**The Embedded and Extended Character Hypotheses**

Mark Alfano & Joshua August Skorburg

Forthcoming in J. Kiverstein (ed.), *Philosophy of the Social Mind*. Routledge.

Draft: Please do not quote or cite without permission

**Introduction**

This chapter brings together two erstwhile distinct strands of philosophical inquiry: the extended mind hypothesis and the situationist challenge to virtue theory. According to proponents of the extended mind hypothesis, the vehicles of at least some mental states (beliefs, desires, emotions) are not located solely within the confines of the nervous system (central or peripheral) or even the skin of the agent whose states they are. When external props, tools, and other systems are suitably integrated into the functional apparatus of the agent, they are partial bearers of her cognitions, motivations, memories, and so on. According to proponents of the situationist challenge to virtue theory, dispositions located solely within the confines of the nervous system (central or peripheral) or even the skin of the agent to whom they are attributed typically do not meet the normative standards associated with either virtue or vice (moral, epistemic, or otherwise) because they are too susceptible to moderating external variables, such as mood modulators, ambient sensibilia, and social expectation signaling.

We here draw on both of these literatures to formulate two novel views – the embedded and extended character hypotheses – according to which the vehicles of not just mental *states* but longer-lasting, wider-ranging, and normatively-evaluable *agentic dispositions* are sometimes located partially beyond the confines of the agent’s skin. Put another way, we will examine the ways in which moral and intellectual character are dependent on, or constituted by, one’s social environment. We believe this is a natural but underexplored next step in both the extended mind and situationist research programs.

Virtues and vices can be understood as dispositions to token a suite of occurrent mental states and engage in signature behaviors in response to configurations of external and internal variables. If those mental states are sometimes extended, perhaps the dispositions to have them are too. Presumably, the dispositions don’t extend in every case, just as the states don’t extend in every case. Perhaps some people are honest all on their own. Perhaps some people are intelligent all on their own. But if our suggestion is on the right track, in some cases, a person is honest because (among other things) she is suitably integrated with props, tools, or other people outwith her brain and body. Likewise, if our suggestion is on the right track, in some cases, a person is intelligent because (among other things) he is suitably integrated with props, tools, or other people outwith his brain and body.

Here is the plan for this paper. We begin by briefly explaining the situationist challenge, with special emphasis on the ways in which social expectations influence people’s thought, feeling, and behavior. Next, we explore two phenomena in which character, or at least the manifestation of character, is dependent on or constituted by such influence: stereotype threat and friendship. We argue that, construed correctly, people who are susceptible to stereotype threat have *socially embedded intellectual character* and that people who have robust friendships have *socially extended moral character*. The similarities and differences between these two phenomena are then used to help map the conceptual space of embedded and extended character, with reference to dynamical systems theory. We conclude with a few brief remarks situating the embedded and extended character hypotheses in the larger context of skepticism about traits and character.

**1. The situationist challenge**

At least some neo-Aristotelian virtue-theoretic views (e.g., Foot 2003) are proudly naturalistic. On the one hand, this makes them methodologically attractive to those of us with independent naturalistic commitments. On the other hand, it means that these views face empirical critiques that non-naturalistic normative theories can sidestep. The situationist challenge to virtue ethics began when Harman (1999) and Doris (2002) argued that the dominant neo-Aristotelian conception of moral virtue was empirically inadequate. Alfano (2012, 2013a, 2014c) took the challenge a step further, arguing that virtue epistemology faces an empirical challenge.[[1]](#footnote-1) Contemporary virtue ethicists and virtue epistemologists typically combine the ancients’ idea that virtues are admirable, cross-situationally consistent traits of character acquired through habituation, with the more modern egalitarian assumption that almost everyone can – at least at an early enough stage of their life – become virtuous. Unfortunately, seemingly trivial and normatively irrelevant situational influences such as ambient sounds, ambient smells, ambient light levels, mood elevators, mood depressors, and social expectation signaling seem to be at least as powerful predictors and explainers of someone’s thoughts, motivations, feelings, deliberations, and behaviors as any traits they may have (Alfano 2013a, 2013b). Given the eminently plausible assumptions that parents and other educators try their best to instill virtue in their wards and that the vast majority of adults aim to be virtuous, these results from social psychology are worrisome. It’s not for lack of trying that we fail to be virtuous in the traditional sense.

Situational influences on moral and intellectual conduct are many and diverse. The best-supported model of situational influences in social psychology seems to be the “Situational Eight DIAMONDS” model (Rauthmann et al. 2014), which stands for:

* *Duty*: a job must be done,
* *Intellect*: the situation affords a chance to demonstrate one’s intellect,
* *Adversity*: one reacts either prospectively or retrospectively to blame,
* *Mating*: one modulates one’s behavior because potential romantic partners are present,
* *pOsitivity*: the situation is potentially enjoyable,
* *Negativity*: the situation is potentially unenjoyable or anxiety-provoking,
* *Deception*: it is possible to deceive someone, and
* *Sociality*: social interaction is possible.

Together, these eight kinds of situational influences account for a large amount of the variance in people’s behavior (24-74%), much more than trait dimensions do (3-18%). Here, we focus on a type of situation that harnesses several of these situational influences in what can fairly be described as a perfect storm: the signaling of expectations by another person with whom one may continue to have ongoing social contact. Your boss asks you to do something you find problematic. Your friend raises her eyebrows expectantly. Your child gives you a plaintive look. Such situations invoke duty: we tend to feel that we owe people with whom we have this sort of relationship particular duties. Such situations invoke adversity: we feel put upon. They invoke mating when the other person is a (potential) sexual partner. They invoke positivity to the extent that we find it rewarding to meet the expectations others have for us (a common preference). They invoke negativity to the extent that we find it unenjoyable or anxiety-provoking to flout someone’s expectations (again, a common preference). They involve sociality by definition.

Much of the groundbreaking social psychology of the second half of the twentieth century investigated the power of expectation confirmation. The most dramatic demonstration was of course the Milgram paradigm (1974), in which roughly two thirds of participants were induced to put what they thought was 450 volts through another participant (actually an actor who was in on the experiment) three times in a row. While there were many important features of this study, the key upshot was that the participants were willing to do what they should easily have recognized was deeply immoral based on the say-so of a purported authority figure. Blass (1999) shows in a meta-analysis that Milgram’s results were no fluke: they have been replicated all around with the world with populations of diverse age, gender, and education level.

Another example of the power of social expectations is the large literature on bystander apathy (Darley & Latané 1968; Latané & Nida 1981). It turns out that the more bystanders are present in an emergency situation, the lower the chances that even one of them will intervene. What seems to happen in such cases is that people scan others’ immediate reactions to help themselves determine what to do. When they see no one else reacting, they decide not to intervene either; thus everyone interprets everyone else’s moment of deliberation as a decision not to intervene.

Reading off others’ expectations and acting accordingly doesn’t always lead to bad outcomes, though. Recent work on *social proof* shows that the normative valence of acting in accordance with expectations depends on what’s expected. For instance, guests at a hotel are 40% more likely to conserve water by not asking for their towels to be washed if they read a message that says, ‘75% of the guests who stayed in this room participated in our resource savings program by using their towels more than once’ than one that says, ‘You can show respect for nature and help save the environment by reusing towels during your stay’ (Goldstein, Cialdini, & Griskevicius 2008).

Psychologists and behavioral economists have also investigated the effect of subtle, thoroughly embodied, social distance cues on moral behavior. In a string of fascinating studies, it’s been shown that people are more willing to share financial resources (Burnham 2003; Burnham & Hare 2007; Rigdon, Ishii, Watabe, & Kitayama 2009), less inclined to steal (Bateson, Nettle, & Roberts 2006), and less disposed to litter (Ernest-Jones, Nettle, & Bateson 2011) when they are “watched” by a representation of a face. The face can be anything from a picture of the beneficiary of their behavior to a cartoon robot’s head to three black dots arranged to look like eyes and a nose.[[2]](#footnote-2)

One way to understand what’s going on in these scenarios is that people’s character is flimsy, that their behavior, thought, and feeling is largely explained by seemingly trivial and normatively irrelevant situational influences.[[3]](#footnote-3) But another way to understand what’s going on in these scenarios is that people’s character is thoroughly embedded in a social context or even partially constituted by the social context. When the bonds holding them in that context are tight and modally robust, perhaps it makes sense to think of their character as *extending* out into the social environment. When the bonds holding them in that context are relatively looser and less modally robust, perhaps it makes sense to think of them as merely *embedded* in the social context, with some properties that are metaphysically independent of it.[[4]](#footnote-4) To quote Shepard (1984, p. 436) slightly out of context, our contention here is that perception – including social perception – is “*externally guided hallucination.*” Our friends, parents, children, enemies, co-workers, bosses, and underlings *make us up*. They revise their view only when the hallucination they’ve projected onto us doesn’t induce us to act as expected.

In any event, the only point we need to make here is that people’s dispositions to manifest what are traditionally considered character traits are highly socially dependent, to different extents and in different ways. The devil is in the details. In the remainder of this chapter, we explore some of those details to argue that character is sometimes embedded and sometimes extended.

This way of understanding the Milgram paradigm and related phenomena accords with Alfano’s (2014a) suggestion: instead of thinking of virtue in the traditional way, as a monadic property of an individual agent, perhaps we should think of it as a relation between the agent and another agent, between the agent and a broader social milieu, or among the agent, a social milieu, and an asocial environment. When the agent is suitably integrated with these externalia, the situational influences that otherwise interfere with or thwart the possession or expression of virtue might systematically support it. For instance, Alfano (2013a) argues that, even if people lack virtues as they are conceived in neo-Aristotelian orthodoxy, they may have “factitious” or artificial virtues that simulate traditional virtues but are partially externally located. A factitious virtue is supported both by the agent’s self-concept (thinking of herself as, say, generous) and, more importantly, by the social expectations signaled to her by her friends, family, colleagues, and acquaintances (realizing that others think of her as generous, expect her to act accordingly, and knowing that she knows this about them). On this view, virtue inheres “in the interstices between the person and her world. The object that possesses the virtue [is] a functionally and physically extended complex comprising the agent, her social setting, and her asocial environment” (2013a, p. 185). In other words, the embedded and extended character hypotheses are the best hopes for virtue ethics and virtue epistemology to defend against situationism.

Under what conditions is it the case not just that an agent’s social environment *causes* her to think, feel, desire, deliberate, and behave as a virtuous person would, but also that these environmental features *are part of* her virtue? We contend that when an agent is functionally integrated through ongoing feedback loops with her social environment, the environment doesn’t just causally influence her but becomes part of her character, for good or ill (Alfano 2015). What her good friends, her romantic partners, her domestic abuser – anyone with whom she has a deep, ongoing relationship – expects influences what she thinks they expect, which influences (among other things) what she expects of herself, the reasons she’s sensitive to, her levels of motivation, and her behaviors; this in turn confirms and strengthens (or disconfirms and undermines) her associates’ expectations, which are again transmitted to her, further shaping her thought, feeling, deliberation, and behavior, which again influences her associates’ expectations, and so on.

In the next section, we explore two examples of this kind of functional integration with the social environment. We believe these cases provide intuitive support for the hypotheses of embedded and extended character.

**2 Two examples: stereotype threat and friendship**

In the previous section, we introduced the situationist challenge, along with Alfano’s claim that the challenge can be deflected by an interactionist theory of virtue, which makes character dependent on the environment, especially the social environment. If this is on the right track, then perhaps long-lasting, wide-ranging, and normatively-evaluable *agentic dispositions* are sometimes best understood as embedded or even as extended into the social environment. In the present section, we offer an example which we believe helps to illustrate how some components of one’s intellectualcharacter — that is, agentic dispositions to perform in certain characteristic ways on intellectual tasks — might best be understood as embedded in this way. We then turn to friendship to argue for the more audacious claim that moral character is sometimes extended.

*2.1 Stereotype threat and embedded character*

First, a few brief remarks are in order regarding the connection between stereotypes and intellectual character. If there are such things as intellectual character and intelligence, then intelligence is surely one of the most important components of intellectual character (along with curiosity and intellectual humility). If you have to predict how two people will perform on a problem-solving task, it makes sense to bet on the one who’s more intelligent. If intelligence is a psychologically credible construct,[[5]](#footnote-5) then someone’s performance on a given problem-solving task can be partially explained by reference to their intelligence.[[6]](#footnote-6) To what does this construct refer, though? One tempting answer, which has often and notoriously been proposed by researchers and politicians with racist and sexist agendas, is that someone’s intelligence inheres inside them. Perhaps it’s hardwired into the brain. Perhaps it’s hard-coded by genetics. Perhaps it ossifies during early childhood under the influence of parenting and culture. Whatever its details, the denouement is almost always the same: some people are more intelligent than other people. They always will be. Their intelligence is their own and inflexible. In the United States, it almost always turns out that the more intelligent group is white males (though who counts as ‘white’ changes over time), while other groups (especially blacks, but also Latinos and women) are less intelligent. A good example is Herrnstein & Murray (1996), though other examples can easily be multiplied.

It’s not hard to see how such interpretations of the psychological evidence both draw on and reaffirm existing stereotypes. Stereotypes characterize members of a group as typically having particular configurations of properties in virtue of their group membership, which confers on them an essence. If our discussion of social dependence in the previous section is on the right track, however, stereotype-influenced perception maybe a kind of hallucination. Someone who typically faces expectations based on their group membership may end up acting in a way that confirms the stereotype because they find it is too burdensome and futile to try to oppose it.

To better understand how this might play out, Taylor and Walton (2011) ask us to imagine a black student at a predominantly white university enrolled in what is widely known to be an intellectually challenging course. Further, this course is meant to diagnose which students can advance to the next required course in the series. When it comes time for an exam in this setting, the student’s awareness of the negative stereotypes associated with his racial group may be heightened. He may worry that any confusion he feels, any questions he asks, any mistakes he makes, will serve to confirm the negative stereotypes associated with his racial group. This is the “social-psychological predicament” that, in a series of seminal experiments, Steele and Aronson (1995) dubbed *stereotype threat*: “the existence of such a [negative] stereotype means that anything one does or any of one’s features that conform to it make the stereotype more plausible as a self-characterization in the eyes of others, and perhaps even in one’s own eyes” (p. 795).

In one study, black and white students were given an exam consisting of questions from the verbal section of the GRE. Differences in individual skill level were controlled for by reference to SAT verbal scores. In the stereotype threat condition, the exam was described as diagnostic of intellectual ability. In the control condition, the exam was described as a problem solving task that was not diagnostic of intellectual ability. What makes the experimental condition threatening is the extant stereotype that black students underperform in school. Thus, a poor individual performance by a black student would be perceived as a confirmation of this stereotype. Being consciously aware of the stereotype and one’s own relation to it, Steele and Aronson hypothesized, would lead to a decrease in performance. And indeed, they found that “Black participants performed worse than White participants when the test was presented as a measure of their ability, but improved dramatically, matching the performance of Whites, when the test was presented as less reflective of ability” (p. 801).

In order to test whether this effect was related specifically to stereotypes, rather than something like test anxiety, Steele and Aronson conducted another experiment. Before taking the same test, one group was required to fill out a demographic questionnaire. In the second group, participants were required to fill out the questionnaire after completing the exam. If the threat experienced in the first study was racially specific, then students who were required to call to mind their membership in a negatively stereotyped racial group before taking the exam should score worse than those who were required to do so after taking the same exam. Not only did Steele and Aronson find this to be the case, but they also found that “priming racial identity depressed Black participants’ performance on a difficult verbal test *even when the test was not presented as diagnostic of intellectual ability*”(p. 808, emphasis added). Thus, a basic demographic survey seems to be sufficient to activate the kind of threat that negatively affects intellectual performance. Indeed, in questionnaires, participants in the stereotype threat condition reported greater cognitive activation of racial stereotypes, greater concerns about their intellectual ability, greater tendencies to make excuses in advance for their test performance, and a greater reluctance to have racial identity linked to performance, than participants in the non-threat condition (p. 805).[[7]](#footnote-7)

All of this suggests that, in the case of negatively stereotyped racial minorities, the more reliably present the negative stereotypes, the more likely the test-taker will encounter low expectations. In a social environment with a highly reliable stereotypic signal – a society where it is frequently borne in on people what groups they do and don’t belong to, and what is expected of them in virtue of their group membership – the hallucinatory self-confirmation of stereotypes will be rife. This predictably leads to a decrease in performance. While the precise mechanisms responsible for this performance decrease are not known,[[8]](#footnote-8) there is little doubt as to the reality and efficacy of stereotype threat for racial minorities in academic contexts.[[9]](#footnote-9)

There are many ways one can be reminded of one’s membership in a negatively stereotyped group, and hence, many ways for stereotype threat to be activated. Demographic questionnaires, preference questionnaires, music, movies, video games, television shows, political institutions, peer chatter, jokes, and other relevant examples could be easily multiplied. The environment the test-taker inherits, that is, is one in which racial stereotypes are readily and reliably present. We can think of the sum total of these features as the level of threat present in the environment. And we can think of the sum total of the worries, anxieties, and distractions the test-taker experiences as the perceived level of threat. As a wealth of empirical research has shown, these worries, anxieties, and distractions can lead to a decrease in intellectual performance. When the test-taker’s perceived level of threat is high, he is more likely to perform poorly on the exam. When this disposition to perform poorly on exams in threatening conditions is manifest, it then provides feedback to the extant stereotypes in his environment: “Blacks consistently score lower on this exam because they are poor students.” If the effects of stereotype threat are not accounted for, then, the exam score starts to look like supporting evidence for the previous statement. As this kind of information builds up, perhaps in the context of a single course, but also over the course of a college career, it feeds forward to create a higher level of threat in the environment. And when the level of threat in the environment is high, the perceived level of threat fed back to the test taker is also heightened. When he faces a similar task, he becomes more likely to underperform as a result of the threatening conditions.

Worse still, it is unlikely that an exemplary performance by our test taker - perhaps even multiple exemplary performances - will do much to drown out the negative stereotypes in his environment. The presence of stereotypes in the environment frequently and reliably influences the test-taker, while the test-taker has very little such influence on the stereotypes present in his environment. His scoring in the ninety-fifth percentile on the midterm will not change stereotypic media portrayal by finals week. That is to say, the feedback loops between stereotypes and their targets are largely asymmetric and unidirectional.

How ought we conceive the relationship between stereotypes and their target’s intellectual character? There are (at least) four relevant routes available. First, one could ignore the phenomenon of stereotype threat and point to standardized test scores as evidence of the inferior intellectual character of racial minorities. On this view, intellectual character is viewed as an individual, innate disposition. Second, one could draw a skeptical conclusion from the social psychological premises: the phenomenon of stereotype threat proves that intellectual character is flimsy, and too susceptible to things like stereotypes, expectations, noises, worries, distractions, anxieties, etc. to be of much use. On this view, intellectual character is perhaps viewed as a convenient fiction. Third, one could hold onto a notion of intellectual character by suggesting that the locus of intellectual character is wider than has been previously assumed. On this view, intellectual character might sometimes *depend on* the stereotypes, expectations, noises, worries, distractions, and anxieties in the social environment. Put another way, the connections and feedback loops between an agent and the relevant features his social environment might be tight enough and reliable enough to think of his intellectual character as *embedded in his social environment*. Fourth, one could take a more radical route still, and claim that intellectual character is quite literally *constituted by* the stereotypes, expectations, noises, worries, distractions, and anxieties in the social environment. On this view, the connections and feedback loops between the agent and his social environment are so tight and so reliable that his intellectual character *extends* to include these features of his social environment as proper components.

We believe the third route is the best for understanding intellectual character in the context of stereotype threat. That is, intellectual character might sometimes *depend on* the stereotypes in the social environment, and the feedback loops between an agent and the relevant features his social environment might be tight and reliable enough to think of his intellectual character as *embedded*. On the one hand, we’re skeptical of views that treat intellectual character as an innate, monadic disposition. On the other hand, its not clear that the feedback loops between an agent and the stereotypes in his social environment are so tight and reliable that his character would *extend* to include them. The signal from the social environment to the target of stereotype threat is relatively reliable, but the feedback from the agent to the environment is much less so. We think the framework of embedding best captures this asymmetric and largely unidirectional relationship.

*2.2 Friendship and extended character*

In the previous section, we argued that the phenomenon of stereotype threat shows that intellectual character is sometimes embedded because the agent and his environment form a coupled system in which the social expectations directed at the target are near-ubiquitous and highly reliable (leading to self-confirming effects), whereas the feedback and behavior from the target to the social environment is highly unreliable. We don’t mean to imply that intellectual character is never extended and at most embedded, but other examples would have to be employed to demonstrate extension. In this section, we go a step further, arguing that moral character is sometimes extended. Our example here is friendship.[[10]](#footnote-10)

Imagine two agents, Ashley and Azim. Ashley and Azim are best friends. They spend as many as three or four days a week with each other. They care deeply about each other – not just about whether the other is suffering or feeling good, not just about whether the other is getting what he or she wants. Beyond these more mundane concerns, Ashley cares about whether Azim is a morally good person, and cares whether Azim thinks that she is a morally good person. Likewise, Azim cares whether Ashley is a morally good person, and cares whether Ashley thinks of him as a morally good person. Moreover, Ashley knows that Azim cares about her and her opinion of him; likewise, Azim knows that Ashley cares about him and his opinion of her. Indeed, Azim knows that Ashley knows that Azim cares about her and her opinion of him, and Ashley knows that Azim knows that Ashley cares about him and his opinion of her. There may even be common knowledge between them of their caring attachments: he knows she cares, and she knows that he knows that she cares, and he knows that she knows that he knows that she cares, and so on. Insisting on this might seem a bit precious, but we think it’s important, and that it characterizes many real friendships. Imagine how you would feel if your friend said, “I don’t even know whether you care about me.” You might respond, “You may not realize it now, but I do care about you, and it’s important to me not only that you see that, but also that I can rest assured that you see it.”

Like everyone, Ashley and Azim have their flaws, and they’re not foolish enough to think themselves perfect. They rely on each other to – gently, and in a spirit of friendship – point out these flaws from time to time. When Ashley is headed down a particular course of action, she infers from the fact that Azim hasn’t tried to convince her to change course that he approves, or at least doesn’t disapprove too strongly. When Azim is unsure of himself, when he fears that he may have acted badly, he looks to Ashley for reassurance, or at least for lack of condemnation.

In their deliberations, each of them weighs reasons like the rest of us, but they have also internalized each other’s voices. Ashley consults her internal-Azim: What would he tell her to do? How would he feel about her plans? How would he react to her behavior? What emotion would his face register if he were watching right now? Likewise, Azim consults his internal-Ashley: How will he feel if and when he tells her about what he just did? How will she react when he tells her how he feels right now? Their internalized models of each other are imperfect, of course. Everything is. But they’re not too shabby, either. After all, Ashley’s internalized Azim gets updated every time she gets actual feedback from him. If internal-Azim tells her to do one thing but actual-Azim says the opposite, she updates internal-Azim. Likewise, Azim’s internalized Ashley gets updated every time he gets actual feedback from her. If internal-Ashley reacts with approbation but actual-Ashley reacts with shock, he updates internal-Ashley.[[11]](#footnote-11) Along these lines, Adam Morton (2013) argues that what distinguishes moral emotions from garden-variety emotions is that the former essentially involve imagining a perspective from which an emotion is directed at you. For instance, guilt is the state of imagining a perspective from which anger is directed at you; shame is the state of imagining a perspective from which contempt is directed at you. The perspective from which the self-directed emotion emanates can be a desiccated ideal observer, but it can also be one’s internal model of a particular person.

None of this is meant to suggest that Ashley slavishly follows Azim’s or internal-Azim’s advice (or vice versa). Nevertheless, both Ashley and Azim trust each other enough to treat the other’s (dis)approval of an action or plan as a *pro tanto* reason for (against) it. And retrospectively, they treat each other’s (dis)approbation as evidence that an action was right (wrong). Indeed, each of them regards the other’s (dis)approval as both an instrumental and an intrinsic reward (punishment). The instrumental value of others’ good opinion is obvious: they’ll be more inclined to trust and cooperate with you if they think well of you. Beyond that, if they broadcast their view of you, they may induce still others to take up the same opinion. And if they broadcast their view to you, you gain information about how you are – or at least about how you are perceived. Likewise, the instrumental disvalue of others’ bad opinion is obvious: they’ll be less inclined to trust and cooperate with you, and more inclined to sanction you if they think ill of you. If they broadcast a negative view of you, they may induce still others to take up the same opinion. Interestingly, if they broadcast it to you, you still gain potentially useful information about how you are – or at least about how you are perceived.[[12]](#footnote-12)

But the (dis)approbation of others may have intrinsic worth as well. As Philip Pettit (1995) points out, among the things people (dis)value is the (dis)approbation of others. This moral psychological fact can be given a cynical reading, on which people are vain esteem-seekers. It can also, though, be given a more positive reading, on which the good opinion of a good (enough) person is intrinsically valuable. This is perhaps most obvious when one considers that the good opinion of a bad (enough) person is often regarded as an insult.

Furthermore, just as there are multiple levels of mutual knowledge between Ashley and Azim (she knows he cares about her, and he knows that she knows he cares about her, and she knows that he knows that she knows he cares about her, and so on), so they often find themselves in episodes where they direct higher-order emotions at one another. Robert Roberts (2013) explores the ways in which emotions and emotional feedback loops strengthen and desiccate such relationships as friendship, enmity, civility, and incivility. For example, consider a sister who generously and in a spirit of friendship gives her brother her own ticket to a concert that he would like to attend. He feels the emotion of gratitude for this gift, which he expresses with a token of thanks. Satisfied that her generosity has hit its mark, she is “gratified by his gratitude. […] And he may in turn be gratified that she is gratified by his gratitude” (p. 137). Despite the fact that this is a tiny schematic example, it plausibly contains a fourth-order emotion (he is gratified that she is gratified that he is gratified that she was generous). Such episodes are, in Roberts’s view, constitutive of friendship and other normative personal relationships (pp. 140-1). They naturally fit into the framework discussed here. Not only is the friendship between Ashley and Azim partly constituted by such emotional ping-pong, but the ongoing feedback such episodes embody makes each of their moral dispositions more modally robust. When Azim plans, he is guided by his internal Ashley. When he acts, he often gets direct feedback from her. If he acts badly (in her eyes), she makes him know it. If he continues to act badly (in her eyes), she makes him know that too. Thus, there are multiple opportunities for correction and adjustment built into their relationship. Azim may never avail himself of the fourth or fifth or sixth contingent intervention, but *were* he to need it, it would be there. Likewise for Ashley. Unlike in the case of stereotype threat, where the signal in one direction is strong and reliable whereas the signal in the other direction is weak and noisy, friendship (ideally) involves strong reliable signaling (and attuned receiving) in both directions.

Next, consider the truism that one’s possibilities for action are constrained by one’s modal knowledge. If you think that something is impossible – even if it’s not – you can’t try to accomplish it. Ashley’s impression of her own possibilities for action (and thus the range of actions she can actually take) is expanded by Azim’s confidence in her. When he signals that he thinks, trusts, or hopes that she can do X, he opens up the possibility of X for her. Likewise, Azim’s impression of his own possibilities for action (and thus the range of actions he can actually take) is expanded by Ashley’s confidence in him. When she signals that she thinks, trusts, or hopes that he can do Y, she opens up the possibility of Y for him. As Victoria McGeer (2008) reminds us, human motivation is often complicated and confusing. Sometimes we don’t know what we really desire, like, or love. Sometimes, we forget what we really value. Sometimes, we don’t know what we’re capable of. In those cases, it’s helpful to refer to a normative lodestone, a model of good conduct. Here we quote at length:

For help in this regard, we are sometimes encouraged to look outside ourselves for role models, finding in others’ thoughts and actions laudable patterns on which to fashion our own. And this may serve us pretty well. However, something similar can occur, often more effectively, through the dynamic of hopeful scaffolding. Here we look outside ourselves once again; but instead of looking for laudable patterns in others’ behavior, what we find instead are laudable patterns that others see – or prospectively see – in our own. We see ourselves as we might be, and thereby become something like a role model for ourselves. The advantage in this is clear: Instead of thinking, ‘I want to be like her,’ – i.e., like someone else altogether – the galvanizing thought that drives us forward is seemingly more immediate and reachable: ‘I want to be as she already sees me to be’ (pp. 248-9; see also James 1978/1896).

Hope of this kind might best be construed not as feedback but as feedforward: Ashley’s model of Azim is robust to his momentary self-doubt, and when she signals her ongoing confidence in him, she nudges him back towards a confident equilibrium (and, once again, vice versa). This recalls Shephard’s (1984) characterization of perception – including, we want to emphasize, social perception – as externally-guided hallucination. Importantly, though, such hallucinations sometimes influence the empirical facts on the ground and thereby the sensory data. This is an example of what Tad Zawidzki (2013; see also Mameli 2001) refers to as ‘mind-shaping’.

If these remarks on friendship are on the right track, they show how friendship can be modeled as a coupled system with strong reliable signaling (and attuned receiving) in both directions. David Wong explores such influences in *Natural Moralities* (2006, pp. 133-7). Drawing on Confucian ethics, he explores the ways in which children learn norms, rules, and values through ongoing interactions with family members. This learning is sometimes explicit but more often implicit. It essentially depends on the existence of regular, cross-situational, and extensive interactions in a trusting relationship embodying (ideally) shared norms. But Wong emphasizes that such interpenetration of moral character occurs not only in childhood but also in adulthood, arguing that others

help to shape and crystallize traits and desires that are especially congruent with our most important ends. Or rather, there are often times when increased self-knowledge merges with the crystallization of a trait or desire – when, for instance, understanding oneself better is at the same time making more determinate tendencies and impulses within one’s character that are in some degree inchoate. I have in mind ways that others can help us through some insight as to what our “real” feelings and motivations are, where that insight is partly an accurate portrayal of what is already there but also helps to reinforce and make more determinate what those feelings and motivations are. A friend who points out to a person that she is more compassionate than she understands herself to be, who points to certain recurring instances of compassionate behavior as evidence, may not just be pointing to what is already there but crystallizing and making more motivationally salient that trait in his friend (p. 136).

Friendship and close relationships on these accounts seem importantly different from stereotype threat. In the case of stereotype threat, there’s not much the target of threat can do to influence the stereotypes in his environment. In the case of friendship, by contrast, Ashley can pull Azim’s levers, and Azim can pull Ashley’s. They are sensitive to each other in real time and respond differentially to the other’s behavior and intentions.

Azim’s expectations for himself, his self-knowledge, his understanding of which actions are available to him, his motivation, the reasons that appear salient to him and their weights, and his deliberative strategies – all of these are influenced in a systematic and ongoing way by Ashley. Likewise, Ashley’s expectations for herself, her self-knowledge, her understanding of which actions are available to her, her motivation, the reasons that appear salient to her and their weights, and her deliberative strategies – all of these are influenced in a systematic and ongoing way by Azim. [[13]](#footnote-13)

Because there are multiple feedback contingencies for both of them, their dispositions become modally robust. Or, in the language of dynamical systems theory (e.g., Palermos 2014), they erect attractors and repellors. What Ashley considers bad behavior, thought, feeling, etc. is a repellor for Azim because when he veers that way, she gives him multiple, increasingly strong nudges back towards equilibrium. What Ashley considers good behavior, thought, feeling, etc. is an attractor for Azim because when he starts acting, thinking, and feeling in these ways she gives him ongoing feedback that reinforces these dispositions. Likewise, what Azim considers bad behavior, thought, feeling, etc. is a repellor for Ashley because when she veers that way, he gives her multiple, increasingly strong nudges back towards equilibrium. What Azim considers good behavior, thought, feeling, etc. is an attractor for Ashley because when she starts acting, thinking, and feeling in these ways he gives her ongoing feedback that reinforces these dispositions.

Given the tight coupling and reliable feedback present in cases of friendship, we contend that friendship can be understood as a case of *extended* moral character. In the case of stereotype threat, the feedback loops between agent and environment were largely asymmetric and unidirectional. For this reason, we argued the framework of embedding was appropriate. In robust friendships, however, these feedback loops are much tighter and more reliable. They are largely bidirectional and symmetric, and for this reason, the framework of extension is appropriate.

**3 Conclusion**

To conclude, it will be worthwhile to consider how the arguments advanced here fit into the larger debates about traits and character. Does social psychological evidence dictate that we abandon the idea of robust character traits? Or is there reason to think we should retain a place for character traits in our moral psychology?

This chapter has aimed less at staking out a novel position in the character skepticism debates, and aimed more at precisely mapping the ways in which character is sometimes dependent upon, or constituted by, the social environment.[[14]](#footnote-14) On the one hand, we remain skeptical about character if what is meant by character is a strictly monadic, internal, individualist, disposition. On the other hand, we can see a place for character in moral psychology if the concept can be operationalized to include the dependence and constitution relations with the social environment we have described here.

In the end, our discussion raises more questions than it answers. Are we responsible for our own embedded character in the same we that we’re allegedly responsible for our own internal character? If we’re not, who is? Perpetrators of stereotypes? Does this make us, in a way (worrisome or encouraging), our brothers’ and sisters’ keepers? Are we responsible for our own extended character in the same way that we’re allegedly responsible for our own internal character? If we’re not, who is? Our friends? Does this make us, in a way (worrisome or encouraging), our friends’ keepers? The embedded and extended character hypotheses, if true, seem to make us both more vulnerable (regarding our own character) and more responsible (regarding the character of others).

**References**

Alfano, M. (2015). Friendship and the structure of trust. In J. Webber & A. Masala (eds.), *From Personality to Virtue: Essays in the Psychology and Ethics of Character*. Oxford University Press.

Alfano, M. (2014a). What are the bearers of virtues? In Sarkissian & Wright (eds.), *Advances in Experimental Moral Psychology*. 73-90, New York: Continuum.

Alfano, M. (2014b). Stereotype threat and intellectual virtue. In Flanagan & Fairweather (eds.), *Naturalizing Epistemic Virtue*. 155-174, Cambridge University Press.

Alfano, M. (2014c). Extending the situationist challenge to reliabilism about inference. In A. Fairweather (ed.), *Virtue Epistemology Naturalized: Bridges Between Virtue Epistemology and Philosophy of Science*. 103-122. Synthese Library.

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

Alfano, M. (2013b). Identifying and defending the hard core of virtue ethics. *Journal of Philosophical Research*, 38, 233-260.

Alfano, M. (2012). Extending the situationist challenge to responsibilist virtue epistemology. *Philosophical Quarterly*, 62:247, 223-249.

Assad, K., Donnellan, M. B., & Conger, R. (2007). Optimism: An enduring resource for romantic relationships. *Journal of Personality and Social Psychology*, 93(2), 285-97.

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters, 12,* 412-414.

Blass, T. (1999). The Milgram paradigm after 35 years: Some things we now know about obedience to authority. *Journal of Applied Social Psychology, 29(5),* 955-978.

Burnham, T. (2003). Engineering altruism: A theoretical and experimental investigation of anonymity and gift giving. *Journal of Economic Behavior and Organization*, *50*, 133-144.

Burnham, T. & Hare, B. (2007). Engineering human cooperation. *Human Nature*, *18(2)*, 88-108.

Croizet, J., & Claire, T. (1998). Extending the concept of stereotype threat to social class: the intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin*, 24, 588-594.

Darley, J. & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*, 377-383.

Doris, J. (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge University Press

Ernest-Jones, M., Nettle, D., & Bateson, M. (2011). Effects of eye images on everyday cooperative behavior: A field experiment. *Evolution and Human Behavior*, *32(3)*, 172-178.

Foot, P. (2003) *Natural Goodness*. Oxford University Press.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, 35, 472-482.

Harman, G. (1999). Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society*, New Series 199, 316-31.

Hernstein, Richard J., & Murray, Charles. (1996). The Bell Curve: Intelligence and Class Structure in American Life. New York: Free Press.

James, W. (1979/1896). *The Will to Believe*. Harvard University Press.

Kane, M., Hambrick, D., & Conway, A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, 131(1), 66-71.

Kovacs, K., & Conway, A. (forthcoming). Process overlap theory: a unified account of human intelligence. *Behavioral and Brain Sciences*.

Kray, L., Galinsky, A., & Thompson, L. (2002). Reversing the gender gap in negotiations: an exploration of stereotype regeneration. *Organizational Behavior and Human Decision Processes*, 87, 386-409.

Kray, L., Thompson, L., & Galinsky, A. (2001). Battle of the sexes: gender stereotype confirmation and reactance in negotiations. *Journal of Personality and Social Psychology*, 80, 942-958.

Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin*, *89*, 308-324.

Mameli, M. (2001). Mindreading, mindshaping, and evolution. *Biology and Philosophy*, 16: 591-628.

McGeer, V. (2008). Trust, hope, and empowerment. *Australasian Journal of Philosophy*, 86(2), 237-254.

Menary, R. (2007). *Cognitive Integration: Mind and Cognition Unbounded*. New York: Palgrave Macmillan.

Milgram, S. (1974). *Obedience to Authority*. New York: Harper Collins.

Morton, A. (2013). *Emotion and Imagination*. London: Polity.

Nguyen, H., & Ryan, A. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, 93(6), 1314-1334.

Oberauer, K., Schulze, R., Wilhelm, O., & Süss, H. (2005). Working memory and intelligence -their correlation and their relation: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, 131(1), 61-65.

Palermos, O. (2014). Loops, Constitution, and Cognitive Extension. *Cognitive Systems Research*, 27, 25-41.

Pettit, P. (1995). The cunning of trust. *Philosophy and Public Affairs*, 24(3), 202-225.

Rauthmann, J., Gallardo-Pujol, D., Guillaume, E., Todd, E., Nave, C., Sherman, R., Ziegler, M., Jones, A., & Funder, D. (2014). The situational eight DIAMONDS: A taxonomy of major dimensions of situational characteristics. *Journal of Personality and Social Psychology*, 107(4): 677-718.

Richard, D., Bond Jr., C., & Stokes-Zoota, J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, 7(4), 331-363.

Rigdon, M., Ishii, K., Watabe, M., & Kitayama, S. (2009). Minimal social cues in the dictator game. *Journal of Economic Psychology*, *30(3)*, 358-367.

Roberts, R. (2013). *Emotions in the Moral Life*. Cambridge University Press.

Rupert, R. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy*, 101, 389-428.

Schmader, T. & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, 85(3), 440-452.

Schmader, T., Johns, M., & Barquissau, M. (2004). The costs of accepting gender differences: the role of stereotype endorsement in women’s experience in the math domain. *Sex Roles*, 50, 835-850.

Shepard, R. (1984). Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review*, 91: 417-47.

Sprevak, M. (2010). Inference to the hypothesis of extended cognition. *Studies in History and Philosophy of Science*, 41(4), 353-362.

Srivastava, S., McGonical, K., Richards, J., Butler, E., & Gross, J. (2006). Optimism in close relationships: How seeing things in a positive light makes them so. *Personality Processes and Individual Differences*, 91(1), 143-53.

Sterelny, K. (2010). Minds: extended or scaffolded? [*Phenomenology and the Cognitive Sciences*](http://philpapers.org/asearch.pl?pub=768) 9 (4), 465-481.

Stone, J., Lynch, C., Sjomeling, M., & Darley, J. (1999). Stereotype threat effects on black and white athletic performance. *Journal of Personality and Social Psychology*, 77, 1213-1227.

Strawson, P. (1974). *Freedom and Resentment and Other Essays*. London: Methuen.

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

Sutton, J., Harris, C., Keil, P., & Barnier, A.. (2010). The psychology of memory, extended cognition, and socially distributed remembering. *Phenomenology and the Cognitive Sciences*, 9 (4), 521-560.

Taylor, V., & Walton, G. (2011). Stereotype threat undermines academic learning. *Personality and Social Psychology Bulletin*, 37(8), 1055-1067.

Wong, D. (2006). *Natural Moralities*. Oxford University Press.

Yeung, N., & von Hippel, C. (2008). Stereotype threat increases the likelihood that female drivers in a simulator run over jaywalkers. *Accident Analysis & Prevention*, 40, 667-674.

Zawidzki, T. (2013). *Mindshaping*. MIT Press.

1. See Alfano (2013a), esp. pp. 111-139, for a more fine-grained discussion of the situationist challenges to responsibilist and reliabilist brands of virtue epistemology. [↑](#footnote-ref-1)
2. For further discussion of such influences, see Alfano (2014a). [↑](#footnote-ref-2)
3. We pause to note here that we’re not talking only about behavior: behavior is influenced in large part *because* thought and feeling are also influenced. This should settle once and for all the charge – often unfairly leveled against philosophical situationists and their sympathizers – of being “behaviorists” in some pejorative sense. [↑](#footnote-ref-3)
4. Our distinction between embedding and extension of character mirrors the debate in the extended mind literature between the hypothesis of embedded cognition (e.g., Rupert 2004) and the hypothesis of extended cognition (e.g., Sprevak 2010). [↑](#footnote-ref-4)
5. And it certainly seems to be so. It is probably the most-studied individual difference in scientific psychology, with a history of over a century and robust predictive and explanatory power (Kovacs & Conway forthcoming). [↑](#footnote-ref-5)
6. Though, if Kovacs & Conway (forthcoming) are on the right track, the ultimate psychological explanation will always appeal in turn to more specific cognitive and neural mechanisms that tend to have a high degree of overlap and thus to explain the “positive manifold” that is so well documented in intelligence research. [↑](#footnote-ref-6)
7. It is easy to underestimate the effect of stereotype threat here in at least two ways: first, the SAT scores meant to control for differences in skill level were presumably acquired under the same kinds of threatening conditions that lead to a decrease in performance. Second, even when an exam is presented as non-diagnostic of intellectual ability, it is likely that a student who has taken an SAT or ACT exam would recognize GRE questions as intellectually diagnostic. More generally, it is worth noting that in a meta-analysis of stereotype threat effects conducted by Nguyen and Ryan (2008), the overall mean effect size was |.26|, and in some cases as high as |.64| To put this into perspective, |.10| is considered to be a small effect size relative to most social psychological effects, |.20| medium, and |.30| large. (Richard, Bond Jr., & Stokes-Zoota 2003, p. 339). [↑](#footnote-ref-7)
8. One possible explanation, offered by Schmader & Johns (2003) is that a decrease in performance is mediated by a decrease in working memory capacity. Given that working memory capacity is a limited resource that is highly correlated with fluid intelligence (meta-analyses by Kane, Hambrick, & Conway (2005) and Oberauer, Schulze, Wilhelm, & Süss (2005) estimate the correlation at *r* = .72 and *r* = .85, respectively), it would be unsurprising that the more of it that is allocated to worrying about one’s group identity and one’s performance, the less would be available for the processing demands of intellectually challenging tasks. [↑](#footnote-ref-8)
9. In academic contexts, the effect extends beyond racial minorities to women – especially in STEM fields (Schmader, Johns, & Barquissau 2004), and also to low SES individuals (Croizet & Claire 1998). Stereotype threat is also experienced outside of academic contexts, such as in negotiations (Kray, Thompson & Galinsky 2001, 2002), athletics (Stone, Lynch, Sjomeling, & Darley, 1999), and driving tests (Yeung & von Hippel 2008). [↑](#footnote-ref-9)
10. Much of this section is informed by and expands on Alfano (2015). Beyond friendship, examples of potentially more complicated dyadic extension in this vein might include romantic partnerships and domestic abuse. [↑](#footnote-ref-10)
11. Thanks to J. Adam Carter and Andy Clark for emphasizing the importance of updating internal models. [↑](#footnote-ref-11)
12. Presumably, this is one of the reasons why people may prefer to have anger directed at them rather than being treated as, in Strawson’s (1974) words, someone “to be managed.” [↑](#footnote-ref-12)
13. The existence and impact of such ongoing feedback loops has been empirically investigated in the context of romantic partnerships (Srivastava et al. 2006; Assad et al. 2007). Further work should examine similar effects in other close relationships. [↑](#footnote-ref-13)
14. Our treatment is by no means exhaustive of the possible relations with the social environment. Given limited space, we have not considered, for example, Sterelny’s (2010) notion of scaffolding, Menary’s (2007) notion of integration, nor the notion of distribution (e.g., Sutton et al. 2010). [↑](#footnote-ref-14)