior are often ignored. One such tool is character-trait attribution: the explanation and prediction of behavior in terms of enduring internal properties of individuals that lead to stable behavioral tendencies. This omission from the theory-of-mind

literature is quite curious: one of the most robust findings in social psychology research is that we often interpret behavior on the basis of stable personality traits (D. L. Ames et al. 2011; Gilbert et al. 1995). Character also figures prominently in moral philosophy (Anscombe 1958; Foot 1967; Miller 2013), and has begun to garner attention in empirical moral psychology research as well (Sripada 2012; Uhlmann et al. 2015). Yet in spite of its presence in neighboring disciplines and a large body of data on the subject, character-trait attribution does not figure in classic and contemporary theories of mindreading.

My goal in this paper is to provide a framework for integrating our understanding of character-trait attribution with other aspects of theory of mind. I will propose that we use representations of a person's stable character traits to infer which hypotheses about that person's more transient mental states – namely, their beliefs, goals, and intentions – are more probable; we then use these mental-state hypotheses to directly predict their behavior. Trait attribution thus forms the upper level of an action-prediction hierarchy, wherein the hypotheses at higher levels inform the hypotheses at lower levels. Feedback from observable behavior then leads us to make revisions to our mentalistic hypotheses, which might occur at either the belief-desire levels or at the level of character traits. This basic inferential structure is best understood in terms of a hierarchical Bayesian model of cognition.

In section 2, I will briefly review part of the empirical literature on the attribution of character traits, and the role that these representations play in predicting and interpreting behavior. In section 3, I will discuss recent "pluralist" accounts of folk psychological reasoning (Andrews 2008, 2012; Fiebich and Coltheart 2015), which *do* acknowledge the role of character-trait attribution in folk psychology, but fail to explain its relationship to other forms of mindreading. In sections 4 and 5, I will outline an account in which character-trait attribution stands in a systematic, hierarchically structured relation to belief and desire attribution. In section 6, I suggest several ways to empirically test this account, as well as ways to apply it to other, related domains.

First, however, a word about how we think of character traits. Like beliefs and desires, character traits are believed to be causally related to behavior, and it is not uncommon to explain behavior by referring to a character trait (e.g. "She turned in the lost wallet because she is an honest person") (Malle 2004). Some traits, such as selfishness and greed, seem to

possess a strong volitional element, and thus seem closely related to desires. Others, such as intelligence, cleverness, and gullibility, seem distinctively epistemic, and thus more related to beliefs. However, traits also differ from beliefs and desires in several important ways. First, beliefs and desires figure in practical reasoning, and lead directly to action. Character traits, on the other hand, do not seem to figure in practical reasoning, and it is less clear how they translate into particular actions. Second, beliefs and desires can easily change. When we acquire new relevant information we regularly change our beliefs; when we successfully act out our plans, our desires and goals are fulfilled. Character traits, conversely, are much more persistent. They do not update or go away as the result of individual actions, but rather last through significant portions of an agent's lifetime. Third, beliefs and desires can be about particular states of affairs. Character traits, on the other hand, relate to the world in a very general way, and tend to be relevant across a wide range of situations. In short, character traits are temporally stable mental properties that relate to action in an opaque general manner across a wide range of situations.[1]

Citing evidence from social psychology, some philosophers have questioned whether people actually possess character traits as we ordinarily think of them (Doris 2002; Harman 1999). These arguments begin with findings showing that subtle manipulations in situational factors lead to dramatic effects on behavior. For instance, Isen and Levin showed that finding a dime in a phone booth makes people much more likely to help a stranger pick up dropped papers – a finding that seems to show that "generosity" is, contrary to common belief, a fickle, variable trait (Isen and Levin 1972). Based on these and other experimental results, "situationists" have argued that it is situations, and not stable character traits, that really cause our behavior. These arguments have sparked a great deal of controversy, and a number of philosophers have mounted defenses of the reality of character traits(Miller 2013; Sabini and Silver 2005; Sreenivasan 2002).

---

[1] John Doris offers a similar, though not identical, analysis of character traits. According to his view, "global" traits have two primary features:

1. *Consistency*. Character and personality traits are reliably manifested in trait-relevant behavior across a diversity of trait-relevant eliciting conditions that may vary widely in their conduciveness to the manifestations of the trait in question.
2. *Stability*. Character and personality traits are reliably manifested in trait-relevant behavior over iterated trials of similar trait-relevant eliciting conditions. (Doris 2002, p. 22)

Doris also mentions a third feature of character traits, *evaluative integration*, which is not relevant for our current purposes.

The situationism debate is about the metaphysical reality of character traits, and whether stable character traits ought to figure in mature scientific explanations of behavior; situationists think that insofar as character traits exist, they are situation-dependent, and that stable character traits have no real explanatory value for psychology. This is not a paper about the metaphysical reality of character traits, however. Rather, it is about people's *representations* of character traits, and the role that these representations play in *folk-psychological inference*. Notoriously, folk reasoning about the world is often inaccurate, and frequently invokes entities that do not stand up to scientific scrutiny. For instance, when reasoning about the motion of objects, even educated adults seem to rely on a quasi-medieval impetus principle, and predict that (for example) a ball spinning around the end of a string will continue to follow a curved path when it is released (McCloskey et al. 1980). Impetus principles no longer play any role in our mature physics, but they still play a role in folk physics. Likewise, stable character traits might have no place in our mature psychology, but they clearly still play a role in our folk psychology. Thus, the situationists might be right that our behavior is never caused by stable character traits, but they could still allow that representations of stable character traits play an important role in our folk psychology. A dedicated situationist could thus happily accept the foregoing account as an error theory explaining why we think people have stable character traits, even though there are none. The two views are, in principle, mutually compatible.

However, while the present account is not committed to the existence of stable character traits as such, it does imply that our representations of character traits have *some* predictive value; otherwise, their prominent role in our cognitive economy would be mysterious. The most obvious explanation for this predictive role would be that stable personality traits do in fact exist, despite the evidence cited by the situationist. Another explanation would be that trait representations serve as a kind of inferential heuristic: even if they fail to track anything real in the world, they may earn their predictive keep by conferring some sort of information-processing benefit on other socio-cognitive processes (such as belief-desire reasoning). Along these lines, I suggest that one function of trait attributions is to provide us with initial prior probability distributions over mentalistic hypotheses, which then undergo

further updating in response to experience. Another explanation would be that trait representations roughly track *something* in the world – just not character traits. Like the impetus principle, which roughly tracks the real physical principle of inertia (but systematically errs in certain cases), it may be that our trait representations roughly correspond to some predictively relevant property of the social environment, which we *systematically misrepresent* as stable character-traits. In this vein, I suggest in section 6 that our trait attributions may sometimes track relational social properties such as status and intergroup threat.

In order for my account to be right, at least one of these explanations must be correct. It could be that all of them are right: perhaps our trait attributions sometimes track real personality traits, while also tracking relational social properties, while simultaneously conferring an information-processing benefit on our overall action-predictions. However, I need not take a strong stand on this issue at this time. All that matters for my current purposes is the role that trait information plays in the structure of mentalistic action-prediction.

2. Impression formation and mindreading

In social psychology and socio-cognitive neuroscience, reasoning about character traits is most often referred to as 'impression formation' and 'person perception' (D. L. Ames et al. 2011; Trope and Gaunt 2007). While we attribute a wide range of specific traits to others, it appears that the kinds of traits we appeal to tend to be organized along two particular social dimensions: warmth and competence (Fiske et al. 2007). The warmth dimension captures attributions of traits such as friendliness, sincerity, trustworthiness, and seems to track whether we expect an individual to be positively or negatively disposed towards us. The competence dimension, in contrast, captures attributions of traits such as intelligence, impulsivity, and social dominance, and seems to track our estimation of an individual's ability to successfully achieve their goals. When trait attributions are analyzed in terms of these two dimensions, they are highly predictive of our reactive attitudes towards both individuals and groups (Cuddy et al. 2007), even across a wide range of cultures (Cuddy et al. 2009).

Interestingly, these two trait dimensions do not just emerge in people's judgments of individuals: they also emerge in stereotypes about social groups. For instance, common anti-Semitic stereotypes tend to invoke low warmth traits, such as deceptiveness and miserliness, but also high competence traits, such as intelligence. Misogynistic stereotypes, in contrast, invoke high warmth traits, such as helpfulness, but also low-competence traits, such as frivolity and superficiality. Stereotypes about very low status groups, such as the homeless, tend to contain low competence traits, such as stupidity and laziness, and low warmth traits, such as dishonesty. Both social in-groups (e.g. students if one is a student) and societal prototype groups (in Western cultures, the White middle class) tend to be rated as both high competence and high warmth (Cuddy et al. 2009; Fiske 2015; Fiske et al. 2007). This suggests that we may use a person's social group membership as a source of evidence about her character traits (see also Fiebich and Coltheart (2015)).

Many of the methods used to study trait attributions involve explicit, linguistically based measures (e.g. Ross 1977); however, character-trait attributions can also be extremely rapid and unconscious. In particular, we seem to use low-level perceptual cues such as facial appearance to make character-trait attributions within 100 milliseconds of encountering someone (Bar et al. 2006; Todorov 2013). Incredibly, these rapid, perceptually based trait attributions also vary along the dimensions of warmth and competence (or, as they are referred to in this literature, trustworthiness and dominance). In particular, neutral-expression faces with wider jaws, heavier brows, and smaller eyes tend to be judged as more dominant, while "baby faces" tend to be judged as less dominant; similarly, neutral-expression faces with downturned brows and lips tend to be judged as less trustworthy, while faces with high brows and upturned lips tend to be judged as more trustworthy (Todorov et al. 2008). Thus, from the first second that we encounter someone, we begin to form a representation of his or her character traits along the warmth and competence dimensions.

Of course, we do not infer personality traits from appearance alone: we also use a person's behavior and second-hand information to inform our representations of their character. The most well-known finding in this vein is that we tend to over-attribute the causes of behavior to underlying traits or dispositions rather than situational factors – a phenomenon known as

the 'correspondence bias' or the 'fundamental attribution error' (Gawronski 2004; Gilbert et al. 1995; E. Jones and Harris 1967; Ross 1977). For instance, when participants passively observe one confederate quizzing another, they tend to rate the questioner as more intelligent than the test-taker, even though the questioner has clearly been provided with the answers, while the test-taker has not (Ross 1977). This bias can be mitigated by prompting participants to explicitly attend to situational factors (e.g. that the questioner has been provided with all the answers, whereas the test-taker has not); however, participants will default to the correspondence bias when placed under cognitive load, even if the very same situational information is made explicitly available (Gilbert et al. 1988, 1995). This suggests that the correspondence bias is the product of a relatively efficient cognitive process, while correcting it requires cognitive control.[2]

Much of the research on the correspondence bias has occurred separately from research on mindreading, focusing instead on the distinction between situation-based and disposition-based explanations of behavior. But some social psychologists have begun to explore the connection between representations of traits and other mental states, such as intentions. For instance, Krull and colleagues found that participants were less likely to exhibit a correspondence bias when an actor performed a helpful action if the actor showed signs of unwillingness (Krull et al. 2008). Likewise, a number of authors have found that the correspondence bias was attenuated when participants were given reason to think that a given action may have been performed for an ulterior motive (D. R. Ames et al. 2004; Fein 1996; Reeder et al. 2004). Hooper and colleagues also found that participants who were primed to engage in explicit perspective-taking displayed a diminished correspondence bias compared to a control group (Hooper et al. 2015). Thus, thinking about mental states seems to mediate inferences from behavior to character (Reeder 2009).

[2] There are also cross-cultural differences in the extent to which individuals fall prey to the correspondence bias. While the correspondence bias is present to some extent across cultures (Choi et al. 1999; Krull et al. 1999; Norenzayan et al. 2003), it appears that members of "individualist" societies are particularly susceptible to it; meanwhile, members of "collectivist" societies seem to pay more attention to situational factors and the presence of social constraints (Choi and Nisbett 1998; Miyamoto and Kitayama 2002). This is consistent with the broad finding that members of collectivist cultures display a habitual tendency to attend to situational factors and contexts (Kitayama et al. 2003). These habitual patterns of attention seem to make members of "collectivist" cultures better able to correct their initial dispositionalist attributions.

This relationship between trait attribution and other forms of mental-state attribution is reflected in the neural correlates of both processes, which overlap substantially and appear to be functionally related. Many neuroimaging studies have confirmed the existence of a distinctive network of brain regions that are consistently recruited when we reason about the thoughts and behavior of others: the temporal-parietal junction (TPJ), the posterior superior temporal sulcus (pSTS), the medial prefrontal cortex (mPFC), the precuneus (PC), and the temporal poles (TP) (Van Overwalle 2009). All of these regions are implicated in impression formation and updating, both under intentional and spontaneous conditions (Cloutier et al. 2011; Ferrari et al. 2016; Harris et al. 2005; Hassabis et al. 2013; Kestemont et al. 2013; Ma et al. 2011). The dorsal region of the mPFC (dmPFC) in particular seems to be centrally implicated in the representation of stable personality traits (Ferrari et al. 2016; Ma et al. 2012). When subjects are explicitly prompted to reason about traits, this region is highly active; in contrast, when subjects are prompted to reason about "situational" factors, this region is less active, while regions associated with goal and belief attribution, such as the TPJ and the pSTS, are more so (Kestemont et al. 2013). However, when subjects learn that a person holds a belief or performs an intentional action that is inconsistent with a previously formed impression, both the TPJ and the dmPFC show increased activity (Cloutier et al. 2011; Ma et al. 2011). Thus, the neuroimaging data, like the behavioral data, suggest that mental state information interacts with the trait attribution process.

There is also some indication that representations of character traits can bias our mental-state attributions. This evidence comes from a debate about how to interpret the side-effect effect, which is when participants seem to over-attribute intentionality and blame to agents whose actions have negative (but not positive) side-effects (Knobe 2003). Chandra Sripada has suggested that this effect may be driven by an initial negative judgment of the agent's character, or "Deep Self" (Sripada 2009). According to this view, participants incorrectly judge that the agent intentionally caused a particular outcome because this intentionality attribution seems to follow from their previous impression of the agent's character; in other words, they interpret the agent's actions (and their consequences) as flowing from their deeper personality traits. To test this theory, Sripada asked participants who had completed a side-effect effect task to give an explicit evaluation of the agent's character and core values;

sure enough, these predicted their intentionality judgments (Sripada and Konrath 2011; Sripada 2012).

To summarize: upon first encountering an individual, we rapidly construct a representation of their character that is especially sensitive to particular trait-dimensions, namely warmth and competence. We use various sources of information to update this representation, and are biased towards interpreting behavior as reflecting stable character traits. However, these inferences are mediated by inferences about mental states: when information about the motivations and beliefs of others is available, we update or refrain from updating our character models accordingly. Moreover, background knowledge about character also seems to affect our intentionality attributions, suggesting that we expect not just behavior, but also intentions to accord with character. Inferences about character and inferences about mental states, in short, appear to inform one another.

3. Character-trait attribution and theories of folk psychology

Traditional accounts of mindreading, such as the simulation theory and the theory-theory, have not typically addressed how we attribute character traits. The notion of character seems particularly hard to integrate into a simulation-based account. According to the simulation theory, if an interpreter has enough information about another agent's beliefs and desires, she can make a successful behavioral prediction by simulating in an offline manner how she would behave if she had those same beliefs and desires (Goldman 2006; Heal 1996). But it is not at all clear how this strategy could extend to trait attribution. Character traits are not the sorts of things that could figure in practical reasoning.[3] Any effect of character on practical reasoning is bound to be oblique: it may affect the kinds of beliefs and desires we form in the first place, the extent to which we deliberate before acting, or the relative importance that we assign to particular desires. Thus, it is not clear where – if anywhere – character traits could fit into a pure simulationist account.

[3] Beliefs about one's own character traits could figure in practical reasoning. But this observation is of little help to the simulation-theorist: surely, this kind of self-reflection is uncommon in the first-person case, and it would be bizarre if we nevertheless believed that other people frequently engage in it. Moreover, beliefs about one's character seem like they would have a very different effect on behavior than character itself. If I reflect on my own impulsivity, for instance it will probably lead me to be less impulsive.

The only plausible option for the simulation theorist would be to endorse a hybrid account. Instead of holding that character enters into the simulation process itself, a hybrid simulation/theory-theorist could hold that character information is used to infer the *inputs* to the simulation procedure. This solves the simulationist's character problem, but only at the cost of conceding that simulation theory is poorly equipped for reasoning about traits. It also raises a new question: how would a theory-theorist explain the effects of character on practical reasoning?

Fortunately, the theory-theory seems better equipped to deal with trait attribution. Traditional theory-theory accounts focus on generalizations about how beliefs and desires combine to produce behavior, treating both as underlying causal variables (Gopnik and Wellman 2012; Wellman 2014). Quite conceivably, character traits could be treated as another kind of underlying variable, albeit one that has a much less direct effect on behavior than beliefs and desires. But any such account would need to do more than just posit an additional variable: it would also have to tell us just how traits relate to mental states, and how they help us to predict behavior. The account that I propose in this paper, which could be construed as a version of the theory-theory, will attempt to do just this.

There is, however, one group of mindreading theorists that has already explicitly addressed the role of character-trait attribution in action prediction and interpretation: the folk-psychology pluralists (Andrews 2008, 2012; Fiebich and Coltheart 2015). The basic goal of the pluralists is to offer an alternative to simulation theory and theory-theory accounts that invokes a broader set of procedures and representations. Andrews (2008, 2012) suggests that in addition to belief-desire reasoning, we also use social norms, stereotypes, situation-based schemas, and trait attribution to predict and explain behavior. Likewise, Fiebich and Coltheart (2015) propose that we use trait-based inferences to predict and interpret behavior, in addition to theory-based and simulation-based procedures. They also situate these various socio-cognitive strategies within a two-systems framework[4] (Apperly and Butterfill 2009; Kahneman 2011). Thus, when predicting and interpreting another agent's behavior, we may use either System 1 or System 2 versions of simulation, theory, or trait attribution.

[4] System 1 strategies are "fast, relatively effortless routines that occur without our awareness," while System 2 strategies are "slow routines which require the expenditure of mental effort and are subject to consciousness and deliberative control" (Fiebich and Coltheart 2015, p. 238).

While they differ in several respects, these pluralist accounts treat reasoning about behavior in terms of character traits as a socio-cognitive alternative to belief-desire predictions and explanations. Even if an agent did not possess the concepts of BELIEF and DESIRE, according to the pluralists, they might still successfully predict behavior by using their knowledge of a target's traits. This is possible, the pluralists argue, because trait-based interpretations rely on associations between behaviors and situations. For instance, a trait like generosity might form the central node[5] in a network of behavior-situation pairings: *leaving large tips* and *restaurants*, *carrying heavy boxes* and *friends moving house*, and so on. These associative networks would enable agents to attribute traits whenever an individual demonstrated one of the relevant behavior-situation pairings, and then use this information to predict that individual's behavior in other situations in which generosity might be possible.

While the pluralists should be given due credit for emphasizing a role for traits in our folk psychology, this particular approach to trait attribution has two important limitations. The first is that the predictive utility of trait-based reasoning will depend heavily on how 'situations' get represented. If trait-behavior associations are tied only to situations that we have previously experienced, then it will be inert whenever we encounter a novel situation. Andrews (2008) recognizes this fact, but suggests that it is not a big problem, because we spend most of our time in relatively familiar situations. But this reply raises an important question: how do we parse situations for the purpose of trait-based predictions? If we parse situations at a fairly coarse level, then Andrews might be right; however, this would make the corresponding predictions far less reliable, since they would be insensitive to important situational differences. For instance, if one forms the association between *leaving large tips* and *restaurants*, then we would predict that a generous person would leave a large tip even when

---

[5] Fiebich & Coltheart distinguish between non-linguistic trait attributions and linguistic trait attributions. Non-linguistic trait attributions occur when an agent does not possess a linguistic concept of a trait (i.e. the word 'generosity'). These only consist in associations between particular behaviors, situations, and agents, and would only allow for predicting similar behaviors in similar situations. Linguistic trait attributions, in contrast, involve the possession of a linguistic concept of a trait, and would facilitate a whole network of predictions.
I am skeptical of this distinction for two reasons. First, non-linguistic trait attribution, on this account, does not seem to involve trait-based reasoning at all: traits are supposed to be enduring, internal properties of individuals, but these non-linguistic trait attributions seem to consist only in superficial behavioral associations. Second, this distinction implicitly assumes that the only way to possess a concept of a trait is through language. But there is ample reason to think that even pre-linguistic or non-linguistic entities can possess concepts (e.g. Call and Tomasello 2008; Carey 2009). While linguistic concepts undoubtedly enrich and expand our trait attribution abilities, there is no reason to think that non-linguistic trait attribution is as impoverished as Fiebich and Coltheart (2015) make it out to be.

she has received poor service, or when a friend is treating her. Conversely, if situations are parsed very finely, then we will treat otherwise familiar situations as novel, given a very slight change. Thus, we might expect a generous person to leave large tips in *sushi restaurants with good lighting*, but not in *sushi restaurants with bad lighting*. In this case, most trait attributions would be predictively inert. Unless we parse situations just right, in other words, the pluralist strategy will either lead to inaccurate overgeneralizations, or inflexible under-generalizations.

Some pluralists may simply wish to concede the limited reliability of trait attributions. Andrews (2008) suggests that when we make inaccurate predictions, we may simply respond by forgetting them, or by giving a post-hoc explanation of our failures, and then carry on with our business. This is not a problem, the pluralist argues, because we have lots of different strategies for folk-psychological prediction: when trait attribution fails us, we may simply try a different one. This response is unsatisfying: even if trait-based reasoning is often inaccurate, it seems that one should be able to learn from these inaccuracies in order to inform future predictions, rather than simply discard them. Indeed, the evidence reviewed in the previous section indicates that we actually pay close attention when our trait-based expectations are violated. But if traits were truly an unreliable way to predict behavior, then why would we continue to track character information? If the pluralist story about trait attribution is correct, then this seems like a bizarre use of limited cognitive resources.

The second, related limitation of the pluralist account of character traits is that it cannot explain the empirical relation between trait attribution and mental-state attribution. As we saw in the previous section, these two forms of reasoning seem to be causally interrelated, both at the behavioral and neural levels. But on the pluralist account of trait reasoning, mental state information is never involved. This is by design: the pluralist's goal is to show that behavioral prediction and interpretation can happen in the absence of mental-state attribution. Indeed, Andrews (2008) argues that there is really a double dissociation between belief-attribution and trait-based reasoning. First, while children are able to reason explicitly about beliefs from an early age, they do not explicitly mention traits in their explanations and predictions of behavior until much later (Kalish 2002). Thus, it is possible to reason about mental states even if one cannot reason about character traits. Second, interventions for children with autism (who lack the ability to reason about beliefs) often rely upon training

children to associate traits and behaviors, such as the term 'happy' with smiling and laughing (Gray 2007). Thus, one can also reason about traits without being able to reason about mental states.

There are a few problems with this argument. First, there is now positive evidence that three to four year-old children respond in an adult-like manner to facial features associated with warmth and competence, despite the fact that they do not refer to such traits in their explanations of behavior (Cogsdill et al. 2014). Second, the autism intervention Andrews describes does not seem to be about trait-reasoning at all, but rather reasoning about emotions. But even if it were an instance of trait-reasoning, this would then be an exception that proves the rule: in the absence of the capacity to reason about beliefs, it seems that children with autism are only able to use trait information through explicit laborious training, whereas it comes naturally to neurotypical individuals.

However, even if we were to accept Andrews' double dissociation argument in its entirety, all it would show is that character reasoning and belief-desire reasoning are not *identical*, and that neither is *necessary* for the other. But these are only the strongest possible relations that could hold between these two processes. Even if, in principle, there could be double dissociations between character reasoning and belief-desire reasoning, it might still be true that the two processes are causally and functionally intertwined.

I suggest that the solution to the pluralist's first problem may lie in developing an answer to the second. What the pluralist proposal lacks is a principled basis for parsing situations for the purpose of behavioral prediction. But if we consider an agent's beliefs and desires, the solution to this problem is obvious: the 'situation' will consist in those features of the local context that the agent believes are relevant to her goals. Moreover, this approach would facilitate predictions even in highly unfamiliar situations. This is because mental-state reasoning is a highly flexible, generative framework for predicting and interpreting behavior. By employing causal models that treat mental states as variables that can take on a broad range of values, mentalistic reasoning is capable of generating behavioral predictions about an indefinitely large number of novel situations, even if they have yet to be encountered (Christensen and Michael 2015). Thus, the predictive link between traits and behaviors postulated by pluralists only makes sense when it is mediated by belief-desire reasoning,

because belief-desire reasoning provides us with a principled basis for parsing situations. However, this leaves us with the same lingering question that we started with: how are trait representations related to representations of beliefs and desires?

4. The action prediction hierarchy

In this section, I introduce hierarchical predictive coding accounts of cognition, and how they have been applied to theory of mind research. This will lay the groundwork for my positive account of how representations of character relate to belief-desire reasoning. I begin with a brief digression about the nature of mirror neurons. The purpose of this digression will be to illustrate how predictive-coding approaches are poised to explain the cognitive underpinnings of action prediction. In particular, these accounts posit that we represent intentional states in a hierarchical fashion, and that mirror-neuron activity reflects the way we exploit this hierarchy when predicting intentional actions. Ultimately, I will argue that our representations of character are a part of this action-prediction hierarchy.

4.1. Mirror neurons and action hierarchies

Mental states vary with respect to their temporal stability. For instance, some desire-like states, such as motor intentions – the intention to make a particular bodily movement, such as reaching or grasping – are highly transient. We also chain together many individual motor intentions in order in order to fulfill particular action-goals, as when we walk across a room to pick up a tool; these goals last longer than the individual motor intentions that comprise them, but are still extinguished relatively quickly. Many of these action-goals can be chained together to achieve more complex, temporally extended goals, such as building a house or fixing a car. These broader goals can in turn serve as sub-goals for still larger projects, and so on. Desire-like states, in other words, seem to form temporal hierarchies: more stable goals are comprised of more transient sub-goals, which are comprised of even more transient sub-sub-goals, eventually bottoming out in very low-level motor plans.

This property of desire-like states has not gone unnoticed by mindreading theorists. In particular, it has caught the attention of several authors who were unsatisfied with the standard, "direct-mapping" interpretation of mirror-neuron activity endorsed by simulation

theorists (Gallese and Goldman 1998). According to this standard interpretation, when we observe the low-level visual properties of another agent's movements, we automatically form an offline representation of those actions in the pre-motor cortex, where we normally represent our own action plans. Based on this representation, the story goes, we are then able to deduce the higher-order intentions that would have caused this action plan, and thereby infer the agent's goals, effectively using our own motor planning system in reverse (Jeannerod et al. 1995; Rizzolatti and Craighero 2004).

There is a big problem with this account: the inference from low-level visual properties of behavior to goals is vastly under-determined. This is because a single behavior is, in principle, compatible with a wide range of underlying motivations. One could, for instance, raise one's hand with an open palm because one is about to salute, to give a high five, to wave in greeting, to tell someone to stop, or to deliver a slap. The same behavioral effect, in other words, could have indefinitely many different mental causes (a predicament known as an 'inverse problem') (Csibra 2008; Jacob and Jeannerod 2005; Kilner et al. 2007). This means that for any given behavior that the mirror neuron system represents, we must sort through an indeterminately large space of possible goals that might have caused it. Thus, the direct-mapping account quickly leads to computational intractability.

This observation has led several authors to argue that mirror-neuron activity does not reflect a bottom-up mapping from motor-intentions to goals, but rather a top-down prediction about an agent's likely behaviors based on a prior hypothesis about its goals – what Csibra (2008) calls the 'predictive action monitoring hypothesis'. This solves the aforementioned under-determination problem, because all it involves is *checking* whether an observed behavior would be made likely given a hypothesized goal, rather than *solving* for a unique goal from an ambiguous behavior. And since goals are a more abstract kind of representation than motor intentions, they tend to be consistent with a fairly wide range of more concrete physical behaviors. For instance, the goal of eating an apple makes a number of behaviors more likely: reaching over to grab the apple and bringing it to one's mouth, or reaching over to grab an apple and then grabbing a knife to peel it, or using the knife to cut it into wedges, etc. Our action-prediction system solves this problem by selecting the behaviors that are most probable given the goal in question. The computational dilemma faced by the direct-

mapping account is thus avoided by taking a top-down, predictive approach that exploits the hierarchical structure of goal-directed action.

The predictive action-monitoring hypothesis also seems to be more consistent with the existing mirror-neuron data: Gallese and Goldman found that monkeys' mirror neurons show no activity for mimicked actions, as when an experimenter pretends to grasp a non-existent object (Gallese and Goldman 1998). Umiltà and colleagues also found that monkeys' mirror neurons do respond to actions where low-level visual input is unavailable, as when they watch an experimenter reach behind an occluder to grasp a hidden piece of food (Umiltà et al. 2001). In humans, motor-priming (an effect of mirror neuron activity) seems to be sensitive to background knowledge about the intentional status of an bodily movement: if participants believe that a movement is forced, rather than goal-directed, no motor priming occurs (Liepelt and Brass 2010; Liepelt and Cramon 2008). Thus it seems that mirror neurons are responsive to *expectations* about goal-directed action, rather than the low-level visual properties of action.

4.2. Hierarchical Predictive Coding and theory of mind

The predictive action-monitoring hypothesis reflects a growing trend in cognitive science and neuroscience towards predictive models of cognition (Clark 2015; Seligman et al. 2013). The brain, in a very general sense, needs to be in the business of making predictions about the world: without being able to predict what's coming next, planning one's future actions is impossible. These predictions need to happen at multiple timescales simultaneously, whether we are predicting the objects in the space before us as we move through it, predicting where to find food in our local environment, or predicting events in the distant future. Predicting the behavior of other agents, in this sense, is just one part of the larger cognitive challenge of planning one's actions. As creatures that engage in complex forms of social coordination, this kind of prediction is especially important for human beings. The predictive action-monitoring hypothesis thus accounts for a key aspect how human beings are able to form plans at multiple timescales in a highly social environment.

Hierarchical predictive coding theories (HPC) have proven a fruitful way to translate this broad insight about the importance of prediction in cognition to specific hypotheses about

neural processing. According to HPC theorists, our expectations about the environment begin on the shortest possible timescale, with predictions about the causes of our present sensory experiences. On this view, neural systems do not just respond to incoming environmental information in a bottom-up manner, but also make forward-looking predictions about what that information will be, which they pass down the cortical hierarchy to the relevant input systems. Incoming information is then checked against the prediction signal; if the two do not match, an error signal is propagated back up the hierarchy, and checked against the higher order prediction. If these error signals are large, then the information they carry is incorporated into a revised internal model of the causal structure of the world, which then generates new predictions about incoming information. This process then repeats itself iteratively until prediction error signals are minimized. Formally, this account is said to be equivalent to a Bayesian updating procedure, wherein the posterior probability of a given hypothesis is a function of the prior probability of that hypothesis and the probability of a given observation (Bar 2007; Clark 2015; Friston and Kiebel 2009; Hohwy 2013; Spratling 2008).[6]

Building on initial applications of the HPC framework to mirror neurons (Kilner et al. 2007), a number of authors have now proposed HPC accounts of mindreading (de Bruin and Strijbos 2015; Hohwy and Palmer 2014; Koster-Hale and Saxe 2013). The most detailed of these proposals to date is that of Koster-Hale and Saxe (2013). Reviewing a wide range of neuroscientific evidence, they argue that much of the neural activity during mindreading tasks displays the signature of a predictive-coding architecture – namely, greater responsivity to unexpected stimuli than expected stimuli (i.e. prediction error signals). For example, they describe how the STS, which is associated with the processing of biological motion and goal-directed action, displays enhanced responses to unexpected behaviors, either because they are inefficient (Brass et al. 2007; Deen and Saxe 2012) or inconsistent with previously displayed desires (Jastorff et al. 2011). Likewise, the TPJ, which is known to respond to information about beliefs (Saxe and Kanwisher 2003), displays a stronger response to belief

[6] There is considerable variation amongst the different versions of predictive coding. Some theorists have taken the extreme position that prediction error signals are the *only* information carried via bottom-up input systems (Clark 2015; Friston and Kiebel 2009; Hohwy 2013), while others allow that traditional bottom-up information-processing compliments top-down prediction (Bar 2007; Spratling 2016). On my account, trait information (e.g. via facial features) is sometimes initially processed in a bottom-up fashion; as such, I disavow the idea that bottom-up input systems only carry prediction errors.

ascriptions that are surprising than those that are expected, given one's background beliefs about an individual (Cloutier et al. 2011; Saxe and Wexler 2005).

Koster-Hale and Saxe also argue that the data on trait-sensitive activity in the dmPFC is also consistent with a prediction-error minimization account. For instance, after Ma and colleagues provided participants with verbal information about the behavior of an individual (from which various character traits could be inferred), they presented them with test sentences that were either consistent or inconsistent with these descriptions (e.g. "Tolvan gave her brother a *hug*" versus "Tolvan gave her brother a *slap*"). They saw increased responsivity in the dmPFC for trait-inconsistent behaviors (Ma et al., 2011; see also Behrens, Hunt, & Rushworth, 2009; Kestemont et al., 2013; Mende-Siedlecki, Cai, & Todorov, 2013). Thus, the dmPFC seems to be sensitive to prediction errors related to personality traits.

Currently, the HPC approach to theory of mind is still in its infancy. As Koster-Hale and Saxe note, more empirical work needs to be done to develop the positive predictions of this kind of account in detail. However, HPC gives mindreading theorists a well-supported, general empirical framework for explaining the nature of action-prediction that has already been fruitfully applied to a number of different cognitive domains. It also coheres with a broader consensus among cognitive scientists about the centrality of predictive processes in cognitive systems. Although it is not without its controversies[7], the HPC approach in general is currently a progressive scientific research program (Lakatos 1970), and a promising way to pursue questions about the nature of social cognition. In the next section, I use this approach to develop an empirically supported conjecture about the relationship between character-trait attribution and other forms of mindreading.

[7] For example, the explanatory status of the Bayesian aspect of these models is a vexed question. Some theorists are explicit that the Bayesian formalism is intended to capture only the computational level of description, abstracted away from implementational, mechanistic details (Chater et al. 2006), while others seem to be making claims about the actual algorithms that support predictive processes (Friston and Kiebel 2009). While some have charged that ultimately, Bayesian models amount to "just-so" stories with little explanatory value (M. Jones and Love 2011), there are reasonable answers to such challenges (Zednik and Jäkel 2016), and plausible ways to interpret the various aspects of Bayesian models that render them empirically tractable (Icard 2016).

5. Character and the action-prediction hierarchy

Within a hierarchical Bayesian framework, a possible relationship between character traits and other mental states starts to emerge. As we saw initially with the case of mirror neurons, predictions about more transient states of affairs, such as motor intentions, tend to be informed by hypotheses about more temporally stable goal states. Hypotheses about goals, in turn, are informed by representations of more enduring desires. At each subordinate level in the predictive hierarchy, expectations about more transient states are shaped by superordinate hypotheses about more enduring states. As I discussed in the introduction, a key feature of character traits that distinguishes them from beliefs and desires is their greater temporal stability. As such, traits seem to fit naturally into the upper levels of the hierarchy for action-prediction. Background beliefs about character traits could thus inform and constrain predictions about more transient mental states, which then inform predictions about observable behavior.

To illustrate: suppose that you are observing Tom, whom you believe to be dishonest. A woman walks past, and accidentally drops her wallet in front of him. Tom looks toward the wallet, and then looks back at the woman. Because you know him to be dishonest, you assign a high probability to the hypothesis that Tom desires to steal the wallet. Given this desire-attribution, you might then expect that Tom will perform a series of actions: look around to see if anyone is watching, bend over discretely by the wallet as if tying his shoe, pick up the wallet and put it in his pocket. The prior trait attribution – dishonesty – thus serves as an over-hypothesis, raising the prior probability of mental-state hypotheses that are consistent with the trait in question – namely, self-interested desires (Kemp et al. 2007).[8] Thus, when we observe Tom's actions in a particular scenario, the first desire-hypotheses that we are liable to make will be based on this prior probability distribution. If we predict that Tom will form some particular self-interested desire, this will then raise the probability

[8] This is one way in which character traits may serve as an inferential heuristic: without this over-hypothesis, mindreaders would begin their action-predictions with a flat probability distribution over all the mental state hypotheses consistent with their current behavioral observations, which would give rise to an inverse problem. Instead, trait attributions bias the prior probability distribution towards a subset of mental-state hypotheses, which the predictor can proceed to test. Even if this distribution is in fact erroneous, it still serves as a means of bootstrapping our initial mental-state predictions, which then allow us to update our priors accordingly.

of certain hypotheses about Tom's actions. Trait-attribution thus has a cascading effect on the kinds of mental-states that we attribute, and ultimately on action-prediction.

How we predict that Tom will act on this initial desire-attribution will also be affected by other psychological and situational factors besides Tom's character. For instance, if there are people watching him, we may predict that when Tom looks around, he will refrain from further action. Likewise, if he sees the woman suddenly turn back, we might predict that he will form a new plan: to act as though he were intent on returning the wallet all along, and hand it back to her. Thus, the effects of trait-attributions on specific action-predictions are likely to be moderated by actively updating belief-attribution procedures that respond to immediate situational factors (Kovács 2015). Importantly, this shows that knowing both 1) that Tom is dishonest, and 2) that someone has dropped a wallet in front of him does not lead to any particular behavioral prediction. Rather, action predictions are produced via an initial trait attribution followed by a series of mental-state inferences at lower levels in the hierarchy.

Notably, if we antecedently believed that Tom were an honest person, then we might attribute to him the desire to return the wallet to its original owner. Ironically, this might generate a similar series of predicted behaviors as in our original prediction: looking around, reaching down to pick up the wallet, and then giving it back to the woman when she returns, or else pocketing it in order to bring it to the police later. Given two opposite trait attributions in the same situation, there might be no difference whatsoever in the actual behaviors initially predicted; all that would differ would be the kinds of intervening mental states that we ascribe to Tom. Within the pluralists' association-based model, we could never capture this difference. But on a hierarchical predictive model where traits inform mental-state attributions, which in turn inform predictions about behavior, we can.

One might object that these toy examples only really shows that character traits help us to predict desires, but that they don't seem to help us predict beliefs. But there are several ways in which trait attributions might make belief-hypotheses more probable. Traits relating to epistemic agency, such as gullibility or suspiciousness, will affect the priors we assign to hypotheses about beliefs formed on the basis of testimony, for example. Other traits may lead to predictions about how ambiguous situations will be interpreted: we may infer that a

paranoid individual will interpret two people whispering one way, an easygoing person another. And even when traits only lead to desire attributions, these might generate expectations about how a person will allocate their attention, which would in turn result in new perceptual beliefs, as when Tom saw the wallet, and then looked around to see if he was being watched.

This hierarchical, predictive approach to trait attribution also helps us make sense of some of the empirical data on trait attribution described in section 2. For instance, if trait attribution is higher up in the action-prediction hierarchy, and has a cascading effect on lower-order mentalistic and behavioral predictions, then we should expect that upon encountering someone new, trait attribution should be prioritized. The more quickly we start to construct a model of a person's character, the faster we will be able to use that information to predict and interpret their behavior. This means that within milliseconds of encountering someone, we need to start to gather whatever information is available to build up a representation of their stable character traits. Facial structure is well-suited for this purpose, because it can be processed extremely rapidly as coarse-grained, low spatial frequency information (Bar et al. 2006). This kind of input can be used for an initial conceptual categorization of a stimulus, which can then be used to generate predictions about subsequent input (Bar 2007; Chaumon et al. 2014). In other words, we can use facial information to form our first impressions of a person's character, which can then inform our subsequent expectations about intentional behavior.

Of course, initial trait attributions based on faces are neither accurate nor particularly informative for predictive purposes. But in a HPC framework, this is not a problem: learning from mistakes is what Bayesian systems do best. If an initial model results in a prediction error, this information can be used to update the model accordingly. For instance, if one encounters a person with a facial structure associated with trustworthiness, one might initially expect her to be generally well intentioned. However, if one witnessed such a person do something obviously cruel (e.g. abusing an animal), one would of course update one's model of that person's character (Mende-Siedlecki et al. 2013; Tannenbaum et al. 2011). As we accumulate new information about a person's behavior, we may iteratively revise our initial model of their character, leading to increased accuracy (Cunningham et al. 2007). This

helps to resolve one of the major puzzles of the pluralist account: even if trait attributions are initially unreliable, they might still serve as a basis for social learning and prediction, leading to increasingly accurate models of a person's character.

This complementary prioritization and updating of trait information can also shed light on the cognitive processing underlying the fundamental attribution error. Most impression-formation tasks that lead to the fundamental attribution error introduce participants to new people. If constructing a character model is prioritized by the mindreading system, then we should expect interpreters to use whatever behavioral information is available to construct that model as fast as possible. But when our attention is drawn to mitigating situational factors, our initial personality models are updated, and the behavioral evidence is discounted (Gilbert et al. 1988).[9] Likewise, when we are provided with additional mental state information (Krull et al. 2008; Reeder et al. 2004), or primed to think about mental states (Hooper et al. 2015), we use that information to update our character models accordingly.

However, in a HPC framework, not all prediction errors lead to updating. The world, after all, is messy and complex. Even a highly accurate causal model is liable to make mistakes. Many of these mistakes will be due to noise in the input, rather than a problem with the model. Updating the model to accommodate every piece of information it encounters would result in overfitting, and thus diminish its overall predictive accuracy (Hohwy 2013). Moreover, updating the model is likely to be cognitively costly, since it would require additional memory searches and generative procedures. Updating, in other words, can be a bad thing. Sometimes, prediction errors ought to be discounted as noise.

Modeling character traits is no different. One might have a fairly accurate representation of a person's character, and still occasionally be surprised by their behavior. For instance, one might be quite surprised to learn that Adolf Hitler was a vegetarian. Such information could be used to update one's model of his moral character, but this seems unlikely. Rather, one would simply ignore this information, and continue to rely on one's prior model. But this raises an interesting question: when do we update our character models in the face of new, conflicting behavioral information, and when do we treat it as noise?

[9] Members of "collectivist" cultures, who habitually attend to contextual factors, no doubt benefit from such attentional effects in their comparative resistance to the correspondence bias.

A recent study by Daniel Ames and Susan Fiske hints at an answer (D. L. Ames and Fiske 2013). Given that updating character models is likely to require the use of additional, limited cognitive resources, they hypothesized that people should selectively allocate those resources towards targets that are most behaviorally relevant to them. To test this hypothesis, they first introduced subjects to two confederates, who were described as "expert consultants" with whom the participants would later collaborate with in a joint project after first performing a solo task. Participants in the outcome-dependent condition were told that based on their performance in this joint task, they would be considered for a $50 prize. Participants in the outcome-independent condition, in contrast, were told that their eligibility for the prize would be based on their performance in the solo task only. Subjects then underwent fMRI scanning and were shown statements about the two confederates that were either consistent or inconsistent with what they had previously been told about them.

Ames and Fiske found inverse patterns of activity in the dmPFC for the two conditions: participants in the outcome-dependent condition displayed more responsivity to inconsistent information, whereas participants in the outcome-independent condition showed more responsivity to the consistent information. The authors suggest that the outcome-dependency manipulation led participants to use different updating strategies: when achieving their goal (the reward) depended upon interacting with the confederate, they used the inconsistent information to update their character model, so as to better predict and adjust to their partner's behavior. In the outcome-independent condition, in contrast, participants tended to dismiss the inconsistent information as noise, and thus conserve cognitive resources.

From a HPC perspective, we can interpret this outcome-dependent updating as reflecting higher-order predictions about the action-relevance of a prediction error. When surprising information is particularly important for action planning (e.g. when we expect to interact with a person in the future), we are more likely to incorporate that information into our predictive models, instead of dismissing it as noise. On the other hand, when a bit of surprising information is not action-relevant (e.g. when we do not expect to interact with a person in the future), we are less likely to devote resources to updating our predictive models, and more likely to dismiss the prediction error as noise. In other words, when a

prediction error is more relevant to our goals, we "raise the volume" on that signal; when it is less relevant, we "turn the volume down." In neuro-cognitive terms, this modulation of expected precision translates into shifts in attention. This may explain why participants under cognitive load are more likely to fall prey to the fundamental attribution error: when their attention is directed towards another task, they fail to attend to update their character models in response to error signals coming from situational information.

In sum: temporally stable character traits are represented at the upper level of an action-prediction hierarchy, and are used to generate prior probability distributions for hypotheses about more transient mental states, including beliefs and desires. These hypotheses are then used to inform hypotheses about even more transient states, which are in turn used to predict or interpret behavior. The downstream effects of trait-attributions on action-prediction are liable to be modulated by active belief-attribution procedures operating at lower-levels in the hierarchy. Prediction-error signals are conveyed back up the hierarchy, and are either used to revise the model at the appropriate level, or dismissed as noise. The action-relevance of an input can modulate whether prediction errors are treated as noise or used to revise the model by changing the expected precision in of an input signal.

6. Future directions

Adopting an integrative hierarchical approach to reasoning about character traits enables us to make a number of novel predictions. Broadly speaking, we should expect that manipulating background information about a person's character should lead to differences in the kinds of mental states that we attribute to them, especially when interpreting ambiguous actions. More specifically, manipulating trait attributions along the warmth dimension should lead to either more negative or more positive desire and intention attributions. For instance, individuals presented as low-warmth should be interpreted as having harmful or self-serving desires, while individuals presented as high warmth should be interpreted as having helpful or altruistic ones (as we saw in section 2, this kind of effect is likely to be responsible for the side-effect effect (Sripada and Konrath 2011; Sripada 2012)). Manipulations along the competence dimension should lead to differences in the amount of knowledge that we attribute to them: more competent individuals should be more likely to be viewed as having the appropriate beliefs and making the appropriate inferences, while less

competent individuals should be more likely to be viewed as ignorant. Warmth and competence information could be conveyed through facial information, through the observation of diagnostic behaviors, interactive experiences, or through testimony.

Beyond these initial predictions, this account of character-trait attribution offers us a new way to connect the study of mindreading with our understanding of stereotyping, prejudice and implicit bias (Spaulding 2016). As was noted earlier, the contents of common stereotypes are characterized by variation along the same two dimensions as ordinary trait attributions, suggesting that we use cues of group membership to infer that an individual will possess a particular set of character traits (e.g. an elderly person will be viewed as kind, but also as incompetent). Cues to group membership, such as skin color or accent, thus seem to play a similar role as facial features in conveying trait information; however, stereotypes seem to contain clusters of trait information, rather than just single traits. If trait attribution affects mental-state attribution as I've described, and stereotypes allow us to attribute clusters of character traits to individuals, then it would follow that stereotypes should also affect mental-state attribution.

As it happens, there is already some evidence that this is the case. Sagar and Schofield showed sixth-grade children vignettes depicting ambiguously aggressive dyadic interactions between students, such as one student bumping into another in a hallway, asking for food in the cafeteria, poking another student, and taking a pencil without asking. The authors also systematically manipulated the race of the actor in each dyad, such that some participants saw a white actors bumping, asking for food, poking, etc., while others saw black actors doing so. They found that the behaviors of black actors were interpreted as more mean and threatening than the identical behaviors from white actors (Sagar and Schofield 1980). Thus, when the intention underlying an action is ambiguous (e.g. intentionally threatening and aggressive versus neutral), observers fell back on stereotypes about black aggressiveness to interpret it. These results suggest that ascertaining the role of character-trait attribution in mindreading may also help us to better understand the cognitive basis for implicit racism.

The connection with stereotyping also raises the possibility that, in addition to their role in action-prediction, trait attributions may also serve a social function.[10] One of the explanations that has been offered for why we represent traits along the warmth/competence dimensions is that warmth helps us to keep track of potential threats, while competence helps us to keep track of agents' social status (Fiske et al. 2007). Notably, threat and status are not intrinsic properties of individuals, but rather relational, social properties that tend to vary with context: who counts as threatening or high status often depends upon one's own group identity and social rank. Thus, while we tend to represent traits as intrinsic, stable properties of individuals, it may be that what we are really tracking are social relationships.[11] These factors will still be highly relevant to action-prediction, however: whether an individual is higher or lower-status than us, or a member of the in-group or out-group, will have a significant effect on how they decide to act towards us. This may explain the predictive utility of trait-attributions: even if the situationists are right, and stable character traits do not really exist, we can still use trait representations as a proxy to help us factor social identity into our action-predictions.

7. Conclusion

Integrating character-trait attribution into a hierarchical Bayesian account of theory of mind promises to enrich our understanding, not just of these two sets of phenomena, but also a network of related phenomena of substantial social and philosophical importance. Traditional accounts of mindreading have paid little heed to character-trait attribution, focusing instead on the attribution of beliefs and desires. Folk psychology pluralists have rightly pointed this oversight, and taken important steps towards drawing attention to the significance of trait reasoning in folk-psychological inference. However, the pluralist account treated trait reasoning as a completely independent form of behavioral prediction, which does not fit well with the empirical data. In contrast, I have argued that the attribution of character traits is systematically related to the attribution of other forms of mentality, and that a hierarchical Bayesian architecture is a promising way to explain that relation. This

[10] Thanks to an anonymous reviewer for suggesting this.

[11] This may be an instance of what Cimpian and Salomon call the 'inherence heuristic' – a "fast, intuitive heuristic leads people to explain many observed patterns in terms of the inherent features of the things that instantiate these patterns" (Cimpian and Salomon 2014, p. 461) – and a precursor to psychological essentialism about certain social categories (Gelman 2004; Haslam et al. 2006; Rhodes et al. 2012).

account yields a number of novel empirical predictions about mindreading, and also has the potential to further unify the study of mindreading with neighboring empirical domains, such as stereotyping and implicit bias.

References:

Ames, D. L., & Fiske, S. T. (2013). Outcome dependency alterns the neural substrates of impression formation. *NeuroImage*, *83*, 599–608. doi:10.1016/j.neuroimage.2013.07.001.Outcome

Ames, D. L., Fiske, S. T., & Todorov, A. (2011). *Impression formation: A focus on others' intents. The Oxford handbook of social neuroscience*. Oxford: Oxford University Press.

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. doi:10.1177/0146167203261890

Andrews, K. (2008). It's in your nature: A pluralistic folk psychology. *Synthese*, *165*(1), 13–29. doi:10.1007/s11229-007-9230-5

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

Anscombe, G. E. M. (1958). Modern moral philosophy. *Philosophy*, *33*(124), 1–19.

Apperly, I., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–970. doi:http://dx.doi.org/10.1037/a0016923

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

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269–278. doi:10.1037/1528-3542.6.2.269

Behrens, T. E. J., Hunt, L. T., & Rushworth, M. F. S. (2009). The computation of social behavior. *Science (New York, N.Y.)*, *324*(5931), 1160–4. doi:10.1126/science.1169694

Brass, M., Schmitt, R. M., Spengler, S., & Gergely, G. (2007). *Investigating Action Understanding: Inferential Processes versus Action Simulation. Current Biology* (Vol. 17). doi:10.1016/j.cub.2007.11.057

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in cognitive sciences*, *12*(5), 187–92. doi:10.1016/j.tics.2008.02.010

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

https://books.google.com/books?hl=en&lr=&id=J5flK50tDaIC&pgis=1. Accessed 29 April 2015

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*. doi:10.1016/j.tics.2006.05.007

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. doi:10.1093/cercor/bht146

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

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

Christensen, W., & Michael, J. (2015). From two systems to a multi-systems architecture for mindreading. *New Ideas in Psychology*, *40*(A), 48–64. doi:10.1016/j.newideapsych.2015.01.003

Cimpian, A., & Salomon, E. (2014). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences*, *37*(05), 461–480. doi:10.1017/S0140525X13002197

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*. doi:10.1016/j.neuroimage.2011.04.051

Cogsdill, E. J., Todorov, A., Spelke, E. S., & Banaji, M. R. (2014). Inferring Character From Faces: A Developmental Study. *Psychological Science*, *25*(5), 1132–1139. doi:10.1177/0956797614523297

Csibra, G. (2008). Action mirroring and action understanding: an alternative account. In P. Haggard, Y. Rossetti, & M. Kawato (Eds.), *Sensorymotor Foundations of Higher Cognition*.

Attention and Performance XXII (pp. 435–459). Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199231447.003.0020

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: behaviors from intergroup affect and stereotypes. *Journal of personality and social psychology*, *92*(4), 631–48. doi:10.1037/0022-3514.92.4.631

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J.-P., et al. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, *48*(1), 1–33. doi:10.1348/014466608X314935

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. doi:10.1521/soco.2007.25.5.736

de Bruin, L., & Strijbos, D. (2015). Direct social perception, mindreading and Bayesian predictive coding. *Consciousness and Cognition*, *36*, 565–570. doi:10.1016/j.concog.2015.04.014

Deen, B., & Saxe, R. R. (2012). Neural correlates of social perception: The posterior superior temporal sulcus is modulated by action rationality, but not animacy. In *Proceedings of the 33rd Annual Cognitive Science Society Conference* (pp. 276–281).

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

Fein, S. (1996). Effects of suspicion on attributional thinking and the correspondence bias. *Journal of Personality and Social Psychology*, *70*(6), 1164. doi:10.1037/0022-3514.70.6.1164

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, & Behavioral Neuroscience*, 626–634. doi:10.3758/s13415-016-0419-2

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. doi:10.1111/mila.12079

Fiske, S. T. (2015). Intergroup biases: A focus on stereotype content. *Current Opinion in*

Behavioral Sciences, *3*(April), 45–50. doi:10.1016/j.cobeha.2015.01.010

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. doi:10.1016/j.tics.2006.11.005

Foot, P. (1967). *Theories of ethics*. Oxford: Oxford University Press.

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).

Gallese, V., & Goldman, A. (1998). Mirror neurons and the mind-reading. *Trends in cognitive sciences*, *2*(12), 493–501.

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*. doi:10.1080/10463280440000026

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in cognitive sciences*, *8*(9), 404–409.

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. doi:10.1037/0033-2909.117.1.21

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology*, *54*(5), 733–740. doi:10.1037/0022-3514.54.5.733

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press. https://books.google.com/books?hl=en&lr=&id=gRlnfe67ZAQC&pgis=1. Accessed 5 May 2015

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

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

Harman, G. (1999). Moral Philosophy Meets Social Psychology : Virtue Ethics and the

Fundamental Attribution Error Author (s): Gilbert Harman Source : Proceedings of the Aristotelian Society , New Series , Vol . 99 (19. *Proceedings of the Aristotelian Society, New Series*, *99*, 315–331.

Harris, L. T., Todorov, A., & Fiske, S. T. (2005). Attributions on the brain: Neuro-imaging dispositional inferences, beyond theory of mind. *NeuroImage*, *28*(4), 763–769. doi:10.1016/j.neuroimage.2005.05.021

Haslam, N., Bastian, B., & Kashima, Y. (2006). Psychological Essentialism, Implicit Theories, and Intergroup Relations. *Group Processes and Intergroup Relations*, *9*(1), 63–76.

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2013). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex*, *24*(8), 1979–1987. doi:10.1093/cercor/bht042

Heal, J. (1996). Simulation, theory, and content. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 75–89). Cambridge, UK: Cambridge University Press.

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

Hohwy, J., & Palmer, C. (2014). Social Cognition as Causal Inference: Implications for Common Knowledge and Autism. In M. Gallotti & J. Michael (Eds.), *Perspectives on Social Ontology and Social Cognition* (pp. 167–189). Dordrecht: Springer Netherlands. doi:10.1007/978-94-017-9147-2_12

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.

Icard, T. (2016). Subjective Probability as Sampling Propensity. *Review of Philosophy and Psychology*, *7*(4), 863–903. doi:10.1007/s13164-015-0283-y

Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology*, *21*(3), 384–388. doi:10.1037/h0032317

Jacob, P., & Jeannerod, M. (2005). The motor theory of social cognition: A critique. *Trends in Cognitive Sciences*, *9*(1), 21–25. doi:10.1016/j.tics.2004.11.003

Jastorff, J., Clavagnier, S., Gergely, G., & Orban, G. A. (2011). Neural Mechanisms of Understanding Rational Actions: Middle Temporal Gyrus Activation by Contextual

Violation. *Cerebral Cortex*, *21*(2), 318–329. doi:10.1093/cercor/bhq098

Jeannerod, M., Arbib, M. A., Rizzolatti, G., & Sakata, H. (1995). Grasping objets: the cortical mechanisms of visuomotor transformation. *Trends in Neurosciences*, *18*(7), 314–320. http://www.sciencedirect.com/science/article/pii/016622369593921J

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

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *The Behavioral and brain sciences*, *34*(4), 169–88; disuccsion 188–231. doi:10.1017/S0140525X10003134

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Macmillan.

Kalish, C. W. (2002). Children's predictions of consistency in people's actions. *Cognition*, *84*(3), 237–265. doi:10.1016/S0010-0277(02)00052-5

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

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. doi:10.1093/scan/nss022

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive Processing*, *8*(3), 159–166. doi:10.1007/s10339-007-0170-2.Predictive

Kitayama, S., Duffy, S., Kawamura, T., & Larsen, J. T. (2003). Perceiving an object and its context in different cultures: A cultural look at new look. *Psychological Science*, *14*(3), 201–206.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. *Neuron*,

79(5), 836–848. doi:10.1016/j.neuron.2013.08.020

Kovács, Á. M. (2015). Belief files in theory of mind reasoning. *Review of Philosophy and Psychology*.

Krull, D. S., Hui-Min Loy, M., Lin, J., Wang, C.-F., Chen, S., & Zhao, X. (1999). The Fundamental Fundamental Attribution Error: Correspondence Bias in Individualist and Collectivist Cultures.pdf. *Personality & social psychology bulletin*, *25*(10), 1208–1219.

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. doi:10.1016/j.jesp.2007.05.004

Lakatos, I. (1970). Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos (Ed.), *Criticism and the Growth of Knowledge* (pp. 91–196). Cambridge, UK: Cambridge University Press.

Liepelt, R., & Brass, M. (2010). Top-Down Modulation of Motor Priming by Belief About Animacy, *57*(3), 221–227. doi:10.1027/1618-3169/a000028

Liepelt, R., & Cramon, D. Y. Von. (2008). What Is Matched in Direct Matching ? Intention Attribution Modulates Motor Priming, *34*(3), 578–591. doi:10.1037/0096-1523.34.3.578

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. doi:10.1093/scan/nsr064

Ma, N., Vandekerckhove, M., Van Hoeck, N., & Van Overwalle, F. (2012). Distinct recruitment of temporo-parietal junction and medial prefrontal cortex in behavior understanding and trait identification. *Social Neuroscience*, *7*(6), 591–605. doi:10.1080/17470919.2012.686925

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction.* Cambridge, MA: MIT Press.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, *210*(4474), 1139–1141.

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person

impressions. *Social Cognitive and Affective Neuroscience*, *8*(6), 623–631. doi:10.1093/scan/nss040

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

Miyamoto, Y., & Kitayama, S. (2002). Cultural variation in correspondence bias: the critical role of attitude diagnosticity of socially constrained behavior. *Journal of personality and social psychology*. doi:10.1037/0022-3514.83.5.1239

Norenzayan, A., Choi, I., & Nisbett, R. (2003). Cultural similarities and differences in social inference: evidence from behavioral predictions and lay theories of behavior. *Human Resource Abstracts*, *38*(2).

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–8. doi:10.1126/science.1107621

Rakoczy, H. (2015). In defense of a developmental dogma: children acquire propositional attitude folk psychology around age 4. *Synthese*. doi:10.1007/s11229-015-0860-8

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

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. doi:10.1037/0022-3514.86.4.530

Rhodes, M., Leslie, S.-J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(34), 13526–31. doi:10.1073/pnas.1208951109

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual review of neuroscience*, *27*, 169–92. doi:10.1146/annurev.neuro.27.070203.144230

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

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

Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in Black and White children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology*, *39*(4), 590–598. doi:10.1037/0022-3514.39.4.590

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking peopleThe role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. doi:10.1016/S1053-8119(03)00230-1

Saxe, R., & Wexler, A. (2005). *Making sense of another mind: The role of the right temporo-parietal junction*. *Neuropsychologia* (Vol. 43). doi:10.1016/j.neuropsychologia.2005.02.013

Seligman, M. E. P., Railton, P., Baumeister, R. F., & Sripada, C. (2013). Navigating Into the Future or Driven by the Past. *Perspectives on Psychological Science*, *8*(2), 119–141. doi:10.1177/1745691612474317

Spaulding, S. (2016). Mind Misreading. *Philosophical Issues*, *26*.

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research*, *48*(12), 1391–1408. doi:10.1016/j.visres.2008.03.009

Spratling, M. W. (2016). Predictive coding as a model of cognition. *Cognitive Processing*, *17*(3), 279–305. doi:10.1007/s10339-016-0765-6

Sreenivasan, G. (2002). Errors about Errors: Virtue Theory and Trait Attribution. *Mind*, *111*(441), 47–68. doi:10.1093/mind/111.441.47

Sripada, C. S. (2009). The Deep Self Model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, *151*(2), 159–176. doi:10.1007/s11098-009-9423-5

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

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). *Moral signals, public outrage, and immaterial harms*. *Journal of Experimental Social Psychology* (Vol. 47). doi:10.1016/j.jesp.2011.05.010

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

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. doi:10.1016/j.tics.2008.10.001

Trope, Y., & Gaunt, R. (2007). Attribution and person perception. In M. Hogg & J. Cooper (Eds.), *The Sage handbook of social psychology* (pp. 176–194). London: Sage Publications.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A Person-Centered Approach to Moral Judgment. *Perspectives on Psychological Science*, *10*(1), 72–81. doi:10.1177/1745691614556679

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001). I Know What You Are Doing. *Neuron*, *31*(1), 155–165. doi:10.1016/S0896-6273(01)00337-3

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

Wellman, H. M. (2014). *Making Minds: How Theory of Mind Develops*. Oxford: Oxford University Press.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child development*, *72*(3), 655–84. http://www.ncbi.nlm.nih.gov/pubmed/11405571

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. doi:10.1016/0010-0277(83)90004-5

Zednik, C., & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese*, *193*(12), 3951–3985. doi:10.1007/s11229-016-1180-3