Stereotypes, Theory of Mind, and the Action-Prediction Hierarchy

Evan Westra

Forthcoming in *Synthese*

Abstract: Both mindreading and stereotyping are forms of social cognition that play a pervasive role in our everyday lives, yet too little attention has been paid to the question of how these two processes are related. This paper offers a theory of the influence of stereotyping on mental-state attribution that draws on hierarchical predictive coding accounts of action prediction. It is argued that the key to understanding the relation between stereotyping and mindreading lies in the fact that stereotypes centrally involve character-trait attributions, which play a systematic role in the action-prediction hierarchy. On this view, when we apply a stereotype to an individual, we rapidly attribute to her a cluster of generic character traits on the basis of her perceived social group membership. These traits are then used to make inferences about that individual's likely beliefs and desires, which in turn inform inferences about her behavior.

1. Introduction

Interpreting behavior in terms of underlying mental causes, or 'mindreading,' is widely agreed to be crucial to our ability to succeed in complex social environments: in order to predict and interpret behavior, we need to be able to reason about the hidden, mentalistic causes of action (beliefs, desires, intentions, etc.). But would-be mindreaders face a persistent challenge: behavior is quite often ambiguous, and consistent with many different possible mental causes. A smile from a stranger on the subway, for instance, could be a signal of recognition, an act of flirtation, an absent-minded reverie, or simple politeness. A shout from a neighbor's apartment might be an outburst of rage from a domestic disturbance or

excitement at a sudden turn of events in a football game. Inferences from behavioral effects to mental causes are always underdetermined.

When navigating social environments, we must somehow sort through these potential mentalistic causes, and arrive at the most probable interpretation. These abductive inferences require us to draw on our own background knowledge to fill in the gaps between behavioral observation and mental cause. Sometimes, we may fill in these gaps with our knowledge of the mindreading target herself and her individual history: if we know someone well, we are often able to infer what she is thinking quite accurately. But just as often, we interact with complete strangers, about whom we know nothing. In these cases, we may instead fall back on stereotypes about the target's social group membership. And this is a point where pernicious social biases can enter into the mindreading process distorting our interpretations of the social world.

It is not common to view stereotyping through the lens of theory of mind. Part of this may be an artifact of disciplinary boundaries: while theory of mind tends to fall under the scope of cognitive and developmental psychology, stereotyping is more often approached via social psychology. This is less true of social neuroscience, however, where there is some recognition that these two processes at least share overlapping neural substrates: stereotyping is significantly associated with activity in the dorsal medial prefrontal cortex and the anterior temporal lobe, which, together with the temporal parietal junction, superior temporal sulcus and precuneus, form the so-called 'mentalizing network' (Amodio 2014; Van Overwalle 2009). But more importantly, we also know that stereotyping and mindreading are both things that we do spontaneously whenever we observe or interact with other people (Bargh et al. 1996; Mason et al. 2006; Samson et al. 2010; Schneider et al. 2012). During social encounters, we rapidly and unconsciously retrieve information about an

individual's social category while simultaneously keeping track of her current mental states. This raises the important question of whether and how this information is integrated in the service of planning our own actions.

There are a number of behavioral findings scattered throughout different empirical literatures suggesting that stereotypes and mental-state attributions may interact in a very concrete way. For instance, Sagar and Schofield showed sixth-graders images and vignettes of ambiguous dyadic interactions between students, such as a student bumping into another in a hallway, asking for food in the cafeteria, poking another student, and taking a pencil without asking. These are behaviors that could be interpreted as the product of either a harmful or benign intention, depending on the participants' background assumptions. Critically, the authors systematically manipulated the race of the actor in each dyad. Nothing else about the observable behavior changed across these scenarios except the actors' race. The authors 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). That is, participants seemed more inclined to attribute harmful rather than benign intentions to the black actors than to the white actors.

Similarly, McGlothlin and Killen showed first- and fourth-graders a series of ambiguous images (e.g. a child picking up money on the ground behind another child, or a frowning child sitting on the ground in front of a swing with another child standing behind it) (Mcglothlin and Killen 2006; McGlothlin and Killen 2010). Once again, these images could be interpreted as depicting a benign action or a harmful action. As with Sagar and Schofield, McGlothlin and Killen varied the race of the actors, leaving everything else about the images exactly the same. They found that children were more likely to interpret the image as depicting a scenario in which a moral transgression had taken place (e.g. the child

picking up the money from the ground was *stealing* rather than *helping*), and to rate the action depicted as more impermissible when the actor was black than when the actor was white. Considering that children of this age reliably use information about intentions when judging whether an actor is blameworthy for a harmful action (Cushman et al. 2013; Killen et al. 2011; Leslie et al. 2006), it is plausible[1] that these divergent moral judgments are driven by divergent intention attributions. Thus, in both of these studies, knowledge of an actor's race seems to bias participants' mental-state attributions.

Background knowledge about the gender of an individual can also bias how we interpret his or her behavior. Condry and Ross showed college students videos of children wearing gender-disguising snow jackets playing roughly in the snow, and asked them to rate the aggressiveness of a target child's behavior. This situation was intentionally ambiguous, because the same roughhousing behavior could either be the product of playful intentions or harmful intentions. Across conditions, the dyads were labeled as male-male, female-female, or male-female. They found that boy-boy interactions were rated as less aggressive and more playful than girl-boy and girl-girl interactions (Condry et al. 1985). The authors speculated that this was due to the fact that play-fighting is a stereotypical play activity for boys, but not for girls; thus, boys were interpreted as having benign intentions, and girls were interpreted as having harmful ones. Expectations about gender even seem to bias how we interpret the behavior of infants: Burnham and Harris showed both college students and new mothers short videos of ambiguously gendered infants, which were randomly assigned either male or female names. Participants consistently judged the behaviors of infants with male names to

[1] One could come up with other, non-mentalistic interpretations of this result. For example, children might simply be relying on associations between race and moral transgression. But given what we know about children's ability to represent intentions, and their ability to use this information in moral judgments, these alternative interpretations seem rather implausible. Further research would be necessary to rule them out completely, however.

be stronger and more masculine. In both of these studies, participants all saw the same videos. The only things that affected their interpretations were their background stereotypes about gender (Burnham and Harris 1992).

Ageist stereotypes appear to affect the way we judge the accuracy of people's memories. For instance, a wide range of evidence shows that mock jurors tend to treat the eyewitness testimony of younger children as less credible that the identical testimony when given by adults, judging that their memories are in general less reliable, and that they are more prone to manipulation and confabulation (Goodman et al. 1984, 1987). An analogous effect also appears to afflict elderly witnesses, who are perceived by mock jurors as less competent and as having more inaccurate memories than younger adult witnesses; further, these judgments appear to be predicted by measures of ageist stereotypes (Mueller-Johnson et al. 2007). Tellingly, these attitudes also seem to be shared by real police officers (Wright and Holliday 2005). In effect, these studies tell us that we are much more likely to attribute *false beliefs* to both very young and old individuals than to other adults.

In real-world social encounters, this apparent interaction between stereotyping and mental-state attribution could have serious consequences: a teacher might judge a child's misstep as an accident, or as an intentional act of mischief; a doctor might hear her patient describe symptoms, and interpret it as an earnest desire for pain-relief, or as a deliberate deception to get an opioid prescription (Drwecki et al. 2011); a juror might dismiss the eyewitness testimony of an elderly person on account of her age; a police officer might interpret a thrashing man in handcuffs as either attempting to attack or panicking and struggling to breathe, based on a judgment about the man's underlying intentions (Goldstein and Schweber 2014; Spaulding 2017). In short, if stereotypes affect how we represent one

another's mental states, this may provide a potent avenue for discrimination to manifest itself in nearly all of our social interactions.

The central goal of this paper will be to suggest a theoretical account of how stereotypes might fit into the architecture of mindreading. I will propose that the source of this relationship can be traced to a core mentalistic feature of stereotype content, namely, the fact that stereotypes are structured around *attributions of character traits*. I will then introduce a model of action prediction that shows how character-trait attribution can influence how we represent other agents' mental states, and show how this explains the stereotype-mindreading relation (Westra 2017).

In section 2, I will identify several aspects of stereotypes that, I will go on to argue, are relevant to their connection with mindreading. In section 3, I will briefly discuss how stereotypes might fit into existing accounts of mindreading and folk psychology. In section 4, I will introduce a hierarchical predictive coding model of mental-state attribution, and show how character-trait attribution fits into that cognitive architecture. Section 5 will show how stereotypes fit into the action-prediction hierarchy, while Section 6 will discuss the various experiential and motivational factors that might moderate the effects of stereotypes on mindreading.

2. Stereotype content: character traits and essences

Stereotypes are stored bodies of rapidly accessible semantic information about the generic characteristics and attributes of social groups (Amodio 2014).[2] They can manifest themselves

[2] Social psychologists and neuroscientists distinguish stereotypes from 'prejudice': while the former is a semantic structure, and encodes descriptive properties of groups, the latter is an evaluative structure, and encodes valenced information. Prejudice and stereotypes are known to dissociate on a number of behavioral and neural measures (Amodio and Devine 2006; S. J. Gilbert et al. 2012). In this paper, I am specifically focused on stereotypes, and leave prejudice to one side.

as consciously endorsed explicit attitudes, or as unconscious attitudes that one might consciously disavow (Banaji et al. 1993). They can be triggered by perceptual cues, such as facial features, skin color, and accent (Mason et al. 2006). Stereotypic information can be activated very quickly and efficiently, and can have rapid biasing effects on attention, on the encoding and retention of information, and on a wide range of behaviors (Bargh et al. 1996; Correll et al. 2002; Donders et al. 2008; Hehman et al. 2014; Macrae et al. 1994; Rothbart et al. 1979); these effects are especially pronounced when under cognitive load (Macrae et al. 1993; Van Knippenberg et al. 1999; Wigboldus et al. 2004).

Beyond their cognitive profile, stereotypes also possess a distinctive kind of content. Intuitively, stereotypes can contain a wide range of information about a social group: styles of dress, music, food, accent, social practices, and various other kinds of parochial information might be contained in a stereotype. As long as a property can be attributed to members of a social group, it might seem that anything could become part of the content of a stereotype. This may be correct; however, the systematic study of stereotype content has also revealed that they possess an underlying structure and internal logic (Bastian and Haslam 2006; Fiske et al. 2002; Levy et al. 1998). At the core of this structure is the observation that many stereotypes seem to be about *character traits*: temporally stable, unobservable psychological properties[3] that have consistent effects on behavior across a wide range of different situations, such as laziness, intelligence, honesty, aggressiveness, and so on (Doris 2002). Upon a little reflection, this observation proves depressingly intuitive: we can all easily call to mind stereotypes about groups that are viewed as lazy, dishonest, greedy, unintelligent, aggressive, meek, and so on. This characterological dimension of

[3] Note that these are properties that we tend to ascribe to character traits in our folk psychology. We may think of character traits this way even if the reality is quite different, as proponents of situationism about character have proposed (Doris 2002). Also, the notion of character here is not meant to be a specifically moral, evaluative construct, and should be read as roughly synonymous with 'personality.'

stereotypes, I will go on to argue, plays a key role in their connection to mental-state attribution.

That stereotypes are structured around character traits is borne out by prominent theories of stereotype content and person perception. The Stereotype Content Model (SCM), for instance, proposes that most stereotypes are structured around two fundamental dimensions of trait attribution: 1)the warmth dimension, which tracks attributions of traits like trustworthy/untrustworthy, friendly/unfriendly, kind/unkind, and gentle/aggressive; and 2) competence dimension, which tracks attributions of traits like intelligent/unintelligent, skillful/clumsy, confident/meek, and serious/frivolous (Cuddy et al. 2007; Fiske 2015; Fiske et al. 2002, 2007).[4] Across many cultures (Cuddy et al. 2009), most stereotypes contain traits that fall into four distinct clusters: high warmth/high competence; high warmth/low competence; low warmth/high competence; and low warmth/low competence. The high warmth/low competence cluster (which includes traits like friendly or nurturing, but also unskilled and unintelligent) represented *paternalistic stereotypes* that are typically applied to groups perceived to be non-threatening and of relatively low social status: the elderly, homemakers, children, and the mentally disabled. The low warmth/high competence cluster
(e.g. intelligent but untrustworthy) represented *envious stereotypes*. This was applied to social groups viewed as both high status and threatening, such as lawyers, politicians, and professional women. The low competence/low warmth cluster represented *contemptuous*

[4] The warmth and competence dimensions are statistical posits that aim to explain recurring correlations between particular trait attributions (e.g. people who are judged as trustworthy also tend to be judged as friendly, kind, and gentle, and people who are viewed as intelligent also tend to be viewed as confident and serious). These two clusters of correlated traits appear throughout the trait-attribution literature, and have been given many labels besides warmth and competence: warm and cold (Asch 1946), social and intellectual (Rosenberg et al. 1968), self-profitable and other-profitable (Peeters 1983), morality and competence (Wojciszke 1994), and trustworthiness and dominance (Todorov et al. 2008).

stereotypes (e.g. unintelligent, unskilled, and dishonest). This was applied to low status, unthreatening groups, such as the homeless, drug addicts, and welfare recipients. Finally, the high warmth/high competence cluster (e.g. honest, friendly, intelligent, and confident) tended to pick out various *social reference groups* – high status, nonthreatening groups viewed as prototypical of a given society (e.g. in the United States, the white middle-class). In some cultures, this cluster was also ascribed to the participant's own social in-group (e.g. other undergraduate students).

Such stereotype-linked traits also have an effect on the kinds of traits that we infer on the basis of behavior. Normally, when we are given a piece of telling behavioral information about a person, we make a spontaneous trait inference about that individual (Uleman et al. 2008). We see this in the recognition probe paradigm (Uleman et al. 1996), in which participants are first presented with a sentence describing a behavior indicative of an underlying trait (e.g. 'Alice solved the mystery halfway through the book,' which implies that Alice is *clever*); next, subjects are presented with a word, and must judge whether or not it appeared in the sentence that was paired with that photo. When subjects see the trait word in question (e.g. 'clever'), they are far slower to respond "no" than when presented with control sentences. However, if the agent in the sentence belongs to a social category that is inconsistent with the trait inference in question (e.g. 'the *garbage man* solved the mystery halfway through the book,') this effect on reaction time is attenuated (Wigboldus et al. 2003). In other words, stereotypical traits seem to crowd out trait inferences that we might make on the basis of behavioral information alone.

The connection between stereotypes and character traits is also borne out in their relationship to people's tacit beliefs about the nature of traits, or 'implicit person theories' (Dweck et al. 1995). In particular, *entity theorists* – people who believe more strongly in the

immutability of traits and are disposed to infer traits on the basis of slim behavioral evidence – are much more prone to stereotyping. Specifically, entity theorists are more likely to endorse stereotypic trait attributions, to see stereotypic traits as the product of innate biological differences between groups, to accept stereotypic explanations of behavior, to infer new stereotypes about novel groups, and to pay greater attention to stereotype-consistent versus inconsistent information (Levy et al. 1998, 2001). The more strongly a person believes in stable, consistent character traits, it seems, the more likely she is to use stereotypes to reason about social categories.

The connection between stereotyping and belief in immutable character traits reveals another distinctive aspect of their content: stereotypes are *essentialized*. Psychological essentialism is the tendency to view biological categories (such as species) as discrete natural kinds, and to believe that members of these kinds all share hidden, innately specified, immutable, and causally potent properties or essences that explain the observable features of those categories (Gelman 2003). Stereotypes about social categories such as race and gender lead people to view those categories as biologically based natural kinds with hidden, immutable essences (Keller 2005; Prentice and Miller 2007). Stereotypic thinking thus reflects a tendency to subsume some social categories into a broader intuitive conceptual framework for understanding the biological world. In turn, entity theorists' tendency towards stereotyping and beliefs in the immutability and fixity of character traits is explained by a broader tendency towards essentialist thinking, about both individuals and certain social categories (Bastian and Haslam 2006; Haslam et al. 2000, 2002, 2004, 2006).[5] Stereotypes are, at their core, essentialized character-trait attributions applied to social groups.

[5] Most measures of social essentialism involve posing questions that probe beliefs about various components of essentialism for a given social group. For example, Haslam et al. (2000) provided adults with questionnaires

In sum, stereotypes are cognitively efficient mental representations that store information about social categories, which people draw on when making inferences about behavior. This information is organized in an essentialist fashion and has a strong characterological component. These facts will play a key role in the positive account that I present in section 5.

3. Stereotyping and the mindreading literature

Traditional accounts of mindreading, such as the simulation theory (ST) and the theory-theory (TT), have not addressed the role of stereotyping in mental-state inference. However, stereotypes seem to fit naturally into a TT account, insofar as they consist in generalizations that could enter into theory-driven causal inferences (Gopnik and Wellman 1992, 2012); indeed, my account can be viewed as an extension of the theory-theory. If a tacit mindreading theory is able to encode the relationship between these generalizations and the formation of various mental states, then the information encoded in stereotypes could enter seamlessly into theory-driven mental state inference. It would be incumbent upon the theory-theorist, however, to specify the nature of this relationship. It is not enough to say that stereotype-based generalizations affect mental-state attributions in some way or other. The theory-theorist must also explain the manner in which stereotypes (with their specific type of content and processing profile) influence mindreading.

---

that included items about the naturalness, inherence, and immutability of various social categories, including age, ethnicity, religion sexual orientation, etc. For instance, the inherence item asked used the following prompt: "Some categories have an underlying reality; although their members have similarities and differences on the surface, underneath they are basically the same. Other categories also have similarities and differences on the surface, but do not correspond to an underlying reality (Haslam et al. 2000, p. 118)." Participants then rated social categories on a scale of 'underlying reality or sameness.' Another measure of essentialism often used with children is the adoption task, which asks children to imagine an individual from social category A being adopted at birth by a family from social category B, and then asking the child whether the individual will grow up to display more A-traits or B-traits (Gelman and Wellman 1991; Hirschfeld and Gelman 1997; Segall et al. 2015).

In contrast, it is less clear how the ST might accommodate stereotypes. Simulation-based mental state inference relies upon one's own decision-making procedures to ascertain how another person might be reasoning (Gordon 1986; Heal 1996). One way stereotypes might fit into this procedure is if we consider stereotypes about our own social group when we act. For instance, if I am an academic, and there is a stereotype about academics being snobbish, a belief about this stereotype might enter into my decision-making procedure when I act. The literature on stereotype threat suggests that we may sometimes do this (Spencer et al. 1999; Steele and Aronson 1995). For instance, Spencer et al. (1999) found that women tend to perform worse on a math test when they are first reminded of the stereotype that women tend to perform worse than men on such tasks. However, in a recent meta-analysis of stereotype-threat findings, Flore and colleagues determined that stereotype-threat manipulations do not produce statistically significant effects on behavior (Flore and Wicherts 2015). Thus, evidence from stereotype threat can provide only weak support for this simulationist proposal. While it may be true that we sometimes consider stereotypes about our own group when we act, it is not at all clear if this has a reliable effect on our behavior.

Another possibility is that stereotypes might somehow provide the belief/desire inputs for a simulation-based mental-state attribution procedure, but that they do not themselves figure in our simulations. This proposal is more plausible, since it does not require us to posit that we regularly consider stereotypes about our own group in practical deliberation. However, it also departs from a pure simulation-based account of mental state attribution and moves into the territory of an ST/TT hybrid (Goldman 2006; Nichols and Stich 2003). In such an account, stereotype-based generalizations would simply figure into a theory of mental-state attribution, but not play any significant role in simulation-based

behavior prediction. Like the pure TT proposal sketched out above, this hybrid proposal would need to specify the relationship between stereotypes and mental states.

A different possible relationship between mindreading and stereotyping is that they might constitute two *entirely separate* ways of predicting and interpreting behavior: sometimes, we predict behavior by reasoning about mental states; other times, we predict behavior by applying stereotypes. This idea has recently been proposed by defenders of folk-psychological pluralism (Andrews 2008, 2012; Fiebich and Coltheart 2015).[6] The general thesis of folk-psychological pluralism is that human beings do not just rely on mindreading to predict and interpret behavior – we also rely on a wide range of socio-cognitive strategies that do not involve representing mental states at all. Among these alternative socio-cognitive strategies, pluralists have suggested that stereotyping provides us with an entirely non-mentalistic route to predicting and interpreting behavior.

Fiebich and Coltheart (2015) provide a number of proposals for how non-mentalistic, stereotype-based behavior predictions might work. The core mechanism underlying this form of prediction, according to their account, lies in the associations that we form on the basis of social group categories. At the most basic level, these might consist in associations between external cues to social group membership, particular situations or contexts, and particular behaviors (for example, police officers eating donuts in donut

[6] Andrews' (2012) account of the relation between stereotyping and mental-state attribution is not entirely clear. Initially, she presents her account of stereotyping as one of many ways in which, 'entire classes of behavior can be predicted, and even prognosticated, *without the attribution of beliefs and desires* [emphasis added]' (p. 68). But elsewhere, she seems more open to a role for mental-state attributions in stereotype-based behavioral predictions: for instance, she writes, 'when we stereotype others, we form expectations about people's behaviors and their *beliefs* based on their group membership [emphasis added]' (p. 86). One way to make sense of this tension would be if Andrews were distinguishing between mental-state attributions that occur via discrete acts of theorizing or simulation, and mental-state attributions that occur as the result of prior associations. That is, if we automatically apply a stereotype to a target, and that stereotype is associated with certain beliefs, we may incidentally come to attribute that belief to the target as well, without ever specifically reasoning about what their beliefs are. If this interpretation is correct, then my own account can be read as an argument for why the relation between mental-state attributions is not incidental at all, but rather quite systematic.

shops). These associations are automatically triggered when we observe members of the relevant social group (police) in the relevant situation (donut shop), leading to a particular behavioral expectation (eating donuts).

A more complex form of stereotype-based behavioral prediction involves associations between observable group-membership cues, behaviors, situations, and personality traits. On this view, stereotype-based prediction does not rely on direct associations between groups and particular behavior-situation combinations; rather, these associations are built around representations of traits. For instance, traits like generosity might be associated with behaviors like leaving large tips at restaurants. Thus, when we come to associate generosity with a particular social group (say, uncles), and we observe a member of a group in a situation that activates a trait-based behavioral association, we come to expect the individual belong that group to perform that behavior. This proposal is initially promising, since, as I have noted, trait-attribution is a core aspect of stereotyping.

It may be that associations like these ones explain some of the effects of stereotypes on our interpretations and predictions of behavior.[7] However, they do not explain why stereotypes also affect our performance on mental-state attribution measures like the ones described in the introduction of this paper (i.e. that stereotypes seem to lead us to attribute different mental states to otherwise identical behaviors). If stereotypes really are alternative way of reasoning about behavior, and do not involve mental-state attributions at all, why should they also influence the way we attribute intentions and beliefs? This suggests that, *pace* the pluralists, stereotypes and mindreading are in fact causally and functionally related.[8]

[7] Although there are good reasons for thinking that the structure of stereotypes is not based solely on statistical associations (Hammond and Cimpian, forthcoming; del Pinal and Spaulding, forthcoming)
[8] For a more detailed critique of folk psychological pluralism, see Westra (2017).

Stereotypes do not lead us to circumvent mindreading – they actually seem to interact with the mindreading process in a non-trivial way.

Another possibility is that the effects of stereotyping on mindreading are related to our social motivations. Along these lines, Spaulding (2017) has suggested that stereotypes and information about social categories may interact with our social goals to affect both the inputs to the mindreading system and the way we process mental-state information. First, a target's social group membership can provide us with a basis for determining the saliency of particular behaviors, which determines which information gets used in a particular mental-state inference. For example, whether a target is recognized as high or low-status might affect whether they represent a potential threat, which would affect whether or not we allocate attention towards their movements. Second, we may sometimes use stereotypes as a cognitively efficient mindreading strategy, particularly when our social goals lead us to prioritize speed over accuracy in our mental-state attributions.

Spaulding's general point that social category information affects how we allocate cognitive resources towards a particular mindreading problem is an important one: we do not deploy our mindreading abilities in a unitary fashion across social contexts and targets. Rather, we engage in mindreading to the extent that it supports our own action plans, which may require varying degrees of accuracy and efficiency. The suggestion that stereotypes can be used as a strategy for mental-state attribution also seems to fit nicely with the cases of biased mental-state attribution mentioned in the introduction. But this idea also needs to be further developed. Stereotypes, as we saw, have very distinctive contents and conceptual structure: they are organized around clusters of character traits, and they are essentialized. Why does a class of cognitive representations with these specific features facilitate

mindreading? As we search for an answer to these questions, however, we should keep in mind Spaulding's point about the effects of motivational factors upon mindreading.

4. Character, mindreading, and hierarchical predictive coding

In order to understand the relation between stereotyping and mindreading, it is useful to think about mindreading in terms of a hierarchical predictive coding account of cognition. Hierarchical predictive coding (HPC) refers to a family of models of neural information processing that heavily emphasize the importance of predictive processes in the way we represent and interact with the environment (Clark 2015; Friston and Kiebel 2009; Hohwy 2013; Hohwy et al. 2008; Rao and Ballard 1999; Spratling 2016). On this approach, the contents of our perceptual, proprioceptive, and interoceptive representations are not just informed by bottom-up signals from the environment or body to the brain; they are also informed by top-down predictions based on statistically informed expectations about what those incoming signals will be. This is accomplished through the activity of generative predictive models that constantly produce hypotheses about incoming perceptual experiences, which are checked in turn against incoming sensory inputs. These models are hierarchically organized, such that higher-order models make predictions based on highly abstract, more stable regularities in the environment, while lower-order models make predictions based on more local, transient properties of the environment. For example, lower-order predictions about visual input might represent low-level properties like edges, surfaces, and colors, whereas higher-order predictions might represent more abstract properties, such as objects and category-membership. Higher-level predictions are passed down the predictive hierarchy, informing predictions at subordinate levels. And so the prediction that one is looking at a *cup* can inform predictions about the presence of edges

and surfaces, because these low-level properties tend to co-occur with instances of the category *cup*. Thus, the contents of our perceptions of the external environment are partly constituted by these top-down predictions.[9]

4.1. The action-prediction hierarchy

Several authors have used HPC to model theory of mind (Csibra, 2008; Hohwy and Palmer, 2014; Hudson *et al.*, 2016; Kilner *et al.*, 2007; Koster-Hale and Saxe, 2013; Ondobaka *et al.*, 2015; Palmer *et al.*, 2015; Westra, 2017). The key observation underlying this approach is that we tend to represent intentional actions hierarchically. To illustrate, take an ordinary intentional action: getting a glass of water from a pitcher in the fridge. This action begins with the formation of the goal of quenching your thirst. Fulfilling this goal requires you to form a number of sub-goals: walk over to the kitchen, take a cup, get the water from the fridge, fill the cup with water, and take a drink. Each of these sub-goals is in turn achieved via the formation of specific motor intentions: getting water from the fridge involves *grasping* the fridge door, *pulling* it open, *reaching* into the fridge, and so on. At the higher levels of this action hierarchy, the relevant mental states are more temporally stable: the goal to get a drink persists throughout the entire exercise until every sub-component of the action is complete, while the sub-goal to get the water jug from the fridge only persists for a portion of the overall action sequence; individual motor intentions last even less time. Intentional actions,

[9] Because my account of mindreading and stereotyping is informal, it is likely to be consistent with a number of other computational approaches that treat cognition as a form of Bayesian inference, besides HPC (e.g. Gopnik and Wellman 2012; Lochmann and Deneve 2011; Solway and Botvinick 2012; Tenenbaum et al. 2011). The key features of any such model, as far as my account is concerned, would be 1) the hierarchical organization of mental-state inferences, where increasing levels in the hierarchy correspond to generative models producing hypotheses about properties of increasing temporal stability and abstractness, and 2) construing attention in terms of higher-order expectations about the precision of lower-order predictions (Hohwy 2012). My use of HPC reflects the fact that it incorporates these two features, and has also made important inroads into the mindreading literature (especially with respect to goal-based action prediction in the mirror neuron system (Kilner et al. 2007)). It does not entail a commitment to some of HPC's more controversial elements, such as the free-energy formulation of prediction-error minimization (Friston and Kiebel 2009), or the idea that feed-forward neural signals contain *only* information about prediction errors (Spratling 2013).

in other words, begin with more abstract, temporally extended mental states, and are realized through a hierarchy of increasingly transient states.

As mindreaders capable of reasoning about mental states, we can exploit the hierarchical structure of intentional action for predictive purposes (Csibra 2008). To predict an agent's observable behavior, a mindreader can begin by attributing to her an overarching goal (e.g. getting a drink from the kitchen). She can then use this goal attribution to predict the target's likely actions, starting with the sub-goals (e.g. going to the kitchen, retrieving a glass, opening the fridge, selecting a beverage). Each of these sub-goals can then be used to predict specific motor intentions (e.g. *grasping* the handle of the fridge and *pulling* the door open). At each level in this predictive process, the mindreader can use the superordinate, stable mental-state attribution as an overhypothesis, which assigns a prior probability distribution[10] to the many subordinate, transient mental state hypotheses that are consistent with her observable behavior. This prior probability distribution, in turn, constrains which mentalistic hypotheses she actually considers. For example, if the mindreader infers that a target walking towards the kitchen has the goal of getting a drink, then she need only consider those subordinate mental state hypotheses that would lead to the satisfaction of that goal (e.g. getting a can of soda, drinking water from the tap, getting water from the fridge, etc.). She need not consider all the possible mental state hypotheses that would be consistent with the target's observable behavior (e.g. going to make a sandwich, going to empty the dishwasher, rearranging the pots and pans, etc.). This winnowing of the mentalistic hypothesis space can then iterate at each level of the action-hierarchy, yielding a concrete expectation about observable behavior – which, some proponents of hierarchical

[10] This model does not require that the agent literally represent the entire space of possible mentalistic hypotheses for a given behavior, nor assign a prior probability to each of these. Rather, the agent's subjective prior probabilities could be interpreted as their propensity to sample from a hypothesis-generating mechanism, whose representational capabilities constitute the (latent) hypothesis space (Icard 2016; Perfors et al. 2011).

predictive theories of mindreading argue, manifest themselves as mirror-neuron activity (Csibra 2008; Kilner et al. 2007). When an action prediction does not match incoming inputs, an error signal is passed back up the action-prediction hierarchy, causing the internal psychological model that generated the prediction to be revised accordingly. This iterative combination of top-down predictions and error-driven learning provides mindreaders with an active, continually updating strategy for understanding the causal basis of the psychological world.

4.2. Character and mental-state attribution[11]

To see how representations of character traits fit into this hierarchy, we must first look at how they relate to our concepts of mental states like beliefs and desires. Character traits resemble other kinds of mental state representations in a number of ways: we tend to think of them as unobservable, internal properties of individuals that dispose us to act in certain ways. Different traits also seem to possess both cognitive and volitional elements, just like beliefs and desires: while traits like intelligence, paranoia, or gullibility have a distinct epistemic dimension, traits like friendliness and honesty seem almost desire-like. But if there is one property that character traits are thought to possess that distinguishes them from beliefs and desires, it is their temporal stability and consistency (Doris 2002).

The key idea here is that a trait is not the sort of thing that could suddenly shift or disappear, depending on the time or the context. We expect character traits to persist in individuals, and to have reliable effects on their behavior far into the future. If, for instance, a person only intermittently tells the truth, we would not call that person honest. Rather, we

[11] This account of the role of character-trait attributions in mindreading is based on a view developed in Westra (2017).

expect traits like honesty to regularly manifest themselves in people's behavior. In short, character traits are not thought to be readily changeable.

Beliefs and desires are not like this: beliefs can be changed or discarded; desires are can be satisfied. If I incorrectly believe that New York City is the capital of the state of New York, but am told that the capital is really Albany, one would not expect me to persist in my belief, but to change it. Likewise, if I have a desire for a cup of coffee, and then I go and buy one and drink it, one would not expect me to continue to desire coffee, because my desire has been satisfied. Beliefs and desires, in other words, are inherently changeable.[12]

A key difference between beliefs and desires and character traits, then, is that character traits tend to be viewed as highly temporally stable. If, as HPC theorists suggest, our action-prediction systems are organized into temporal hierarchies, then character traits would fit naturally into the upper levels of this temporally structured action-prediction hierarchy. This also suggests that representations of character traits may be used to make inferences about less stable mental states, such as beliefs and desires, which would in turn inform behavioral predictions. In other words, the mindreader's background beliefs about a person's character – their 'inner nature' – may inform the kinds of mental states that she ascribes to them.

---

[12] Granted, we do hold on to some of our beliefs and desires for long periods of time. But this has nothing to do with the nature of beliefs and desires *as such*, and everything to do with independent facts about the world. If I persist in believing that Washington, D.C. is the capital of the United States, or that all bachelors are unmarried, it is because facts about the world (and the meaning of 'bachelor') make these beliefs true. Likewise, I may have standing desires for world peace and to win the lottery; what makes these desires persist is that my winning the lottery and world peace are unlikely to happen, and so my desires are destined to go unfulfilled. This is not so for the stability of character traits.

Also, note that the beliefs and desires that we are often least likely to give up, such as deeply held moral convictions and values, are precisely those that we treat as part of our core identities, as essential to who we are (Strohminger and Nichols 2014). They are, in other words, much more *trait-like* than our other attitudes.

By way of illustration, suppose that you are on a walk with your friend George, and you both observe George's wife speaking with an attractive man. If you antecedently believe George to be insecure, you might expect him to form the belief that his wife is being disloyal to him; you might then expect him to become angry and make a scene. But if you antecedently believe George to be confident, you might instead expect him to believe that his wife is simply engaged in an innocent chat; you might then predict that George will pleasantly greet the pair. In short, depending on your background beliefs about George's competence traits (whether he is insecure or confident), you may end up attributing different stable background beliefs to him, which would then lead to different action predictions.

Similarly, imagine that your friend Claire is either honest or dishonest (i.e. high warmth or low warmth), and that you see her happen upon a wallet full of cash on the street. If she is honest, we might expect her to desire to return the wallet to its owner, then form the intention to pick it up and bring it to the police station, and then pick it up and put it in her pocket. If Claire is dishonest, in contrast, we might expect her to desire to keep the cash for herself, then form the plan to discreetly pick it up, put it in her bag, and walk away.

These examples show how trait attributions (on both the warmth and competence dimensions) could lead to cascading effects upon mental state attributions at various levels in the action-prediction hierarchy, from relatively stable background beliefs and desires to highly transient motor intentions and perceptual beliefs. In a hierarchical action-prediction system, particular trait attributions would affect the probability that an agent might form certain beliefs and desires. Given the many possible mental states that might cause a given behavior, background knowledge about personality traits would serve to render certain mental states more probable. High-warmth traits, for instance, would make desires with helpful contents much more probable, while low-warmth traits would do so for desires with

harmful or self-serving contents. High-competence traits might make true-belief or knowledge attributions more probable, while low-competence traits would predict false beliefs and ignorance. Competence traits might also be related to the attribution of intentions of plans that are likely to be successful. Thus, the attribution of stable character traits changes the probability distributions for hypotheses about beliefs and desires. These beliefs and desires would then inform hypotheses about more transient mental states and, ultimately, our predictions and interpretations of behavior.

The fact that character traits occupy a relatively high position in the action-prediction hierarchy helps to explain a number of puzzling phenomena related to trait-attribution. For instance, it is well known that we often make very rapid, intuitive, 'thin slice' judgments about other individuals (which can be surprisingly accurate). We also use highly superficial information about facial features to make extremely rapid (i.e. under 100ms) inferences about traits such as trustworthiness and dominance (Bar et al. 2006; Todorov 2013; Todorov et al. 2008). Some populations are also prone to interpret behavior as the product of a person's character traits, rather than their situation, a phenomenon known as the 'correspondence bias' or 'fundamental attribution error' (Gawronski 2004; D. T. Gilbert et al. 1995; Jones and Harris 1967; Ross 1977). If representations of character traits sit towards the top of the action-prediction hierarchy, and have significant downstream effects upon other forms of mental-state attribution, then it would make sense for this information to be prioritized, and processed as rapidly and efficiently as possible. Ironically, this means some of the most rapid inferences that we make about people are about what we take to be their deepest, most stable traits.

To be clear, the claim here is not that mental states are attributed solely on the basis of prior character-trait attributions. Attributions of mental states at all levels of the hierarchy

are most likely supported by numerous socio-cognitive mechanisms, such as gaze-following, the detection of biological motion, and emotion-recognition systems, as well as default strategies like the projection of one's own beliefs, and statistical information acquired via experience. The action-prediction system, in short, is likely to receive both bottom-up and top-down inputs from a number of sources. However, characterological information may play an important role in informing the probabilities that we ascribe to alternative mentalistic hypotheses generated at lower levels in the action-prediction hierarchy. Character-trait attributions, on this account, provide the mindreader with an efficient and stable inferential basis for modeling the transient mental states of other agents.

5. Stereotypes in the action-prediction hierarchy

If character traits play this role in the action-prediction hierarchy, and the contents of stereotypes are structured around character traits, then we have the beginnings of a plausible hypothesis about the relationship between stereotypes and mindreading. When we apply a stereotype to an individual, on this view, we draw on a stored model of the generic character traits that we associate with a particular social group. We then use these traits to make inferences about that individual's likely beliefs and desires, which in turn inform inferences about her behavior. So, for instance, if an individual belongs to a group that is stereotyped as high warmth, this raises the prior probability that they will have helpful, honest, and friendly intentions. Conversely, identifying someone as belonging to a low-warmth group would raise the prior probability that she would have harmful, deceitful, and unfriendly intentions. And, tentatively, if an individual belongs to a stereotypically low-competence group, her behaviors are more likely to be interpreted as stemming from ignorance or false beliefs, while an individual from a stereotypically high-competence group will be more likely to be interpreted

as acting on true beliefs or knowledge (for examples of ageist stereotypes about inaccurate memories, see Goodman et al. 1984, 1987; Mueller-Johnson et al. 2007).

This inference may involve several steps: a stereotypic trait attribution might first lead to the attribution of more general, stable beliefs and desires, which might then be used to infer more concrete intentions, and ultimately get used to predict or interpret specific movements. For example, attributing a stereotype of greediness to a person might initially lead us to infer that she has a strong, stable desire for wealth; in a particular social interaction (say, a business negotiation), this broad desire attribution might lead us to a more specific one (e.g. she does not want to give up shares in the company), which might in turn lead us to particular interpretations of visible behavior (e.g. a furrowed brow signifying reluctance rather than contemplation) (D. R. Ames et al. 2012). Note that the stereotype does not, in this case, directly lead us to a specific interpretation of the agent's behavior in a particular context – that kind of situation-specific information is probably not stored in the stereotype itself. Rather, the stereotype points us in the direction of a high-level, general mental-state attribution, which in turn facilitates increasingly concrete mindreading hypotheses.

To take a concrete example, consider the case of Sagar and Schofield (1980). Their primary finding was that the same ambiguous action was rated as more aggressive, mean, and threatening when it was performed by a black actor than when it was performed by a white actor. According to Fiske et al. (2002), black people in the United States are rated as lower-warmth than white people. On the current account, this is what explains the difference in mental-state attribution across the two groups: observers used a stereotype to make an inference about the stable traits of the black actor in the vignette, which led them to see the harmful desire as more probable than the playful/neutral one. A similar explanation can be given for the findings of McGlothlin and Killen (2006, 2010): whereas the ambiguous

actions of white actors in the images are judged to be morally neutral, participants attributed harmful intentions to the black actors, and thus judged their actions to be morally wrong.

On the current proposal, participants in these studies would have first rapidly processed external cues about the actors' group membership (i.e. skin color), which would have activated a stereotype associated with African Americans. This stereotype would include a cluster of personality traits, including low-warmth traits like aggressiveness, which would be used to develop a hierarchical model of that actor's character and mental states. Upon observing the actor's ambiguous action, the mindreader's hierarchical model would generate the intentional interpretation that would be most likely given the prior hypothesis that the actor has the trait of aggressiveness: namely, an intention to cause harm. Thus, prior hypotheses about the actor's character (based on information stored in stereotypes) would have influenced how the mindreader interpreted her ambiguous action. Similarly, when anticipating an unfamiliar agent's actions, stereotypes about that agent's social identity could be used to generate a generic model of their personality traits, which might inform action-predictions.

This account of stereotypes helps us understand how they lead to more computationally efficient (but also less accurate) mindreading. As was mentioned in the introduction, a single behavior can often be given indefinitely many different mentalistic interpretations. Some authors have argued that this problem actually makes mindreading an intractable problem, since every act of mentalistic interpretation will require sorting through an indefinitely large set of potential interpretations (Bermudez 2003; Morton 1996; Zawidzki 2013). What stereotypes do is bias us towards a subset of those hypotheses, which we then use to support our behavioral predictions and interpretations. In other words, stereotypes

save computational resources by providing us with a heuristic for rapidly generating (what we take to be) relevant mentalistic interpretations of behavior.

The high position of stereotypes within the action-prediction hierarchy can also help us to understand the fact that they are essentialized. Recall for a moment some of the characteristic features of essences: they are unobservable, they are immutable, and they are causally potent. These are precisely the features that we should expect to find in the highly abstract and stable representations that occupy the upper level of the action-prediction hierarchy. The upper levels of this hierarchy, after all, are meant to track very stable regularities in the environment, which in turn allow us to make sense of highly variable, transient perceptual inputs. In the domain of action-prediction, I have suggested, these regularities manifest as representations of temporally stable, consistent character traits that reliably dispose agents towards certain kinds of mental states. What stereotypes do is assign sets of these trait-essences to particular social groups for the purposes of predicting the actions of their individual members.

Exposing this kind of relationship between mindreading and essentialist thinking about social groups also raises interesting questions about how the two processes interact over the course of development. Both theory of mind abilities and essentialist thinking about social groups are present in preschool-aged children across cultures (H. C. Barrett et al. 2013; Liu et al. 2008; Rhodes and Mandalaywala 2017); however, there are important differences in how the two forms of social cognition manifest themselves between childhood and adolescence. The basic ability to reason about more transient states, such as beliefs, desires, and emotions, exhibits a more-or-less cross-culturally stable developmental trajectory

(Shahaeian et al. 2011; Wellman et al. 2001, 2006).[13] However, both the content of specific essentialist stereotypes, and their persistence throughout the lifespan, vary widely with culture and social context. For instance, essentialism about gender categories is quite common across most cultures, but essentialism about race and ethno-religious categories is not (Chalik et al. 2017; Diesendruck et al. 2013). And while most children are essentialists about many social groups early on, this tendency will diminish in some populations as time goes on (e.g. children growing up in more urban environments), but persist in others (e.g. children growing up in more rural environments) (Rhodes and Gelman 2009). Famously, essentialist thinking about particular groups also appears to be triggered in part by certain linguistic cues, such as generic statements (e.g. "Boys like trucks and girls like dolls,") (Gelman and Roberts 2017; Rhodes et al. 2012; Segall et al. 2015). Thus, while younger children across different environments develop a similar basic set of theory of mind abilities, their reliance on particular essentialist stereotypes is greatly affected by numerous social factors.

On the present account, this makes sense: while basic forms of mental-state attribution are used across social contexts, essentialist stereotypes provide mindreaders with a heuristic for rapid, specific mental-state attributions for unfamiliar members of known social groups. The particular groups that get essentialized will inevitably depend upon the salience of different group boundaries, which will be affected by a child's first-hand experience with those groups, and by the way that they are talked about in the child's

[13] There is, of course, a huge debate about when certain theory-of-mind abilities (especially belief-attribution) develop (Baillargeon et al. 2010; Heyes 2014; Scholl and Leslie 2001; Wellman et al. 2001). But whether one believes that the core elements of theory of mind develop rapidly in the first year of life (Carruthers 2013), or more slowly over the first five years (Wellman 2014), there is still a general consensus that children possess a wide range of theory-of-mind abilities by at least four-and-a-half (Wellman et al. 2001), and display other relevant abilities quite a bit earlier (Behne et al. 2005; Moll and Tomasello 2006; Repacholi and Gopnik 1997; Wellman and Liu 2004).

environment. This will be something of a learning process for children, who may implicitly try out and then subsequently revise or discard different essentialist ideas as they learn about different groups. Ultimately, the way that stereotypes affect mental-state attribution should depend on the salience of the stereotype in the child's experience. In this manner, a child's basic theory-of-mind abilities get supplemented with higher-order, culturally local generalizations about the character traits of different types of individual. Thus, while we see cross-cultural consistency in the kinds of reasoning that occurs at lower levels of the action-hierarchy, we get much more cultural specificity at its higher levels.

6. Stereotypes, goals, and error-driven updating

The present proposal suggests that stereotypes inform our mental-state attributions, and thereby affect our behavioral predictions. But this raises an important question: what happens when stereotypes yield faulty action-predictions? Shouldn't this lead to prediction error signals that cause us to update the models that generated the faulty prediction – namely, the stereotype itself? In other words, shouldn't a hierarchical predictive coding model of stereotypes ultimately lead to their elimination?

In some circumstances, this may be precisely what happens: positive intergroup contact is known to reduce bias (Pettigrew and Tropp 2000). Indeed, when McGlothlin and Killen (2010) gave their ambiguous pictures task to children from racially heterogeneous schools, children were not more likely to attribute harmful intentions to the African American actor. Notably, with increased intergroup contact, these children would have had more opportunities to engage in cooperative activities with children of different races, during which they would have engaged in mindreading. These same children could also reasonably expect to engage in such cooperative activities in the future. Thus, both their past

experiences and future plans would push them away from inaccurate stereotype-based mindreading (Pettigrew and Tropp 2008). However, in the absence of such regular contact with members of other groups, mindreaders' stereotypes may instead be reinforced by their repetition in public discourse, which creates a biased learning environment.

Several other factors might affect the perseverance of stereotypes in our action-prediction models. One reflects a basic challenge for any attempt to develop a predictive statistical model of the environment: the world is a messy place, and full of noise. If there is noise in the data that gets fed into a model, then it is to be expected that even highly accurate models will sometimes make incorrect predictions. In these cases, adjusting the model so that it fits the data would come at a cost to its predictive accuracy. Thus, even though a predictive model must be sensitive to error signals, some of these must be discounted as noise at least some of the time, or else run the risk of overfitting the data.

In predictive coding models of cognition, which error signals get treated as noise and which ones do not can be determined by second-order predictions about the gains in predictive precision that would be achieved by updating one's existing model. With predictive models that are expected to be largely accurate, the predictive utility of updating in response to an error signal would be relatively low, and so error signals will be more likely to be treated as noise. When models are not expected to be highly precise, the predictive benefits of updating would be higher, and so the system would become more sensitive to prediction errors. In other words, second-order predictions determine whether we 'turn up' or 'turn down the volume' on a given set of error signals. These modulations in the 'volume' or 'gain' on prediction errors manifest themselves as changes in attention (Hohwy 2012, 2013). Thus, when we expect that certain incoming sensory inputs are likely to carry

information that will improve the predictive accuracy of our internal models, we pay more attention to those signals.

A further factor that affects the way we update our models will be the expected value of incoming information with respect to our action plans. Predicting the world, after all, is not an end unto itself; it is a means to support adaptive behavior. Moreover, the computational resources we can devote to any given prediction problem are inherently limited. Thus, not all accurate predictions are of equal value. In general, we should expect the way we devote computational resources towards responding to prediction errors to vary as a function of the adaptive significance of the prediction problem in question. Some prediction problems, such as detecting predators, may not actually benefit from increased accuracy, given the high costs of false negatives; other problems, such as coordinating our behavior with other agents in the service of joint goals, may require highly accurate predictions (Godfrey-Smith 1991). In effect, the gain on prediction errors needs to be modulated by higher order predictions about the expected utility[14] of updating our internal models.

In practice, this will mean that the extent to which we revise our initial, stereotypic models in response to stereotype-inconsistent information will depend upon our goals (D. L. Ames and Fiske 2013; Spaulding 2017; Westra 2017). If our plans happen to depend upon accurately representing an individual's mental states (say, if we think we are likely to

[14] In an HPC framework, estimations of expected utility would need to rely upon affect-based, interoceptive predictions about the somatic and hedonic consequences of a prospective scenario (Barrett, 2017; but see Carruthers (2017) for a non-hedonist account of the function of valence in prospection). Contemplating walking down a dark alleyway in a bad neighborhood, for example, may yield a prediction about the likelihood of a threatening encounter, which would in turn trigger an affective response – namely, a preparation for fight or flight. This affective prediction could in turn support decision-making (Seligman et al. 2013), but also one's subsequent sensitivity to prediction errors via the allocation of attentional resources.

Note also that this construal of affect would also necessarily figure in any HPC account of prejudice (see footnote 1).

cooperate with this person in the future), then we should heighten our sensitivity to stereotype-driven prediction errors, and update accordingly. On the other hand, if we don't expect that accurately representing this individual's mental states will make a difference to our plans (e.g. if we view the individual as belonging to a lower social status than us), then we may be more likely to dismiss prediction errors as noise. This mechanism of increasing and decreasing the sensitivity to prediction errors makes sense of Spaulding's point that our biases affect both mindreading inputs and mindreading processing in terms of HPC: when our social goals are best served by reliance on stereotypes, we ignore stereotype-disconfirming inputs, and instead rely upon generic models of other agents' mental states. When our social goals instead motivate us to accurately represent an individual's mental states, we devote additional cognitive resources towards integrating stereotype-disconfirming information into our model of their beliefs and desires.

Notably, there may be many different contexts in which our plans require greater precision in our mentalistic models of other agents. I have already mentioned the potential for cooperation as a motivating factor: if one must engage in complex forms of coordination with an individual in order to achieve a joint goal (e.g. co-authoring a paper), one will then need a much more complex and precise model of the colleague's beliefs, intentions, and perceptual states. But in less complex, highly familiar forms of cooperation (e.g. paying a cashier at a grocery store), less precision is necessary; in these cases, mental models are likely to be highly schematized and reliant on stereotypes. The amount of precision required for competitive or hostile interactions with out-group members will also vary depending on the context. In a highly strategic competitive interaction, such as a business negotiation, one generally needs a very elaborate model of the other agent's goals and intentions, as stereotyping in these contexts will likely lead to error (D. R. Ames et al. 2012). In contrast,

when one is confronted by a mugger with a gun demanding one's wallet, one does not need to engage in extensive planning, and so the added value of a richer, highly precise mental model will be quite limited. The precision of one's predictive models of other agents' mental states should thus be highly sensitive to the complexity and familiarity of the interaction in question, in addition to the social identities of the interacting agents.

What happens when we do devote more resources towards representing an individual's mental states? One possibility is that, as we are more strongly motivated to engage in accurate mindreading with a particular individual, we are more likely to explicitly, consciously represent another person's experiences by projecting ourselves into their situation (Buckner and Carroll 2007). Explicitly representing the perspectives of other agents via self-projection is known to facilitate more accurate, sophisticated forms of mental-state attribution (Surtees et al. 2012, 2013), although it also draws heavily on working memory resources (Bukowski and Samson 2017; Wardlow 2013). In terms of the action-prediction hierarchy, this kind of perspective-taking would shift cognitive resources towards generating more precise predictions about the target's transient mental states, drawing more heavily on one's own experiences and introspective knowledge. Rather than passively relying upon the top-down effects of stereotypes to shape one's mentalistic hypotheses, this strategy leads mindreaders to consciously construct a richer, more detailed, 'individuated' representation of the target's beliefs, desires, and experiences (Mason and Macrae 2004).

Consistent with this idea, perspective-taking manipulations have been shown to lead people to seek out stereotype-disconfirming information, and to have a better memory for stereotype-inconsistent information. In a series of studies, Todd and colleagues manipulated whether white participants adopted the perspective of a black individual, and then proceeded to show them a series of 30 sentences about that individual describing behaviors that were

consistent, inconsistent, or neutral with respect to African-American stereotypes. After completing a five-minute distractor task, participants were asked to recall as many of the sentences as they could. They found that participants who first took the perspective of the target were more likely to remember stereotype-inconsistent information than those that did not. In another task, Todd et al. also showed that participants were more likely to choose to ask stereotype-disconfirming questions about a target when they first took that target's perspective (Todd et al. 2012). Thus, when participants were motivated to direct more cognitive resources towards representing mental states of the target, they were also more likely to seek out and retain information that disconfirmed stereotype-based predictions.

But while perspective-taking may be effective in mitigating the effects of stereotyping on the interpretation and prediction of behavior, the heavy executive demands of this strategy limit its applicability: in conditions of cognitive load, we should expect stereotype-based bias in mindreading. In other words, even when a mindreader is highly motivated to engage in accurate mindreading, features of the context that place heavy demands on cognitive resources (e.g. when her working memory is occupied with multiple tasks at once) may lead her to rely more heavily on stereotypes.

7. Conclusion

I have proposed that the effects of stereotypes on mental-state attribution can be traced to their characterological content. The clusters of essentialized character traits contained in stereotypes influence other forms of mindreading by informing the relative probabilities ascribed to different mentalistic hypotheses, which in turn influence our predictions of intentional actions. In this hierarchically structured action-prediction system, stereotypes facilitate efficient mindreading, but also lead to biased interpretations of behavior. The

effects of stereotypes on mindreading can be modulated by social learning, motivational factors, and other features of context. One way that stereotype-driven mindreading can be mitigated is through increased intergroup contact; another is via effortful, working-memory based forms of explicit perspective-taking.

This proposal already has some empirical support, but could be more directly tested in a number of ways. The way to do this would be to follow the model of the studies discussed above: present participants with ambiguous scenarios and provide them with mental-state attribution measures, varying the social group membership of the actor across multiple conditions. This basic design could be supplemented with measures of stereotype endorsement, stereotype activation, or essentialist thinking more generally.

Such an approach could prove highly beneficial to our understanding of the cognitive underpinnings of bias. There is an immense and advanced literature on the neural and cognitive underpinnings of mindreading, which can inform and extend our understanding of the way stereotypes operate in everyday situations, and help us to develop targeted intervention strategies to mitigate their effects. Likewise, our knowledge of mindreading will only benefit from a concerted effort to understand how we deploy our social-cognitive abilities in intergroup contexts. By studying theory of mind and stereotyping together, we stand to learn about the various ways that mindreading goes awry, and contributes to pernicious patterns of social bias.

References:

Ames, D. L., & Fiske, S. T. (2013). Outcome dependency alterns the neural substrates of impression formation. *NeuroImage*, *83*, 599–608.

Ames, D. R., Weber, E. U., & Zou, X. (2012). Mind-reading in strategic interaction: The impact of perceived similarity on projection and stereotyping. *Organizational Behavior and Human Decision Processes*, *117*(1), 96–110.

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews: Neuroscience*, *15*(10), 670–682.

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and Evaluation in Implicit Race Bias : Evidence for Independent Constructs and Unique Effects on Behavior. *Journal of personality and social psychology*, *91*(4), 652–661.

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

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

Asch, S. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, *41*(3), 258–290.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in cognitive sciences*, *14*(3), 110–8.

Banaji, M. R., Hardin, C., & Rothman, A. J. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology*, *65*(2), 272–281.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269–278.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230–244.

Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., et al. (2013). Early false-belief understanding in traditional non-Western societies. *Proceedings of the Royal Society of London B: Biological Sciences*, *280*(1755), 20122654.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

Bastian, B., & Haslam, N. (2006). Psychological essentialism and stereotype endorsement. *Journal of Experimental Social Psychology*, *42*(2), 228–235.

Behne, T., Carpenter, M., & Tomasello, M. (2005). One-year-olds comprehend the communicative intentions behind gestures in a hiding game. *Developmental Science*, *8*(6), 492–499.

Bermudez, J. L. (2003). The Domain of Folk Psychology. *Royal Institute of Philosophy Supplement*, *53*, 25–48.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, *11*(2), 49–57.

Bukowski, H., & Samson, D. (2017). New Insights into the Inter-Individual Variability in

Perspective Taking. *Vision*, *1*(1), 8.

Burnham, D. K., & Harris, M. B. (1992). Effects of Real Gender and Labeled Gender on Adults' Perceptions of Infants. *Journal of Genetic Psychology*, *15*(2), 165–183.

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

Carruthers, P. (2017). Valence and Value. *Philosophy and Phenomenological Research*.

Chalik, L., Leslie, S.-J., & Rhodes, M. (2017). Cultural context shapes essentialist beliefs about religion. *Developmental psychology*, *53*(6), 1178–1187.

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

Condry, J. C., Ross, D. F., Condry, J. C., & Ross, D. F. (1985). Sex and Aggression : The Influence of Gender Label on the Perception of Aggression in Children Development *Child Development*, *56*(1), 225–233.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: using ethnicity to disambiguate potentially threatening individuals. *Journal of personality and social psychology*, *83*(6), 1314–1329.

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.

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.

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.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21.

del Pinal, G., & Spaulding, S. (n.d.). Conceptual Centrality and Implicit Bias. *Mind & Language*.

Diesendruck, G., Goldfein-Elbaz, R., Rhodes, M., Gelman, S., & Neumark, N. (2013). Cross-cultural differences in children's beliefs about the objectivity of social categories. *Child development*, *84*(6), 1906–17.

Donders, N. C., Correll, J., & Wittenbrink, B. (2008). Danger stereotypes predict racially biased attentional allocation. *Journal of Experimental Social Psychology*, *44*(5), 1328–1333.

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

Drwecki, B. B., Moore, C. F., Ward, S. E., & Prkachin, K. M. (2011). Reducing racial disparities in pain treatment: The role of empathy and perspective-taking. *PAIN*, *152*(5), 1001–1006.

Dweck, C. S., Chiu, C., & Hong, Y. (1995). Implicit Theories and Their Role in Judgments and Reactions: A Word From Two Perspectives. *Psychological Inquiry*, *6*(4), 267–285.

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.

Fiske, S. T. (2015). Intergroup biases: A focus on stereotype content. *Current Opinion in Behavioral Sciences*, *3*(April), 45–50.

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.

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.

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology*, *53*(1), 25–44.

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).

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*.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford University Press, USA.

Gelman, S. A., & Roberts, S. O. (2017). How language shapes the cultural inheritance of categories. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(30), 7900–7907.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious. *Cognition*, *38*(3), 213–244.

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.

Gilbert, S. J., Swencionis, J. K., & Amodio, D. M. (2012). Evaluative vs. trait representation in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal cortex. *Neuropsychologia*, *50*(14), 3600–3611.

Godfrey-Smith, P. (1991). Signal, Decision, Action. *The Journal of Philosophy*, *88*(12), 709.

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

Goldstein, J., & Schweber, N. (2014, July 19). Man's Death After Chokehold Raises Old Issue for the Police. *The New York Times*, p. A1. New York, NY.

Goodman, G. S., Golding, J. M., & Haith, M. M. (1984). Jurors' Reactions to Child Witnesses. *Journal of Social Issues*, *40*(2), 139–156.

Goodman, G. S., Golding, J. M., Helgeson, V. S., Haith, M. M., & Michelli, J. (1987). When a child takes the stand: Jurors' perceptions of children's eyewitness testimony. *Law and Human Behavior*, *11*(1), 27–40.

Gopnik, A., & Wellman, H. M. (1992). Why the Child's Theory of Mind Really Is a Theory. *Mind & Language*, *7*(1–2), 145–171.

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

Gordon, R. M. (1986). Folk Psychology as Simulation. *Mind & Language*, *1*(2), 158–171.

Hammond, M. D., & Cimpian, A. (in press). Investigating the cognitive structure of stereotypes: Generic beliefs about groups predict social judgments better than statistical beliefs. *Journal of Experimental Psychology: General.*

Haslam, N., Bastian, B., & Bissett, M. (2004). Essentialist Beliefs about Personality and Their Implications. *Personality and Social Psychology Bulletin*, *30*(12), 1661–1673.

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

Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology*, *39*(1), 113–127.

Haslam, N., Rothschild, L., & Ernst, D. (2002). Are essentialist beliefs associated with prejudice? *British Journal of Social Psychology*, *41*(1), 87–100.

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.

Hehman, E., Volpert, H. I., & Simons, R. F. (2014). The N400 as an index of racial stereotype accessibility. *Social Cognitive and Affective Neuroscience*, *9*(4), 544–552.

Heyes, C. (2014). False belief in infancy: A Fresh Look. *Developmental science*, 1–13.

Hirschfeld, L. A., & Gelman, S. A. (1997). What young children think about the relationship between language variation and social difference. *Cognitive Development*, *12*(2), 213–238.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in psychology*, *3*, 96.

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.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687–701.

Hudson, M., Nicholson, T., Ellis, R., & Bach, P. (2016). I see what you say: Prior knowledge of others' goals automatically biases the perception of their actions. *Cognition* 146, 245-250.

Icard, T. (2016). Subjective Probability as Sampling Propensity. *Review of Philosophy and Psychology*, *7*(4), 863–903.

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

Keller, J. (2005). In Genes We Trust: The Biological Component of Psychological Essentialism and Its Relationship to Mechanisms of Motivated Social Cognition. *Journal of Personality and Social Psychology*, *88*(4), 686–702.

Killen, M., Lynn Mulvey, K., Richardson, C., Jampol, N., & Woodward, A. L. (2011). The accidental transgressor: morally-relevant theory of mind. *Cognition*, *119*(2), 197–215.

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.

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. *Neuron*, *79*(5), 836–848.

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect. *Psychological science*, *17*(5), 421–7.

Levy, S. R., Plaks, J. E., Hong, Y., Chiu, C., & Dweck, C. S. (2001). Static Versus Dynamic Theories and the Perception of Groups: Different Routes to Different Destinations. *Personality and Social Psychology Review*, *5*(2), 156–168.

Levy, S. R., Stroessner, S. J., & Dweck, C. S. (1998). Stereotype formation and endorsement: The role of implicit theories. *Journal of Personality and Social Psychology*, *74*(6), 1421–1436.

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. a. (2008). Theory of mind development in Chinese children: a meta-analysis of false-belief understanding across cultures and languages. *Developmental psychology*, *44*(2), 523–31.

Lochmann, T., & Deneve, S. (2011). Neural processing as causal inference. *Current Opinion in Neurobiology*, *21*(5), 774–781.

Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology*, *23*(1), 77–87.

Macrae, C. N., Stangor, C., & Milne, A. B. (1994). Activating Social Stereotypes: A Functional Analysis. *Journal of Experimental Social Psychology*, *30*(4), 370–389.

Mason, M. F., Cloutier, J., & Macrae, C. N. (2006). On construing others: Category and stereotype activation from facial cues. *Social Cognition*, *24*(5), 540.

Mason, M. F., & Macrae, C. N. (2004). Categorizing and individuating others: the neural substrates of person perception. *Journal of cognitive neuroscience*, *16*(10), 1785–1795.

Mcglothlin, H., & Killen, M. (2006). Intergroup Attitudes of European American Children Attending Ethnically Homogeneous Schools. *Child Development*, *77*(5), 1375–1386.

McGlothlin, H., & Killen, M. (2010). How social experience is related to children's intergroup attitudes. *European Journal of Social Psychology*, *40*(4), 625–634.

Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology*, *24*(3), 603–613.

Morton, A. (1996). Folk Psychology is not Predictive. *Mind*, *105*(417), 119–137.

Mueller-Johnson, K., Toglia, M. P., Sweeney, C. D., & Ceci, S. J. (2007). The perceived credibility of older adults as witnesses and its relation to ageism. *Behavioral Sciences & the Law*, *25*(3), 355–375.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and*

understanding other minds. Clarendon Press/Oxford University Press.

Ondobaka, S., Kilner, J., & Friston, K. (2015). The role of interoceptive inference in theory of mind. *Brain and Cognition*.

Palmer, C. J., Seth, A. K., & Hohwy, J. (2015). The felt presence of other minds: Predictive processing, counterfactual predictions, and mentalising in autism. *Consciousness and Cognition*, *36*, 376–389.

Peeters, G. (1983). Relational and informational patterns in social cognition. *Current issues in European social psychology*, *1*, 201–237.

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.

Pettigrew, T. F., & Tropp, L. R. (2000). Does intergroup contact reduce prejudice? Recent meta-analytic findings. *Reducing prejudice and discrimination*, *93*, 114.

Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology*, *38*(6), 922–934.

Prentice, D. A., & Miller, D. T. (2007). Psychological essentialism of human categories. *Current Directions in Psychological Science*, *16*(4), 202–206.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*, 79–87.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, *33*(1), 12–21.

Rhodes, M., & Gelman, S. A. (2009). A developmental examination of the conceptual structure of animal, artifact, and human social categories across two cultural contexts. *Cognitive psychology*, *59*(3), 244–74.

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.

Rhodes, M., & Mandalaywala, T. M. (2017). The development and developmental consequences of social essentialism. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(4), e1437.

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology*, *9*(4), 283–294.

Ross, L. (1977). The Intuitive Psychologist And His Shortcomings: Distortions in the Attribution Process. *Advances in Experimental Social Psychology*.

Rothbart, M., Evans, M., & Fulero, S. (1979). Recall for confirming events: Memory processes and the maintenance of social stereotypes. *Journal of Experimental Social Psychology*, *15*(4), 343–355.

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.

Samson, D., Apperly, I., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1255–1266.

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology: General*, *141*(3), 433–438.

Scholl, B. J., & Leslie, A. M. (2001). Minds, modules, and meta-analysis. *Child development*, *72*(3), 696–701.

Segall, G., Birnbaum, D., Deeb, I., & Diesendruck, G. (2015). The intergenerational transmission of ethnic essentialism: *how* parents talk counts the most. *Developmental Science*, *18*(4), 543–555.

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.

Shahaeian, A., Peterson, C. C., Slaughter, V., & Wellman, H. M. (2011). Culture and the sequence of steps in theory of mind development. *Developmental psychology*, *47*(5), 1239–1247.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychological review*, *119*(1), 120–54.

Spaulding, S. (2017). Do you see what I see? How social differences influence mindreading. *Synthese*, 1–22.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28.

Spratling, M. W. (2013). Distinguishing theory from implementation in predictive coding accounts of brain function. *Behavioral and Brain Sciences*, *36*(3), 231–232.

Spratling, M. W. (2016). Predictive coding as a model of cognition. *Cognitive Processing*, *17*(3), 279–305.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811.

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

Surtees, A., Apperly, I., & Samson, D. (2013). The use of embodied self-rotation for visual and spatial perspective-taking. *Frontiers in human neuroscience*, *7*(November), 698.

Surtees, A., Butterfill, S., & Apperly, I. (2012). Direct and indirect measures of Level-2 perspective-taking in children and adults. *The British journal of developmental psychology*, *30*(Pt 1), 75–86.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., Goodman, N. D., Xu, F., Tenenbaum, J. B., et al. (2011). How to grow a mind: statistics, structure, and abstraction. *Science (New York, N.Y.)*, *331*(6022), 1279–85.

Todd, A. R., Galinsky, A. D., & Bodenhausen, G. V. (2012). Perspective Taking Undermines Stereotype Maintenance Processes: Evidence from Social Memory, Behavior Explanation, and Information Solicitation. *Social Cognition*, *30*(1), 94–108.

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.

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.

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.

Van Knippenberg, A., Dijksterhuis, A., & Vermeulen, D. (1999). Judgement and memory of a criminal act: the effects of stereotypes and cognitive load. *European Journal of Social Psychology*, *29*(2–3), 191–201.

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

Wardlow, L. (2013). Individual differences in speakers' perspective taking: the roles of executive control and working memory. *Psychonomic bulletin & review*, *20*(4), 766–72.

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.

Wellman, H. M., Fang, F., Liu, D., Zhu, L., & Liu, G. (2006). Scaling of theory-of-mind understandings in Chinese children. *Psychological Science*, *17*(12), 1075–1081.

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

Westra, E. (2017). Character and theory of mind: an integrative approach. *Philosophical Studies*, 1–25.

Wigboldus, D. H. J., Dijksterhuis, A., & van Knippenberg, A. (2003). When stereotypes get in the way: Stereotypes obstruct stereotype-inconsistent trait inferences. *Journal of Personality and Social Psychology*, *84*(3), 470–484.

Wigboldus, D. H. J., Sherman, J. W., Franzese, H. L., & van Knippenberg, A. (2004). Capacity and Comprehension: Spontaneous Stereotyping Under Cognitive Load. *Social Cognition*, *22*(3), 292–309.

Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of competence or morality. *Journal of Personality and Social Psychology*, *67*(2), 222–232.

Wright, A. M., & Holliday, R. E. (2005). Police officers' perceptions of older eyewitnesses. *Legal and Criminological Psychology*, *10*(2), 211–223.

Zawidzki, T. W. (2013). *Mindshaping: A New Framework for Understanding Human Social*

Cognition. MIT Press.