Hello, everyone, and thank you for tuning in.

I would like to start by saying the opinions expressed below are purely my own. Josh Knobe and George Newman may or may not be on board with anything I'm about to say, though I do try to give a fair representation of what we, as a team, argue for in our paper.

One issue that seems to have touched a nerve is the cross-cultural generalizability of the true self. This is no great shock. Here we are, after all, at the International Cognition & Culture Institute. Lawrence Hirschfeld, Radu Umbres, and Victoria Fomina all raise points relating to this problem (some more delicately than others).

To be sure, cross-cultural universality of the true self has not been demonstrated. Nor, I hasten to add, has it been claimed. Trying to prove that a cognitive capacity is a human universal is a lot like trying to prove all swans are white. There will always be some Pirahã black swan lurking out there, threatening to bring the whole thing tumbling down. I have no interest in defending views so bold and so profoundly vulnerable.

What we do claim is a cross-cultural *robustness*, since the true self appears across multiple different cultures.

Now, we can argue about how robust is robust, how canalized is canalized. And to a certain extent this question awaits the patient gathering of further data. But the fact that we observe the true self even in cultures with radically different notions of selfhood from our own gives us some sense of this robustness. The true self survives at least some cultural vicissitudes that the self does not.

Hirschfeld writes: "Selves, true or otherwise, are ways of imagining a wealth of options that simply aren't available to most people." He seems quite confident, but where is his evidence? The available data show that the true self concept holds across multiple cultural groups (De Freitas et al., in press-b). Certainly it seems premature to assert otherwise.

Umbres points out that college-educated populations—on which the De Freitas et al. data relies—may differ in important ways from rural populations. It is difficult to argue with this, so I will not. But it bears mentioning that research on the self concept across Eastern and Western cultures relies on college educated populations in both. These populations absolutely can show cross-cultural differences in conceptions of the self; yet they fail to do so for the true self.

*

Gloria Origgi levels the rather heartbreaking accusation that we ignore "literary, philosophical, and artistic sources of evidence." To some extent I disagree; our paper engages with the philosophical literature, and we do mention some examples from the arts. (If Grease doesn't represent the best of American culture I don't know what does.)

Many of the commentators point to qualitative, cultural, or anecdotal experiences as potential counterpoints to arguments in the paper. I think these sources of evidence are fertile sources for idea generation. They are particularly useful breaking out of the tunnel vision Origgi worries about. But as evidence, they stand rather spindly and unsteadily on their own. They require great girders of support, in the form of empirical data. It would be wonderful to follow these leads, these hints and allegations, quantitatively.

*

Ophelia Deroy and Gloria Origgi both advance alternative explanations for the true self. Origgi

suggests that “the true self is at the core of our *social* self-identity, our public self-image; that it is tailored to defend our reputation – we may look so-so, but deep down, we are so good! – and to contribute to how we would like to see ourselves seen by others.” Deroy wonders if the true self isn’t simply a manifestation of the illusory superiority effect: “we tend to be over-optimistic about our own abilities, and diffuse bad evidence when it contradicts this belief. The beliefs about the ‘true self’ resemble a general optimism about human morality.”

The problem with these proposals is the selective positivity of the true self also applies when making evaluations of other people. So the most parsimonious explanation is unlikely one that has to do with social signaling of the self. The positive valence bias in other-person judgments is arguably the most novel and powerful aspect of this emerging literature. It shows that this is not just some bastardized version of the self-serving bias. It seems to reveal a more general cognitive mechanism.

*

Fomina and Umbres each challenge the assertion that the true self is always good.

Umbres identifies the denigration of outgroups (e.g. the Roma and the Jews) as a counterexample. De Freitas and Cikara have unpublished data, which we cite in our paper, showing that encouraging people to focus on the true self of outgroup members decreases negative attitudes towards them, reducing intergroup conflict. This finding suggests that, if pressed, racists may concede that their bogeymen are nonetheless good deep down. (As a nameless politician once said: “Some, I assume, are good people.”)

In the entirely plausible event that some groups are so loathed that they really are seen as bad to the bone, we should regard this as an important boundary condition to the general tendency to conceive of true selves as good. Earlier work has already shown that, if an evil true self is specified, people will go along with this proposition. So we do know that it’s possible for people to at least accept the premise that people can have bad true selves. The outstanding question is, do we ever naturally see some people as bad “deep down”? We suggest psychopaths as one candidate in our paper. Perhaps racial outgroups form another.

Fomina, in her commentary, points to the evil spirits of folklore. Demons certainly seem to be bad—are their true selves bad as well? Culture is also rife with villains with a history of goodness (Darth Vader and Satan both come to mind). Given that nonhuman entities are seen as having a good underlying essence too (De Freitas et al., in press-a), it seems reasonable to expect this effect generalize to spirits.

I would be excited to see studies testing whether the supernatural true self can be seen as morally bad. Until then, I remain cheerfully skeptical.

*

I am grateful to Simon Cullen for articulating a promising new avenue for future research. A necessary first step in research on the true self was to demonstrate that attributions of the true self differ from attributions of the self. But having demonstrated this, how does this pattern of attribution change when situational factors are added into the mix?

Anyone who has waded into the attribution literature knows what a formless mess it is. Bertram Malle makes a valiant effort to rescue some basic conclusions from this morass in his epic 2006 meta-analysis. One of his conclusions was that the classic actor-observer asymmetry—where behavior of other people is attributed to the person, but one’s own actions are attributed to the situation—only holds when the behavior is negative. George and I ran some studies last year that

attempted to expand on this: Surely this effect should be nullified when making valenced attributions about the true self? What we found was a preference to attribute positive behaviors to *both* conceptualizations of the self—much like what Cullen reports. That is, we get the predicted effect for true self attributions, but we fail to replicate the supposed preference for attributing negative behaviors to the actor.

Assuming they are not completely capricious, attribution processes appear to be incredibly sensitive to specific circumstances. The majority of the literature on which Malle based his conclusions were studies dealing with attributions about skills like test performance (e.g. did Mary ace her test because she's smart, or because she studied hard?), not moral behavior. Perhaps it will turn out that a critical factor in attribution is whether we are reasoning about moral behavior. Moral judgment may draw out thinking about the true self, thus leading to a completely different pattern of attribution. (See also Pizarro et al., 2003 for a nice example of the complex interplay between valence and attribution when moral judgment is at stake.)

*

What riled me the most while I was writing this paper was how often psychologists (particularly of the older-school, self-help variety) have treated the "real me" as something lying in wait to be discovered, obscured beneath the schmutz of society and childhood trauma. It is obvious that this conceptualization of the true self is scientifically indefensible, not least because these discussions never seemed to be evidence-based.

I am open, though, to the true self existing in the more limited way that Brent Strickland proposes. The original studies from Knobe and company show that the true self resides in neither first order desire nor second order desire, as various philosophers had proposed. Between warring factions within the self, liberals ascribe the first order desire (homosexual urges) to the true self, whereas conservatives ascribe the second order desire (thinking one ought to resist such urges) to it.

A possible conclusion of findings like this, and the one we advance in the paper, is that the true self is radically subjective, and thus not a scientific concept. (As an aside: the paper's coda was only added under duress by our editor. All of us would have preferred to remain agnostic on whether the true self exists, but in retrospect I'm glad our collective arm was twisted.)

Another possible conclusion is that, discounting the biases and projections of third-party judgments, the first-person understanding of the true self reflects a coherent mentality. It's doubtful, of course, that the content is identical across individuals, but perhaps it is meaningfully similar at a more abstract level, reflecting, for example, one's personal values. An analogy could be drawn here with moral psychology. While there are individual and cultural differences on which issues are moralized, there are patterns in what is moralized, out of which more general rules can be proposed. Morality becomes a psychological concept through a mutual interplay between studying what issues people consider moral and formalizing the patterns that appear in these responses.

My only caveat is that, in such a case, I should want to dispose of the loaded term "true self". It suggests a deep and abiding epistemic reality within, which is a burden this tender concept cannot possibly bear.

Citations

De Freitas, J., Tobia, K., Newman, G., and Knobe, J. (In press-a). The good ship Theseus: The effect of valence on object identity judgments. Cognitive Science.

De Freitas, J., Sarkissian, H., Grossman, I., De Brigard, F., Luco, A., Newman, G., and Knobe, J. (In press-b). Consistent belief in a good true self in misanthropes and three interdependent cultures. Cognitive Science.

De Freitas, J. and Cikara, M. (n.d.). Deep down my enemy is good: Thinking about the true self reduces intergroup bias. Unpublished manuscript.

Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. Psychological Bulletin, 132(6):895–919.

Pizarro, D., Uhlmann, E., and Salovey, P. (2003). Asymmetry in judgments of moral blame and praise the role of perceived metadesires. Psychological Science, 14(3):267–272.