The Embedded and Extended Character Hypotheses

Mark Alfano & J. August Skorburg

Draft: Please do not quote or cite without permission

**Introduction**

This paper 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. This is a natural but underexplored next step in the extended mind research program. Virtues and vices are dispositions to token a suite of occurrent mental states and engage in a signature of behaviors in response to configurations of external and internal variables. If those mental states are 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 recounting the histories of the extended mind debate and the situationist challenge, with an eye to their integration. Next, following Orestis Palermos’s (2014) attempt to recruit dynamic systems theory to help settle the cognitive extension debate, we lay out conditions for the sort of integration that will count as character embedding and extension. Third, we argue that an epistemic example (stereotype threat) meets the criteria for embedding, while a moral example (friendship) satisfies the more exacting conditions for extension.

**1 Historical background**

This section recounts the relevant histories of the extended mind debate and the situationist challenge to virtue theory.

*1.1 The extended mind debate*

We here follow one thread of the extended mind debate in order to point to the need to identify clear criteria for cognitive embedding and extension. While there are historical antecedents[[1]](#footnote-1), the hypothesis of the extended mind was inaugurated by Clark and Chalmers (1998). This paper continues to generate responses, defenses, and epicycles, but we limit our focus here to a single claim:

[In cases of cognitive extension], the human organism is linked with an external entity in a two-way interaction, creating a *coupled system* that can be seen as a cognitive system in its own right. All the components in the system play an active causal role, and they jointly govern behavior in the same sort of way that cognition usually does [….] Our thesis is that this sort of coupled process counts equally well as a cognitive process, whether or not it is wholly in the head (pp. 8-9).

A persistent thorn in the side of extended mind theorists has been their supposed inability to provide a principled criterion for what counts as the right kind of “two-way interaction” or “coupled system.” This criticism is best expressed in Adams and Aizawa’s (2001) formulation of the so-called *coupling-constitution fallacy*:

The fallacious pattern is to draw attention to cases, real or imagined, in which some object or process is coupled in some fashion to some cognitive agent. From this, one slides to the conclusion that the object or process constitutes part of the agent’s cognitive apparatus or cognitive processing […. C]oupling relations are distinct from constitutive relations, and the fact that object or process *X* is coupled to object or process *Y* does not entail that *X* is part of *Y* (p. 68).

We acknowledge the force of this criticism. One main desideratum of a theory of cognitive extension is a principled criterion that excludes implausible cases of extension (e.g., “because I have reliable access to the Internet, the entire Internet is part of my mind”) while still including at least some paradigm cases of extended cognition (e.g., the practices of air traffic controllers described in Noë (2009)). Without a clear criterion for extension, this debate becomes mired in a swamp of contrary intuitions, with advocates pointing to cases that seem to them to involve extension and skeptics questioning whether, by the advocates’ own lights, all interactions between a cognitive agent and her environment amount to extension. This is the well-known problem of “cognitive bloat” (Rowlands 2010, pp. 93-95).

One of the most plausible strategies for dealing with this general worry about inferring constitution from causation is offered by Rupert (2004), who contends that the framework of embedding can account for the same range of cognitive phenomena as the framework of extension, and can do so without having to make the dubious inference from causation to constitution. To say cognition is embedded is to say that “cognitive processes depend very heavily, in hitherto unexpected ways, on organismically external props and devices and on the structure of the environment in which cognition takes place” (p. 393). Importantly, the embedding framework denies that these props, devices, and structures are themselves mental. Put another way, a proponent of embedding is not committed to the claim that the vehicles of cognition are external; she is only committed to the claim that an explanation of cognitive processes will not be an entirely internalist one (Sprevak 2010). If embedding can account for the same range of cases as extension, Rupert plausibly argues, then *ceteris paribus* we should endorse the former over the latter by “the methodological principle of conservatism” (Rupert, p. 395). While we need not take a stand here on the explanatory power of these frameworks in the cognitive sciences, or the value of methodological conservatism, we do take Rupert’s criticism to illustrate a crucial distinction between embedding and extension. For our purposes, extension is committed to the stronger claim that, sometimes, cognition (or character) is *constituted* *by* structures and processes beyond the confines of the agent’s skin, while embedding is committed to the weaker claim that, sometimes, cognition (or character) *merely depends on* structures and processes beyond the confines of the agent’s skin.

Regardless of where one’s allegiances may lie in this debate, it is clear that if this line of inquiry into the ontology of mind is to be a “progressive research program” rather than “the swapping of hunches about what exists” (Yablo 1998, p. 229), what’s needed is a relatively clear criterion for distinguishing cases of extension from cases of embedding, and both from cases of mere causal influence. We believe that Palermos’s *ongoing mutual interdependence on the basis of feedback loops criterion* (henceforth just ‘feedback loops criterion’) is able to do just this. We explain this criterion in more detail below; for now, here’s a thumbnail sketch: two systems, S1 and S2, count as suitably coupled when states of S1 systematically function as parameters of S2, and states of S2 systematically function as parameters of S1. We believe that if, as Palermos claims, the feedback loops criterion provides an objective way of “drawing systemic boundaries,” (p. 34) then, *mutatis mutandis*, the same strategy will support arguments for the extension of character beyond the bounds of skull and skin.

*1.2 The situationist challenge to virtue theory*

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 theory began when Harman (1999) and Doris (2002) argued that the dominant neo-Aristotelian conception of moral virtue was empirically inadequate. Neo-Aristotelians 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.

Alfano (2012, 2013a, 2014c) took the challenge a step further, arguing that virtue epistemology is also empirically inadequate. But Alfano (2014a) also suggested a way out: 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 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).

Despite its appeal, Alfano’s model of extended character is highly speculative, under-informed by the extended mind debate, and subject to the familiar charge of committing the coupling-constitution fallacy. Under what conditions is it the case not just that an agent’s social and/or asocial 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? This is where the feedback loops criterion takes center stage: when an agent is functionally integrated through ongoing feedback loops with her social and/or asocial environment, the environment doesn’t just causally influence her but becomes part of her character, for good or ill (Alfano 2015). What her friends expect 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 friends’ expectations, which are again transmitted to her, further shaping her thought, feeling, deliberation, and behavior, which again influences her friends’ expectations, and so on. In the next section, we explain how dynamical systems theory models such interactions.

**2 The dynamical systems theory approach to extension**

Orestis Palermos (2014) is responsible for the application of dynamical systems theory to cognitive extension, so in this section we follow him.[[2]](#footnote-2) Briefly, a dynamical system S is an ordered pair ({x1, x2, …, xn}, L(*x, u*)), the first element of which is the set of possible states of the system and the second element of which is the dynamical law of the system. The law is typically given by a set of differential equations that take states and parameters as arguments and these govern the transitions between states. Depending on the initial state and the settings of the parameters, L determines the sequence of subsequent states. Some systems demonstrate a tendency to enter and then never leave a constrained set of states, known as attractors; each attractor is surrounded by its basin of attraction (e.g., a watershed, in which all water that flows at all ends up in the same place, or a large star’s gravitational field, which pulls into its orbit everything that gets close enough, but not too close, to be annihilated). Some systems demonstrate a tendency to avoid or at least not remain in a constrained set of states, known as repellors (e.g., the top of a hill, from which objects are likely to roll or slide).

The most important feature of dynamical systems for our purposes is the set of parameters, which can be either static (in an autonomous system) or dynamic (in a non-autonomous system). In a non-autonomous system, the parameters can change so much that new attractors appear, basins of attraction expand and contract, old attractors evaporate, new repellors appear, and old repellors evaporate. This alters the qualitative and quantitative features of the system. How should we conceive of such changing parameters? One convenient way to do so is to treat some of the parameters of the non-autonomous system, S1, as a function of the state variables of a second system, S2 (Beer 1995, p. 181). As S2 evolves, some of its states are or determine some of the parameters of S1.

Consider the following quip from Adams and Aizawa (2010): “Question: Why did the pencil think that 2 + 2 = 4? Clark’s answer: Because it was coupled to the mathematician” (pg. 67). This is, of course, a caricature of Clark’s position, but it does provide an opportunity to emphasize the difference between mere state overlap and parametric coupling. The state of the pencil is part of the more encompassing state of the pencil-and-mathematician. From the point of view of dynamical systems theory, this is uninteresting. The parameters governing the pencil’s trajectory are static laws of physics, uninfluenced by the states of the mathematician. Likewise, the parameters governing the trajectory of the mathematician’s cognition are (as far as this example is concerned, at least) static, uninfluenced by the states of the pencil. Sure, there are causal interactions between the mathematician and the pencil, and the states of each are proper parts of the states of the combined mathematician-and-pencil system, but, as an adolescent at the junior prom would tell you, juxtaposition is not sufficient for coupling.

A pair of juxtaposed but autonomous systems may casually interact with each other in an unsystematic way, but the states of one never or rarely influence the parameters of the other. The example of the mathematician and the pencil seems to satisfy this description. By contrast, a pair of non-autonomous systems doesn’t just have stray interactions; one systematically influences the parameters of the other. This kind of dependence can be asymmetric, such that S2 goes its merry way regardless of what happens in S1, or symmetric, such that S2 is dependent on S1 in roughly the same way that S1 is dependent on S2: some of the parameters of S2 are functions of the states of S1. An example of asymmetric dependence is a plant’s growth over time (S1) and the relative trajectories of the Sun and Earth (S2). The speed of the plant’s growth at time *t* is a function of, among other things, ambient heat and light at *t*. These parameters (time of day, time of year, etc.) are a function of the trajectories of the Sun and the Earth, but the trajectories of the Sun and Earth are, for all practical purposes, uninfluenced by the plant’s growth.

It’s crucial to construe the environment’s effects on the agent and the agent’s effects on the environment not as one-off impacts, but as ongoing pressures, interventions, and adjustments. The temperature of the ambient air doesn’t affect you once and then stop; your fiddling with the thermostat is nearly interminable. We here quote Beer (1995, p. 182) at some length:

I cannot overemphasize the fundamental role that feedback plays in this relationship. Any action that an agent takes affects its environment in some way through [pathway Y], which in turn affects the agent itself through the feedback it receives from its environment via [pathway X]. Likewise, the environment’s effects on an agent through [X] are fed back through [Y] to in turn affect the environment itself. Thus, each of these two dynamical systems is continuously deforming the flow of the other […] and therefore influencing its subsequent trajectory.

Furthermore, within symmetric systems, there is an important qualitative difference between a pair, S1 and S2, where most or all of S1’s parameters are functions of S2’s states while almost none of S2’s parameters are functions of S1’s states, and a pair, T1 and T2, where the parameter-setting is roughly balanced. In artificial intelligence and other fields, S1 has been modeled as an agent and S2 as the agent’s environment. Your environment clearly has the relevant kind of influence on you. Just for instance, like the growing plant, what you can and will do depends on ambient temperature, of which your body temperature is a function, and light, of which your sensory inputs are a function. It’s less clear whether or to what extent you ever have this kind of influence on your environment.[[3]](#footnote-3) In the ordinary course of things, the environmental system is much too large and complex for the agent to have a non-trivial effect on it. There may be small local effects, but these are dwarfed both by the sheer size of the environment and by the much more substantial influence the environment has on the agent.

This imbalance, even within symmetric non-autonomous systems, is illustrated by figure 1. In this figure, E represents the environment, A the agent, pathway X E’s influence on A’s parameters, and pathway Y A’s influence on E’s parameters. Completely asymmetric interactions can be seen as degenerate cases in which Y is trivial or non-existent. Think, for instance, of the gravitational pull of the Sun and Moon on terrestrial oceanic water (X) and terrestrial oceanic water’s gravitational pull on the Sun and Moon (Y). Naturally, these forces are equal, but since the masses of the Sun and Moon are astronomically greater (pun intended) than the mass of the Earth’s oceanic water, X is enough to cause the tides (and hence set some parameters of any decent model of sea level) whereas Y is so trivial that only in very strange circumstances would an astrophysicist attend to it.

Figure 1: An agent and its environment as coupled dynamic systems, adapted from Beer (1995, p. 182)

To move away from astronomical examples and back to the sorts of examples that interest us, consider the coupled system of Twitter (E) and Justin Bieber (A). At the time this paper was drafted, Bieber had over 52 million followers – roughly 7% of all users. What people are tweeting about is to some extent influenced by what Justin Bieber tweets about (*much* more than what some nobody with a dozen followers tweets); what Justin Bieber tweets about is to some extent influenced by what people are tweeting about. This is a symmetric non-autonomous system, but an imbalanced one. Even though the Y pathway is non-trivial, the X pathway is much more influential. When parametric coupling goes only or almost only from one system to another system, the appropriate framework is probably not extension but embedding. We return to this point below.

Symmetric coupled systems in which one system is an agent and the other is the rest of the environment will typically embody this imbalance. That doesn’t mean, though, that there are no cases of balanced coupling. Instead of thinking of the environment as monolithic, we can carve out another agent and then think about the resulting system in terms of three pairwise couplings, as illustrated in figure 2.

Figure 2: A pair of agents and their environment as coupled dynamic systems

In this framework, there are two agents, both of whom are positioned in the same environment. As before, the influence of the environment on the agents (pathways X1 and X2) is much greater than the influence of the agents on the environment (pathways Y1 and Y2). The influence of the agents on each other (pathways Z1 and Z2), however, at least can be roughly balanced. Moreover, it can be substantial indeed when those agents are human animals, given how hypersocial we are. People often try to live up to their family’s, friends’, and acquaintances’ expectations, even without realizing it (Zawidzki 2013). They are attuned to others’ reactive attitudes. They consider prospectively whether others would approve or disapprove of some course of action, which informs their deliberation and behavior. They consult with others explicitly, implicitly, and imaginatively. They revise their beliefs and values in light of others’ feedback. We return to this idea below to argue that friendship is a paradigm case of character extension. In the balance of this section, we illustrate the idea of balanced symmetric coupling with a biological example.

Consider the isopod parasite *Ceratothoa italica*. This parasite usually enters the gill cavity of the host fish and settles on its tongue. Once settled, the parasite degenerates most of the host fish’s tongue, attaching itself to the small remaining tongue stub with little hook-like pereopods. “In this position the isopod superficially resembles the tongue of its host fish” and as Brusca and Gilligan (1983) hypothesized, “these isopods serve as a mechanical replacement for the fish’s tongue and represent the first known case (in animals) of functional replacement of a host structure by a parasite” (pp. 813-14). Importantly, in fish, the tongue plays little or no respiratory role and is “primarily a mechanical device to hold prey in the mouth” (p. 815). All of this leads Brusca and Gilligan to claim that “a fish with an ‘isopod tongue,’ while perhaps not feeding as efficiently as a non-parasitized fish, feeds more efficiently than a fish with no tongue at all and no isopods in its place” (p. 815).

We believe that the host fish (A1) and the *Ceratothoa italica* (A2) exemplify a roughly balanced symmetrically coupled dynamic system (A1-&-A2) embedded in the Mediterranean Sea, an environment (E) on which they have little to no parametric influence. Some of the parameters of the fish (e.g., the position and function of its “tongue”) are functions of the states (width, height, length, strength of attachment to the fish’s tongue stub) of the parasite. Because the parasite effectively supplants the host’s tongue, and because the tongue is crucial to the processing of food, the parasite quite literally determines what and how the host eats. It is alsoimpossible to predict, explain, or control the trajectory of the *Ceratothoa italica* without taking into account the moment-to-moment evolution of the states of the host fish. In the language of dynamical systems theory, some of the parameters of the parasite (e.g., its nutritive inputs) are functions of the states of the fish (e.g., its circulatory system). Moreover, the integration is close to balanced, with both the fish and the parasite exerting large and important influences on each other. As we pointed out above, this mutual dependence is best construed not as a one-off impact of each system on the other, but in terms of ongoing pressures, interventions, and adjustments, where feedback loops are so important that the functioning of neither system can be understood without reference to the other.

In this section, we introduced dynamical systems theory in thumbnail sketch. We then used dynamical systems theory to distinguish the degrees to which systems can be pairwise dependent on one another. The least interesting kind of dependence is mere causal interaction, in which the states of each system affect to some degree the states of the other. If arguments for extended cognition or extended character depended solely on such interactions, they would clearly fail. Next, some of the parameters can be functions of some of the states of another system. This kind of dependence can be either asymmetric, as in the tidal example, or symmetric. We contend that examples of asymmetric dependence at best support arguments for embedded cognition and character, not extended cognition and character. Finally, among systems that are symmetrically dependent, there is an important distinction between imbalanced and balanced coupling. Imbalanced coupling is exemplified by Twitter and Justin Bieber, whereas balanced coupling is exemplified by *Ceratothoa italica* and its host fish. This taxonomy is illustrated in figure 3.

Figure 3: Taxonomy of systems in dynamical systems theory

Imbalanced but symmetric coupling seems to us, like asymmetric coupling, to support arguments only for embedded cognition and character. By contrast, we think that balanced symmetric coupling is not just the best support for the idea of extension but perhaps the only support. It seems to us plausible to say that *Ceratothoa italica*, once it does its dirty work, *is* the host fish’s tongue. We are not convinced that anything shy of such balanced coupling is up to the job.[[4]](#footnote-4)

This sets the bar quite high, since it’s unclear whether most of the orthodox examples of extension involve balanced symmetric coupling. In the next section, we use stereotype threat to illustrate embedding and friendship to illustrate extension. If these examples are as convincing as we think they are, we will have shown how to translate the embedded and extended mind hypotheses into the language of intellectual and moral character.

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

In the previous sections, we introduced relevant distinctions from the extended mind debate and relevant terminology from dynamical systems theory in order to set up the central claims of this paper: long-lasting, wide-ranging, and normatively-evaluable *agentic dispositions* are sometimes best understood as embedded or even as extended. 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.

*3.1 Stereotype threat and embedded character*

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).[[5]](#footnote-5)

All of this suggests that, in the case of negatively stereotyped racial minorities, the more one worries about individual performance being viewed as indicative of the negatively stereotyped racial group to which one belongs, the worse one’s individual performance on a relevant task is likely to be. While the precise mechanisms responsible for this performance decrease are not known,[[6]](#footnote-6) there is little doubt as to the reality and efficacy of stereotype threat for racial minorities in academic contexts.[[7]](#footnote-7)

Alfano (2014b) argues on intuitive grounds that stereotype threat is an example of cognitive extension. With our more nuanced taxonomy, we can now see that it is better understood as an example of embedding. Recall that, for our purposes, the interesting kinds of coupled systems are *non-autonomous* (some of the parameters of S1 are functions of some of the states of S2) where the interaction between S1 and S2 is ongoing. We argued above that making a case for extension will also require that the non-autonomous coupled systems be *symmetric* (some of the parameters of S1 are functions of some of the states of S2 in roughly the same way as some of the parameters of S2 are functions of some of the states of S1) and *balanced* (S1 exerts roughly the same degree of influence on S2 as S2 exerts on S1). We also noted that cases that do not meet both the symmetry and the balance conditions are probably best understood in the framework of embedding. In such cases, the agent may be said to be heavily dependent on, or intimately coupled with the relevant environmental features, but the further claim that these relevant environmental features thus form a proper part of the agent would be unlicensed. In the remainder of this section, we explore how stereotype threat provides a way to understand intellectual character as embedded in this way.

We can represent the anecdote from the beginning of Section 3.1 and the findings from Steele and Aronson with reference to Figure 1. In the figure, where E and A form a coupled system, E represents the social and cultural environment in which the stereotype that “Blacks are poor students” exists and A represents a black test-taker. The X pathway represents all socio-cultural influences of the stereotype on A, and the Y pathway represents the influence of A on the environment.

The evolution of E over time is governed by a dynamical law, as outlined in Section 2. One state variable that characterizes E is the level of threat which is present. The level of threat is a function of, among other things, media portrayal, prior prejudices, implicit biases, evaluative judgments, slurs, and of course, larger economic, political, and legal structures.

Similarly, the evolution of A over time is governed by a dynamical law, and one state variable that characterizes A is a disposition to perform in characteristic ways on exams. This disposition is a function of, among other things, intelligence, previous education, health, fatigue, difficulty of exam, and distractions, such as the perceived level of threat. The perceived level of threat is a state of A which is characterized by A’s awareness of relevant negative stereotypes.

In cases of stereotype threat, E and A can be understood as forming a non-autonomous, symmetric, but imbalanced coupled system. We will consider each of these designations in turn.

First, E and A are non-autonomous because some of the parameters of A (a given characteristic performance on an exam) are functions of some of the states of E (the level of threat). The level of threat sets the parameters for performance on the exam: the higher the level of threat, the lower the exam score is likely to be; the lower the threat, the higher the exam score is likely to be.

The coupling between E and A is symmetric because some of the parameters of E are determined by some of the states of A in roughly the same way as some of the parameters of A are determined by some of the states of E. That is, exam performance also sets the parameters for the level of threat: the lower the exam scores, the higher the threat then becomes; the higher the exam scores, the lower the threat then becomes.[[8]](#footnote-8)

Lastly, what is made clear in Figure 1, and what is important in the context of the embedded/extended distinction, is that the X and Y pathways are imbalanced. A single individual’s performance, good or bad, will not significantly change the stereotypes about the group of which he is a member, in much the same way that tweets to our two dozen followers will not significantly influence what is trending on Twitter. This, however, does not entail that every situation involving stereotype threat is necessarily imbalanced. Imagine a small school with a classroom of 15 students, 12 white and 3 black, where the stereotype that blacks are poor students is prevalent. If 3 black students in the class, despite threatening conditions, regularly kept pace with or outperformed their peers, this should, in theory, help to balance out the X and Y pathways.[[9]](#footnote-9)

With this sketch of stereotype threat as modeled by dynamical systems theory on the table, think back to the anecdote at the beginning of Section 3.1. The environment the test-taker inherits is one in which there are biases, prejudices, judgments, and stereotypes about the intellectual capabilities of his racial group. The sum total of these features is what we have called the level of threat present in his environment. These features of the environment come to our test taker through the mass media, political institutions, chatter from his peers, and the like - what we have called the X pathway. The sum total of the worries, anxieties, and distractions he experiences from the X pathway comprises 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. So 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 – through what we have called the Y pathway – 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 accrues, perhaps in the context of a single course, but also over the course of a college career, it contributes to a higher level of threat in the environment. And when the level of threat in the environment is high, the perceived level of threat that comes to our test taker is also heightened. So when he faces a similar task, he becomes all the more likely to underperform as a result of the threatening conditions. Given the importance of feedback loops in this kind of coupled system, and given the ubiquity and importance of standardized tests, it is no wonder stereotype threat is such a prevalent and pernicious phenomenon.

To return to the language of dynamical systems theory, these feedback loops which lead to underperformance on exams over time can be understood in terms of attractors and attractor basins. Recall that some systems demonstrate a tendency to enter and then never leave a constrained set of states. This space is known as an attractor and the constrained set of states surrounding it is known as its basin of attraction. When we consider the test taker and the environment as a coupled system, this system can enter into a number of possible states with respect to the level of threat in the environment and the corresponding performance on an exam. One such state might be characterized by a high level of threat (and perceived level of threat) present when test takers are required to fill out demographic information before taking an intellectually diagnostic exam. In this state, we know that test takers will routinely underperform. Thus, the demographic information requirement and corresponding drop in exam performance can be understood as an attractor, because once the system enters this state, it will tend not to leave closely related, constrained sets of states in the future. That is, over time (be it a single class or a college career), when the system enters a state where a sufficient level of threat is present, it will tend not to leave the state of corresponding underperformance on exams.

What we hope to have shown here is that stereotype threat can be modeled using the tools of dynamical systems theory. Within this framework, stereotype threat fits into our taxonomy of systems (Figure 3) as a non-autonomous, symmetric, but imbalanced coupled system. Given the imbalanced nature of the coupling between a test taker and his environment, we believe that stereotype threat thus meets the condition for embedding, but falls short of extension.

*3.2 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 symmetric but imbalanced coupled system. 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) As before, we will present the case intuitively at first, then spell it out in more detail using the language of dynamical systems theory.

Imagine two agents, Ashley (A1) and Azim (A2). 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 actually 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. 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 the 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.[[11]](#footnote-11)

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.

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).

If these remarks on friendship are on the right track, they show how friendship can be modeled as a balanced symmetric coupled system, in which some of Ashley’s parameters are functions of Azim’s states and some of Azim’s states are functions of Ashley’s parameters. 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)

In the language of dynamical systems theory, A1’s parameters are partly a function of A2’s states. 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.

Because there are multiple feedback contingencies for both of them, their dispositions become modally robust. Or, in the language of dynamical systems theory, 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.[[12]](#footnote-12)

**4 Programmatic conclusion**

The foregoing 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**

Abraham, R., & Shaw, C. (1992). *Dynamics – The Geometry of Behavior*, 2nd edition. Redwood City, CA: Addison-Wesley.

Adams, F., & Aizawa, K. (2001). Defending the bounds of cognition. In Menary (ed.), *The Extended Mind*. 67-80, MIT Press.

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. (2015). Ramsifying virtue theory. In M. Alfano (ed.), *Current Controversies in Virtue Theory*. New York: Routledge.

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.

Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72, 173-215.

Brusca, R. & Gilligan, M. (1983). Tongue replacement in a marine fish (Lutjanus guttatus) by a parasite isopod (Crustacea: isopoda). *Copeia*, 3, 813-816.

Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis* 58(1), 7-19.

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.

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

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

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.

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.

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

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.

Noë, A. (2009). *Out of Our Heads: Why You Are Not Your Brain, and Other Lessons from the Biology of Consciousness*. New York: Hill and Wang.

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.

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.

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

Rowlands, M. (2010). *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. MIT 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.

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.

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.

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.

Yablo, S. (1998). Does ontology rest on a mistake? *Proceedings of the Aristotelian Society*, 72, 229-261.

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. Broadly, we view the relevant trajectory of externalism as beginning with the externalization of semantic content (Putnam and Burge), running through epistemic justification (Dretske), to mental states (Clark and Chalmers), up to our present position of extended dispositions or character. [↑](#footnote-ref-1)
2. For a short, readable summary of dynamic systems theory, see Beer (1995). For an even more extensive treatment, see Abraham & Shaw (1992). [↑](#footnote-ref-2)
3. Fundamentally, everything in your future light cone is potentially influenced by you. With vanishingly few exceptions, this influence is so trivial that modeling it would be pointless. [↑](#footnote-ref-3)
4. Since the balanced/imbalanced distinction is continuous rather than categorical, the line between balanced and unbalanced is vague, and there will inevitably be borderline cases. We are convinced, though, that there are clear cases on either side of the distinction. [↑](#footnote-ref-4)
5. 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), pg. 339). Given that less than 25% of all effects studied by social psychologists produce a mean effect size of |.30|, this would indicate that the effects of stereotype threat are some of the most powerful in all of social psychology. [↑](#footnote-ref-5)
6. 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-6)
7. In academic contexts, the effect extends beyond racial minorities to women – especially in the 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-7)
8. We believe that Kovacs and Conway’s (forthcoming) process overlap model, in conjunction with the findings of Schmader and Johns (2003), provides a helpful framework for more precisely understanding the neural and cognitive processes which are affected in cases of stereotype threat, and hence, play an important role in the determination of exam scores. [↑](#footnote-ref-8)
9. We say “in theory” here, being well aware that the propagation of negative academic stereotypes may have less to do with analyses of exam scores, and more to do with things like implicit biases (and this is to say nothing of obvious, explicit biases). Indeed, we believe that a dynamical systems approach is particularly well-suited to track these different dimensions of stereotype threat. [↑](#footnote-ref-9)
10. Much of this section is informed by and expands on Alfano (forthcoming a). [↑](#footnote-ref-10)
11. 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-11)
12. 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 friendships, parent-child relationships, sibling relationships, and other close bonds. [↑](#footnote-ref-12)