‘True Self’ Journal Club

The International Cognition and Culture Institute

March 2016

Some recent work in experimental philosophy and in social psychology addresses central issues in cognition and culture. Case in point: an article by Nina Strohminger, George Newman, and Joshua Knobe entitled "The True Self: A psychological concept distinct from the self" (forthcoming in Perspectives on Psychological Science and available here.) We thought discussing both the substance of this article and the place of this kind of work in research on cognition and culture would be worth a "journal club" webinar at CognitionAndCulture.net. It includes a précis of the article prepared by the authors; seven commentaries by Simon Cullen, Ophelia Deroy, Victoria Fomina, Larry Hirschfeld, Gloria Origgi, Brent Strickland, and Radu Umbres; the reply of the authors; and a general discussion.

TABLE OF CONTENTS

PRÉCIS OF "THE TRUE SELF: A PSYCHOLOGICAL CONCEPT DISTINCT FROM THE SELF"
Nina Strohminger, George Newman, and Joshua Knobe *page 2*

THE TRUE SELF AND THE SITUATION
Simon Cuellen *page 5*

TO BE OR NOT TO BE TWO?
Ophelia Deroy *page 9*

THE TRUE SELF, SUPERNATURAL AGENTS, AND THE PROBLEM OF EVIL
Victoria Fomina *page 11*

TRUTH AND CONSEQUENCES
Lawrence Hirschfeld *page 13*

THE “TRUE SELF,” MORE COMPLEX, MORE SOCIAL
Gloria Origgi *page 15*

IS THERE REALLY NO SUCH THING AS THE TRUE SELF?
Brent Strickland *page 17*

ANTHROPOLOGICAL DOUBTS ABOUT THE MORAL “TRUE SELF”
Radu Umbres *page 19*

NINA STROHMINGER'S RESPONSE: A FRIENDLY DESULTORY PHILIPPIC
Nina Strohminger *page 21*

GENERAL DISCUSSION *page 26*

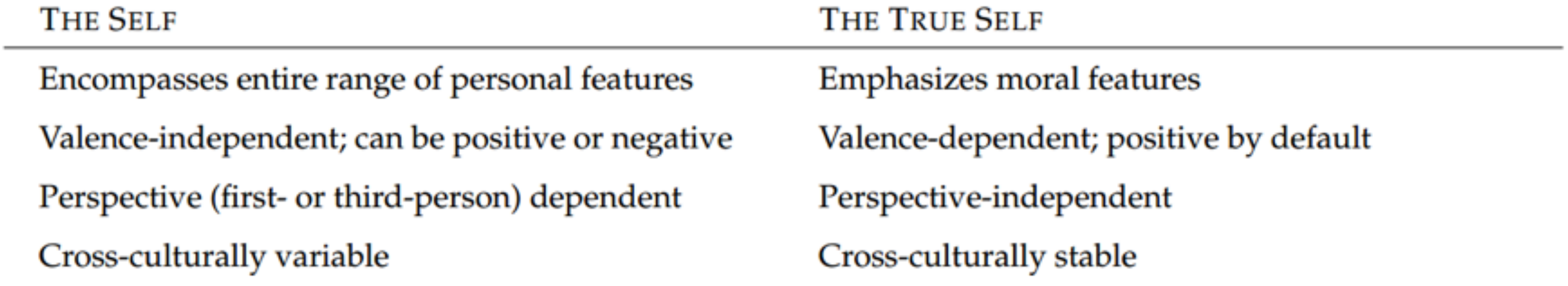
1

PRÉCIS OF "THE TRUE SELF: A PSYCHOLOGICAL CONCEPT DISTINCT FROM THE SELF"

By Nina Strohminger, George Newman, and Joshua Knobe

James Stewart, Donna Reed, and the young Karolyn Grimes in Frank Capra's film It's a Wonderful Life.

In self research, a boundary is typically drawn around the 'self' and everything else (other people, the environment). But emerging research shows that a further distinction can be made. Among those characteristics that are part of the self, a subset are seen as belonging to the *true self.* The contents of the true self are believed to make a person who they really are, deep down. People's understanding of the true self differs from their understanding of the self, in several key respects:



The Self	The True Self
Encompasses entire range of personal features	Emphasizes moral features
Valence-independent; can be positive or negative	Valence-dependent; positive by default
Perspective (first- or third-person) dependent	Perspective-independent
Cross-culturally variable	Cross-culturally stable

While the self is comprised of a long list of mental and physical features, the true self is primarily moral. Moral features contribute to perceived identity more than any other personal feature [1,2]. Moral traits are also considered to be the most deeply rooted, causally central aspect of a person's identity [3]. This pattern is quite robust. It shows up regardless of the context (changes brought on the aging process, medical interventions, supernatural events), and regardless of the type of moral feature (disposition, behavior, or belief); [4,5]).

The true self is not merely moral, but morally good. When asked which part of the self is responsible for a person becoming bad (e.g. a deadbeat dad), subjects attribute this change to the surface self, but becoming a better person (e.g. a loving father) is attributed to the true self [6]. This effect is contingent on the values of the person rendering the judgment: liberals think homosexual urges are part of the true self, but conservatives think it is not. Though we are perfectly willing to conceive of people as bad, we are unwilling to see them as bad deep down.

One of the more fervent research programs of social psychology has focused on actor-observer asymmetries. Yet the true self appears to be perspective invariant. People judge that their own true selves are morally good, but they also judge that other people's true selves are morally good.

A long tradition of research uncovers dramatic differences in conceptions of the self across cultures. Nonetheless, preliminary investigations find that the true self is seen as morally good across a variety of cultures. This finding shows up even amongst the notoriously misanthropic Russians and Tibetan Buddhists, who expressly disavow the existence of the self [7,8].

One possible explanation for these findings is that true self attributions arise as a result of more domain-general cognitive processes. In support of this view, recent work finds that a wide range of nonhuman entities are also seen as essentially good [9,10]. Recent studies show that beliefs about the true self are characterized by telltale features of essentialist reasoning, such as immutability, informativeness, and inherence [11]. The features attributed to the true self might therefore be due to the influence of psychological essentialism.

References

[1] Strohminger, N. and Nichols, S. (2014). The Essential Moral Self. Cognition, 1(31), 159–171.
[2] Strohminger, N. and Nichols, N. (2015). Neurodegeneration and Identity. Psychological Science, 26(9), 1468–1479.
[3] Chen, S., Urminsky, O. and Bartels, D. (2016). Beliefs about the Causal Structure of the Self-Concept Determine Which Changes Disrupt Personal Identity. Psychological Science, Vol. 27(10), 1398–1406.
[4] Heiphetz, L., Strohminger, N., and Young, L. (in press). The Role of Moral Beliefs, Memories, and Preferences in Representations of Identity. Cognitive Science.
[5] Molouki, S. and Bartels, D. M. (in press). Personal change and the continuity of identity. Cognitive Psychology.
[6] Newman, G., Bloom, P. and Knobe, J. (2014). Value Judgments and the True Self. Personality and Social Psychology Bulletin, 40, 203-216.
[7] Garfield, J. L., Nichols, S., Rai, A. K., and Strohminger, N. (2015). Ego, egoism and the impact of religion on ethical experience: What a paradoxical consequence of buddhist culture tells us about mo-ral psychology. The Journal of Ethics, 19(3-4), 293–304.

[8] De Freitas, J., Sarkissian, H., Grossman, I., De Brigard, F., Luco, A., Newman, G., & Knobe, J. (in prep). Is there universal belief in a good true self?

[9] Knobe, J., Prasada, S., and Newman, G. (2013). Dual character concepts and the normative dimension of conceptual representation. Cognition, 127, 242-257.

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

[11] Christy, A., Schlegel, R., and Cimpian, A. (in prep). Why do people believe in true selves? The role of psychological essentialism.

2

THE TRUE SELF AND THE SITUATION

By Simon Cuellen

You are probably already familiar with Darley and Batson's (1973) study. Participants were students at the Princeton Theological Seminary. As part of the study, they were asked to give a short sermon in a nearby building. Half of the seminarians were told they were running late, so they'd better hurry to the building. The other half were told they had plenty of time, but they might as well mosey over. On their way, both the hurried and relaxed seminarians encountered a man slumped in a doorway, groaning.

The overall pattern of results won't surprise you, but the strength of the effect might: relaxed seminarians were *six times* more likely to help the injured man. Even if they were on their way to deliver a sermon on the parable of the Good Samaritan, it didn't matter---hardly any hurried seminarians stopped to help.

Take a moment to consider one of the hurried seminarians, rushing along, trying to figure out what he will say in his sermon. He notices the injured-looking man in the doorway, but he doesn't stop to help. He just keeps walking. Question: what caused this callous behavior? Does it seem to arise from within *the seminarian himself,* or does it seem that something about the external *situation* caused him to ignore the man? The common thing to say, of course, is that the cause does *not* lie within the seminarian himself. After all, this very seminarian would have helped, had he been in the relaxed condition. So, it seems the situation is to blame.

It seems like a natural verdict! But what does it mean? In their excellent review article, Strohminger et al. describe recent research that suggests the natural verdict is importantly ambiguous: do we mean that the cause of the hurried seminarian's callousness lies "on the sunny side of his epidermis," or do we mean that the cause is not a part of his *true self*?

According to Strohminger et al., the true self is assumed (defeasibly) to be morally good. So, if the relevant distinction is between those actions which are caused by the seminarian's true self and those which are not, it is easy to see why people would naturally attribute his callousness to the situation.

But if this is so, it should also seem that the relaxed seminarians, nearly all of whom offer to help the victim, are *not* led to help by the situation. Rather, their kind actions should seem to arise from deep within themselves. Put another way, if the question "Is the action caused by the person or by the situation?" concerns the actor's true self, we should expect to find a surprising asymmetry in how people explain good and bad actions. Actions that are perceived to be good should seem more person-caused than those that are perceived to be bad, even when they are otherwise exactly alike.

In my own studies (unpublished, as of this writing), I have consistently found this pattern in participants' explanations for morally valenced actions. If you want to predict whether an observer will explain an action more in terms of the actor herself or more in terms of the situation, you can't do better than to find out about the observer's moral attitudes towards the action.

Why did the young woman decided to terminate her pregnancy? People who believe that abortion is morally evil are far more likely to say that she aborted because her boyfriend had recently broken up with her. By contrast, people who believe that abortion is morally permissible are much more likely to locate the cause of her decision within *the woman herself*. Why did the evangelical Christian man, who believes homosexuality to be immoral but is himself attracted to other men, give into his erotic urges? People who are disgusted by male homosexuality are much more likely to say that it had something to do with a stressful event he endured earlier that day; people with positive attitudes towards homosexuality think it's because he's gay. Why did the white woman decide to convert to Islam? People who believe that "Muslims are dirty" are much more likely to say it was because she was peer-pressured or because she came from a broken home; people with more positive moral attitudes towards Muslims prefer to say she converted because she's a spiritual person.

These findings make good sense if it is beliefs about the true self that are relevant to whether people perceive actions as arising from within the actor or the situation. On the other hand, if the self is understood in a morally-neutral sense, as is common in social psychology, it is less clear why the data should pattern in this way. Think of it in terms of the Good Samaritan study. The experimenter and experimental conditions are equally 'outside the skin' of the relaxed and hurried seminarians, so if the epidermis-centric concept of the self is the relevant concept, shouldn't both the kind and callous actions seem *equally* caused by the situation? [1]

The research described by Strohminger et al. suggests that ordinary people are largely unconcerned with the boundary between the sunny and meaty sides of the epidermis. By contrast, the list of psychological phenomena in which beliefs about the true self play an important role is growing quickly. Conceptualizing the person/situation distinction in value-neutral terms may therefore be unproductive for social psychology; in fact, it may be seriously distorting. To better understand how people perceive the causes of actions, social psychology should investigate the concept of the true self.

Note

[1] The person/situation distinction is often analyzed in terms of "causal covariation." The rough idea is that an action is caused by those factors with which it covaries. On this view, an action is situationally caused to the degree that the actor would not have performed the action, had the situation been relevantly different. But this analysis also will not do: both the hurried and relaxed seminarians would have behaved differently had they been assigned to different experimental conditions. So this analysis, too, fails to capture the intuition that the callousness of the hurried seminarians, but not the kindness of the relaxed seminarians, is caused by the situation. In presently unpublished studies, I compare the effect of observers' moral attitudes and the effect of their beliefs about covariation on how they explain morally-valenced actions. In the cases I've tested, the effect of covariation information is insignificant, both in absolute terms and by comparison to the powerful effect of participants' moral attitudes.

Comment

Samuel Vessière

When, how, and by whom moral responsibilities and intentions are attributed to agencies other than the 'true self' in the evaluation of a wrongdoing is a very pertinent question for cognition and culture.

Running your experiment with non-WEIRD populations is likely to yield fascinating results.

Alessandro Duranti has called for a timely research project on what he calls the "intentional continuum" across cultures: namely, building on Robert Levy's work on the cognition of emotions in Tahiti, the extent to which agents' intentions and propositional attitudes 'as such' are hyper-cognized to hypo-cognized across cultures.

When Rita Astuti and Maurice Bloch replicated Jonathan Haidt's moral dumbfounding scenarios among the Vezo of Madagascar, for example, they found that, in the accidental 'separated-at-birth' incest scenario in which the child turns out fine, the situation is judged as "wrong" regardless of the actors' intentions, since the act of incest will upset the ancestors and cause catastrophes. We could call this view (my term) a 'cosmic-consequentialist' position, which is likely to be found among many non-WEIRD people, and all other cultures (like the famed 'Opacity of Mind' examples from Melanesia) where percolutionary forces are given more weight that illocutionary ones. Without going as far as Julian Jaynes on the 'unconscious Greeks' or post-structuralist conspiracy theories that present psychological interiority as Evil inventions of Locke, Descartes, and Kant, we have good reason to believe that our Western ancestors did not attribute much causal power (moral or otherwise) to an interior true Self. As Bernard Williams points out, Oedipus blinds himself for acts that he knows he did not intend to commit.

All this to say that I am intrigued by your findings: intuitively, I would have guessed that morally conservative people in intention-centric, epistemologically individualist cultures would have placed a lot of moral and intentional weight on a person's 'true "bad" self' And then again, it also goes to show (perhaps), that conservatives are less individualistic, and closer in folk-intentionality styles to their non-WEIRD cousins.

Astuti, R., & Bloch, M. (2015). The causal cognition of wrong doing: incest, intentionality, and morality. Frontiers in psychology, 6, 136.

Robbins, J., & Rumsey, A. (2008). Introduction: Cultural and linguistic anthropology and the opacity of other minds. Anthropological Quarterly, 81(2), 407-420.

Duranti, Packer, M. J. (2016). The Anthropology of Intentions: Language in a World of Others, by Alessandro Duranti: Cambridge, MA: Cambridge University Press, 2015, 237 pp.,

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological review, 108(4), 814.

Veissiere, S. (accepted with revisions). 'Varieties of the Intentional Experience: Cognitive Ideology, Translucence, Complexity. Anthropological Theory.

3

TO BE OR NOT TO BE TWO?

By Ophelia Deroy

I could have been more courageous I will be more forgiving. Such thoughts are frequent, especially at the beginning of the year, like now. We believe that we stand for certain values, or deeper virtues, which might not be expressed or realized in our day-to-day actions. We fall short of a certain idea of ourselves. What Nina et al.'s paper argues is that we tend to transform this shortfall into the idea of a deep true self, distinct from the superficial or more contextual self we express in everyday life.

This is the claim I want to focus on, before questioning their second claim that this is not just a self-serving bias.

The original claim of the paper is not that we entertain idealized moral versions of our selves. It is about the way we relate to our better selves. As I acknowledge for instance that I could have been more courageous, or will be more forgiving, I run counterfactuals or projections of a higher moral self. My thoughts then refer to another version of myself, or another temporal slice of myself: I am putting my self in the shoes of Ophelia-if-things-had-been-different, or Ophelia-in-2017.These are other selves, which I temporally substitute to my current self.

But that's not how the true self works. What is clear is that the true self is not a *counterfactual, future* or alternative version of me. The true self is thought to be me: It is actual, it is something which exists at the same time as the superficial self that I express in everyday life. When saying it is actual, I take it that it is not even just thought to be a set of moral dispositions or capacities, but a set of actual, fully possessed virtues. But if we all think that we actually have - or are - a true self, are we all so divided? While the paper makes a clear and convincing case for *what* the true self covers (i.e. moral traits), *how* we exactly relate to this true self, needs more clarity. How is the distance between deep and superficial self managed - and what do people think about the relation between these two 'selves' - if they are indeed thinking about this as two different selves?

I have nothing against the idea that concepts like 'personhood' or 'self' are used flexibly; I have nothing either against attributing inconsistent representations to people: Certainly, we might both think that we are one, and yet two. But one would want an explanation of how this 1-2 selves work, especially when there are good precedents on the philosophy-market. Kant thought that there was a transcendental moral self in us as well as in everyone - is the true self similar, and are we natural Kantians? Harry Frankfurt suggested that we have higher-order desires - a tendency to reflect on the larger values that our more worldly desires and motivations lead us to want and do. Isn't it possible to capture the evidence listed in favour of a true self in this more minimal way : We take some distance from our immediate desires and actions, and wish for that what we want and do to be loftier, nobler, more moral. What would be lost if the contrast between 'true; and 'superficial self' was changed for the contrast between the life we would like to live and the life we live? Other candidates also exist, note, in psychology: As shown by Tali Sharot, 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.

These alternatives raise questions: Couldn't the evidence beautifully reported in the paper be explained by ways of thinking about ourselves other than the dichotomy between true and superficial self? Here is another way to raise the question: Accept that the authors are right, and that we evolved to have such belief into a core, true self. What exactly does the belief in a true self do for us - and more crucially, what does it do that all these other ways of thinking about better moral us (as counterfactual or optimistic versions of us, as dispositions, or through higher-order desires) could not do?

Let me suggest a possible direction of defense: thinking about the true self as actual and part of us might be more optimistic than thinking of counterfactual or future versions of ourselves, or higher-order desires, which are more often accompanied with regrets or sense of striving. Our true self is here already - no need to look further. True self is not something abstract or distant - it is us. So representing an actual true self might be more useful to us. It is also a much better argument for reputation management: "Look, I can say, this is not really who I am".

That we extend this optimism about ourselves to others looks like a source of further questions.

4

THE TRUE SELF, SUPERNATURAL AGENTS, AND THE PROBLEM OF EVIL

By Victoria Fomina

Nina Strohminger, Joshua Knobe, and George Newman's compelling and thought-provoking piece on the "True Self" presents an original theoretical intervention into the vast body of literature on the self, which spans across several different disciplinary and epistemological traditions. Aside from making an important contribution to the existing theories of the self, the true self concept also opens up an avenue for raising a number of interesting questions in the domain of moral cognition.

One such question that merits further exploration concerns the relationship between the perception of agency and the attribution of moral qualities. The possibility of moral evaluations and attribution of the true (and the "superficial") self is commonly contingent upon the perception of agency (mental states and intentions), but does this relationship go the other way around? In other words, does the perception of agency immediately triggers the attribution of a self (and a true self, assuming the two always come in one package)? If the attribution of the true self is indeed independent of any culturally specific notions of human nature, then there is no reason for this phenomenon to be restricted to human agents only and one would expect such attributions to extend to any perceived agents, including animals and supernatural beings. The moral taxonomy of the latter, however, poses a challenge for the notion of the inherently good true self. The rich body of ethnographic research on supernatural agents might indeed supply a good chunk of supporting evidence for the case of moral essentialism, in so far as the essence of many supernatural agents is often described through a number of core defining character traits. Nonetheless, these ascribed core traits are far from always being morally good. The supernatural cosmologies of many cultures feature creatures, whose "deeper" self is described as either inherently ambiguous and volatile (all sorts of trickster characters) or as plainly and purely evil (the Devil and all kinds of demons).

To give just one example, consider the case of *exotiká* – the monstrous creatures and malicious spirits described by the anthropologist Charles Stewart in his *Demons and the Devil* (1991), an ethnography of popular supernatural beliefs in modern Greece. According to Stewart, *exotiká*, whose origin can sometimes be traced back to Greek antiquity, not only co-exist with but have been successfully integrated into the Christian Orthodox cosmology, within which they assume the role of mediating the

Church's rather abstract and philosophical notion of evil and personifying the demonic forces. Not all *exotiká* are seen as unambiguously evil. In fact, they are often represented as dangerous trickster characters, an encounter with whom might on a very rare occasion yield some benefit for a person, but in majority of cases would be harmful or deadly (p. 175). This ambiguity, however, leaves little space for the attribution of a morally good true self. Rather, the occasional out-of-character manifestations of benevolence on the part of *exotiká* are seen as a devilish trick – an attempt at disguising their demonic nature in order to deceive, seduce, and manipulate their unsuspecting victims.

Provided that one accepts that people, who believe in supernatural agents, indeed process some of them as inherently bad, what are the implications of the example of the *exotiká* and similar cases for the theory of the true self? One possibility would be to treat such an example as anecdotal counterevidence against the moral goodness of the attributed true self. The other possible interpretation would be to view this case as evidence for the fact that the phenomenon of the true self attribution is robustly manifested in relation to the human agents only. But then, what is it about the specifics of processing of human and non-human agents that makes the attribution of the morally good true self invariable in the former case and optional in the latter? Another related question is whether it is for some reason easier for people to ascribe the bad deeper self to non-human, rather than to human agents? Although there is presently no experimental data on the attribution of moral traits to supernatural agents (especially to the malicious ones), the very fact of cross-cultural and historical persistence of beliefs in evil spirits is suggestive of the ease and readiness with which people are prepared to imagine agents, who are essentially (perhaps even ontologically) bad. The concept of the true self thus inevitably generates its own version of the proverbial problem of evil, which yet remains to be addressed.

5

TRUTH AND CONSEQUENCES

By Lawrence Hirschfeld

I was delighted to discover that deep down the authentic me is a happy go lucky sort of guy, and others recognize this too. I am also a little skeptical, to be honest. The authors of acknowledge that there may be some individual variation in this tableau—although the choice between psychopath and seriously mentally ill is not what I had hoped for. The authors also place great emphasis on moral qualities; indeed, the inner, authentic, true self is bounded by a sort of moral paradise; falling out of the true self represents a morally "bad case" whereas falling into authentic self is a good moral case.

With all this emphasis on self, authentic, and moral, one would imagine that the authors would be moved to interrogate these concepts, but surprisingly they do not. Self, the causal nexus that stops at the skin, is a bit underdrawn. Lots of things stop at the skin which may be relevant to the self but are not in themselves the self; not the least of which is our central nervous system. Few would deny that the self is a function of consciousness which in turn is a function of the central nervous system and its systematic interaction with the world of experience. But this only situates self, it does little to identify it.

Moreover, for many of the world, the self is hardly distinguished by a dermal constraint. Much of the self—much of *who we are* and *who ourselves see us as*—is contingent on the sort of person we are, defined by the kinds of roles we habitually assume, our place in and commitment to affiliative networks, and our roles in opportunistic collaborative activities in which "self-interest" (almost never defined in terms of the self, but the social universe in which the self is active) motivates us. In short, the self is an actor, a series of roles we assume, contingent on context and driven by need to serve various cross-cutting group interests. Almost nothing in the text suggests this possibility—perhaps because middle-class whites in Northern Europe and the United States pretend to be driven by an independent, authentic *me*. It is imperative that we realize that this is not part of a grand theoretical partitioning of the world's ways of being *me*, but a highly peculiar, decidedly not authentic cultural convention, ascribed to a small, distinct, but frighteningly wealthy group. Weird doesn't even cover how peculiar this is. Everyone else in the world, including these folks' wives, children, workers, etc. do not subscribe to this distorted and distorting notion of self.

It’s important not to confuse this with traits—crystallized thoughts that we habitually turn to in order to explain (or explain away) how or why we and others behave. Like other versions of theory of mind or mentalizing, these peculiarly gelled beliefs about ourselves and others are no more accurate other forms of mindreading. Mindreading isn’t a tool for accurately imagining another’s thoughts and using those thoughts to predict or retrodict their behavior; mindreading is a tool for imagining that we can do this. Luckily, we don’t rely on it that much; indeed, it is less precocious than a more important tool: reading successfully the social universe constituted of relations, roles, group membership and affiliative commitments (see Hirschfeld, Bartless, Whited, Frith, 2007; Hirschfeld, 2013). Arguably, it is part of core cognition (Spelke & Kinzler, 2007).

How about morality? Perhaps the least explored notion in the paper. Virtually no definition, or even hint of one, is provided. What is striking is many of the notions of morality that the paper cite are more aptly called norms or often valenced, social expectations which enjoy considerable distributional robustness, quasi-stability, and high degrees of relevance. None of this implies morality—systematic judgements of good (and bad) and the cognitive and cultural mechanisms that support these judgements. One of the striking findings of recent work on morality is a robust willingness to morally parse situations in the same way while explaining these choices in an equally vigorous idiosyncratic manner that belies reduction. Discursive treatments of these images of good and bad often then are not terribly informative beyond the possibility that individual differences in sociocultural contexts may play an important role. What the real self is when an individual is faced with a commitment to the weight of sociocultural commitment versus a powerful sense that this weight constrains a more “honest” (really sought after) *me* is no more obviously about a real versus superficial self than, say, a commitment to seek a professional football career in the face of facts of the matter that predict a very low hit rate. We live mentally and socially conflicted lives because we have shallowed the cool aid of choice. Most people in the world do not have nor pretend to have such choices. A theory that distinguishes the true versus superficial self, and casts normative commitments to the superficial heap, may not be meant as an affront, but it manages to perform one well. Selves, true or otherwise, are ways of imagining a wealth of options that simply aren’t available to most people. They are not cognitive outliers since they include a staggering percentage of the supposedly independent, self-oriented “West.”

6

THE “TRUE SELF,” MORE COMPLEX, MORE SOCIAL

By Gloria Origgi

“*I have entered upon a performance which is without example, whose accomplishment will have no imitator. I mean to present my fellow-mortals with a man in all the integrity of nature; and this man shall be myself.*” Thus wrote Jean-Jacques Rousseau as the incipit of his Confessions, a narrative elaboration of his authentic, “true self” as opposed to a hypocritical social identity. Rousseau confessed many weaknesses and failings but the message was clear: his true self, while rich and complex, was better than some of his objectionable actions.

While “true self” is an emerging concept in contemporary psychology, it played an important role in modern philosophy at least since Rousseau. The modern/romantic paradigm of “authenticity”, of a deep interiority that should guide our actions and is key to understanding who we really are, is one of the mainstream philosophical constructs of the last two centuries. Notions such as “false consciousness” or “bad faith” are at the core of many theories of twentieth century philosophy that contrast an authentic interiority to a social mask.

Here then is a first question. Social-psychological approaches such as the one developed in this article, ignoring literary, philosophical, and artistic sources of evidence, provide a very clear and simple picture of the true self as good an moral. Could it be, however, that simplicity and unambiguous valence of the true self so described is, to some important extent, an artifact of the approach?

A second question raised but not really answered by this article is whether people develop their sense of their “true self” on the basis of personal inner experience and self-reflection guided by biologically inherited essentialism, or whether it emerges and fulfill its function in social interactions, and in particular in the defense of one’s own “character” and reputation?

The evidence invoked by the authors, when taken together with less measurable but richer evidence from literature, history of ideas (or clinical psychology) fails to give clear and strong support to the idea that the “true self” is an essentialist folk concept about the morally good nature of ourselves and people in general. It is quite compatible with the alternative view 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. At the beginning of the twentieth century, the American sociologist Charles Horton Cooley called this “the looking-glass self.” This second ego is woven over time from multiple strands, incorporating how we think the people around us perceive and judge us (or how they should do so). The centrality of the “true self” in making sense of who we are and why we act depends on the crucial importance for us of preserving a positive image of ourselves not just in our own eyes, but also, and no less importantly, in the eyes of others.

7

IS THERE REALLY NO SUCH THING AS THE TRUE SELF?

By Brent Strickland

In “The True Self: A psychological concept distinct from the self,” Strohminger, Knobe, and Newman (henceforth “SKN”) outline a fascinating and compelling body of research on people’s naïve intuitions regarding the “true self.” The evidence suggests that there is a cross-culturally robust notion of the true self, which people conceive of as an intrinsically moral part of the self which causes positive personal changes and importantly contributes to establishing personal identity.

Here I’d like to ask why the true self is generally conceived of in this way? For example, why do people intuitively think that if someone goes from being an evil wife-beating drunk to a model citizen, that this was brought about by their true self? On the other hand, why do people think that if a person changes from being a model citizen to an evil wife-beating drunk, this is likely to be due to external or situational factors? The things to be explained here are thus (1) why is the true self conceived of as an essentially moral and positive aspect of the self and (2) why is this a near cross-cultural and cross-individual universal (excepting outliers like psychopaths but including misanthropes and pessimists)?

I want to explore an answer to this question that was discarded (perhaps too quickly) by SKN: that the term the “the true self” (and its equivalents) refers to something real whose nature is similar across the human species.

How would such a view play out? Imagine that the mind is composed of multiple dissociable parts (as “massive modularity” views suggest; Carruthers, 2006; Sperber, 2002), and can thus contain competing desires and priorities. Imagine further that some subset of these (perhaps unconscious and/or unrealized) desires and priorities aim to achieve a meaningful life. It may be that their collection is the referent of the term the “true self”, even if people may wrongly attribute to it a host of properties. It’s important to be clear about this last point. On this view, it might be that many of the things that people think about the true self (e.g. that it is immutable or immaterial) are just wrong. Nevertheless, the term “the true self” would still refer to something real. By analogy, the Vikings thought that lightning was actually bolts hurled by Thor. They were entirely wrong about this, but “*leiptr*” (Old Norse for “lightning" according to https://glosbe.com/en/non/lightning) still referred to a true type of physical phenomenon.

As SKN state, the idea of the true self shares an intimate connection with a deeper sense of meaning in life. Thus (in the authors' words) "Suppose a person has a desire to make a lot of money, and also a desire to create a beautiful work of art. This person may see both desires as aspects of her self, but to the extent that she sees only the latter as falling within her true self, the satisfaction of this latter desire will contribute to her sense of meaning in life in a way that the satisfaction of the former will not." Perhaps whatever it is that is generating the second type of desire is what our word "the true self" is referring to.

The authors reject the notion there being an actual true self on the grounds that it is unverifiable and radically subjective. However, it seems possible to empirically examine what makes or would make some individual's life seem deeply meaningful (for example, by asking them directly or by listening to their regrets on their deathbed). Moreover, I'm not convinced that the radical subjectivity point holds up to much scrutiny. It is quite plausible that those desires and priorities which contribute to creating a deep sense of meaning in life correlate almost perfectly with what different cultures and individuals consider virtuous. Indeed it seems that (at least within a given culture) deathbed regrets can be stable across individuals. This is according to Australian nurse Bronnie Ware who spent many years caring for the terminally ill in the last 12 weeks of life, and who recorded their deathbed epiphanies (https://www.theguardian.com/lifeandstyle/2012/feb/01/top-five-regrets-of-the-dying).

Moreover, according to Ware, the things that people regret most commonly do in fact appear to correspond to prioritizing the "superficial self" over the "deep self." People regret working too much, chasing money, chasing fame, etc… (which we think of as being superficial associated with the superficial self) while also regretting *not* spending enough time chasing dreams, expressing emotions, and spending time with friends and family (which intuitively correspond to the true self).

How to explain subjective variability in how people attribute properties to the true vs. superficial self? If an observer is lacking complete information about another's person's deep moral sense, but correctly understands that this person's perhaps (implicit) moral compass contributes to their sense of meaning in life, the most rational bet may simply be to attribute to that other person's true self one's own morality. Note that this could be the most rational thing to do even when faced with someone who screams to the rooftop that they think, for example, wife beating or money chasing is a good thing. Based on the above deathbed reports, apparently people commonly behave in ways that contradict their own deep moral compass.

So in short, I think there may be an empirically verifiable and non-radically subjective thing that the term "true self" refers to. If so, this may be sufficient to explain the cross-cultural and cross-individual stability of the true self-concept.

8

ANTHROPOLOGICAL DOUBTS ABOUT THE MORAL “TRUE SELF”

By Radu Umbres

Do people everywhere believe there is a true self – a moral true self, what is more? This is a question of obvious anthropological relevance. Most anthropologists, however, would question the basic assumptions of the hypothesis , not to mention its validity for many of the societies they have studied. After Marcel Mauss’s famous 1938 essay “A category of the human mind: the notion of person; the notion of self” that made this a standard anthropological issue, the universality of the category of self or personhood has been contested. It has been argued rather that the “self” has a particular intellectual genealogy, a history which makes it some version of it available to Western minds but not necessarily in all cultures or in all historical periods.

There is a rich social psychology literature on the notion of the self with a focus on comparisons between so-called interdependent cultures (typically Eastern) and so-called independent cultures (Western, in most cases the US), a contrast that many if not most anthropologists find way too vague and broad to be even worthy of discussion. Out of this literature, Strohminger Newman, and Knobe extract evidence to show that moral features are, across cultures, seen as more central to the self. This speaks only indirectly to the issue of whether the self/true self distinction is a universal one. They also quote a forthcoming article (De Freitas, J., Sarkissian, H., Newman, G. E., Grossman, I., De Brigard, F., Luco, A., & Knobe, J. (2017). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive Science*) presenting experimental evidence in direct support of their thesis. This evidence consists in experiments done in the US, Colombia, Russia, and Singapore (these last three treated as “interdependent” cultures). Participants in these three countries were, as is typical in this social psychological literature, university students.

To what extent do such studies with university students as participants genuinely capture the relevant range of cultural diversity? Can these studies really support the claim that belief in a true self is universal? University students in Bogota, Singapore, or Tomsk are likely to understand what is expected of them in a way more similar to that of American participants than would Colombian, Russian or South-East Asian peasants. They are likely to have their notions of the self strongly influenced by Western inputs. Serious cross-cultural comparisons should, moreover, involve people with little or no Western-type schooling, living in small-scale societies not dominated by a “moral” religion. They should be carried with methods that have some ecological validity, or to put it more simply, that make good sense in such cultural surroundings.

During my own fieldwork in a Romanian village, I studied cultural representations of personhood that are problematic for the idea of a “true morally good self”. For one thing, villagers portrayed certain individuals as truly immoral deep down, and their flawed character was feared and avoided even despite appearances to the contrary. This “misanthropism”, to use the term mentioned in the article, was selective, and did not extend, for example, to many other people believed to be fundamentally good - for example relatives and friends. A further example comes from a study with Catalina Tesar, where we compared these Romanian peasants’ and Romanian Roma (“Gypsy”) beliefs about ethnic essentialism. In our “switched-at-birth” scenarios, we found that Romanian peasants believe that Roma-born Romanian-adopted children will develop characters associated with their ethnic group of birth rather than of adoption. Tellingly, some traits in these hypothetical children (such a lack of inhibition against stealing or begging) are seen as rather immoral by Romanian peasants. Could we say that they are a reflection of a “true self” that villagers associate with Roma (but not Romanian) essence?

Even in Western societies, there is plenty of historical instances where people do not assume that, except for a few individuals, everyone has an essentially good and true self. An essentialized view of the self is compatible with the attribution of essentially bad “true self” (if this is the right way to describe it) to members of despised groups.. To take an extreme case, consider what Anti-Semites in Nazi Germany believed about the “true self” of Jews. For them, while Jews presented themselves as moral and law-abiding, they hiddenly behaved quite immorally, according to their true nature. Even partially Jewish origins were seen as implying that the “true self” was Jewish and hence evil. Note that negative essentialism is orthogonal to misanthropy (as discussed in the article). A Nazi could easily fuse, on the one hand, a deep negative assessment of Jewish essence with, on the other, strong trust and cooperative inclinations towards fellow Aryans assumed to have good, true moral selves.

I end with an exotic, but not less relevant example, of human reasoning about evil deep inside people. In some cultures, this idea of core evil is formulated in terms of either sorcery or witchcraft, two related but importantly different notions, as argued by Evans Pritchard in his classical work on the Zande. The sorcerer, it is thought, intentionally harms his enemy with magical techniques, while the witch may harm them unintentionally and even unknowingly. Yet in both cases there is a source of evil, or wrongdoing deep inside the person. We may not necessarily think of the evil intent of the sorcerer or of the witchcraft substance inside the witch as part of their “true self”, but it is nevertheless thought of an essential aspect of who they are: How do such cases fit, if at all, in the picture proposed by Strohminger, Newman, and Knobe?

9

NINA STROHMINGER'S RESPONSE: A FRIENDLY DESULTORY PHILIPPIC

By Nina Strohminger

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 Pirahãn 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.

References

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.

10

GENERAL DISCUSSION

Noga Arikha: “To thine own self be true”

This paper provides a very interesting account of how people understand a “true self”, based on instances of self-report. In previous work (“Neurodegeneration and Identity”, Psychological Science, 26:9 2015), Strohminger has shown that the families of patients afflicted with frontotemporal dementia, which affects moral judgement, report having lost that person’s “true self” more acutely than those of patients suffering from Alzheimer’s, for instance, which affects cognitive functions, memory and language. This is what has led her to conclude that the moral sense – rather than cognitive faculties – is more likely experienced as an aspect of the “true” self. A radical change to a person’s moral character – from kind and patient to impulsive and rude, for instance – is understandably experienced by others as a fundamental change in “who” that person is.

The use of such clinical data for investigations on the notion of self is of great value and needs to be pursued. The terms, however, cry out for more historically and philosophically grounded definition. This paper shows up via a larger set of studies a widespread psychological need to believe in an equivalence between truth and goodness – a strikingly Platonic, optimistic stance, “a hopeful phantasm”, as Strohminger writes in conclusion (perhaps a heartwarming one in these days when many in political power shun both truth and goodness). But as others in this discussion have pointed out, the questions posed in the various studies and the responses given by the participants need to be parsed and contextualised.

Strohminger herself calls the “true self” a “folk concept”, a “fiction”. It is also, as Larry Hirschfeld observes in his response here, specific to Western modernity – tied into the idea of authenticity, a product of socio-political developments, and so on. (A study of this is Jerrold Seigel, “The Idea of the Self: Thought and Experience in Western Europe since the Seventeenth Century”.) For a long time, the social self was the “true self” – rule-, role- and duty-bound. The idea of a lone, authentic individual, an essence set apart from others, is largely a product of Romanticism. Shakespeare famously has Polonius, in “Hamlet”, say to his son Laertes “to thine own self be true”. But Hamlet later brands Polonius a “knave”, and Laertes demonstrates his loyalty to his dead father by helping avenge his assassination at the hands of Hamlet, without considering what self he is beyond a son and an avenger.

Shakespeare is thus unlikely to have willed this phrase as the wise piece of advice it has since become, twisted to fit modern concerns with self-fulfillment and self-expression. Self-determination is limited where one must stick to an externally or internally given script and social role.

What we do, how we relate to others, and how we choose to act, constitute in part our self, as moral agent, just as, Gloria Origgi reminds us in her comment, the self is profoundly entwined with our social situation and how "we see ourselves being seen".

Strohminger chooses not to address what respondents in her investigation understand by "moral goodness", perhaps because that is not the point of her investigation. From this paper, however, it seems that all humans participate in a "folk" sense of the Platonic equation of truth with goodness, even as to define an action as "morally good" or not depends on innumerable variables, on how consequentialist or utilitarian one will be, etc. Further, to oppose a morally good "true self " to the "self" tout court is to assume that there exists a purer, better self than the complex, multifarious bundle the self is, the "organs and thoughts, desires and intentions, whims and dispositions", in Strohminger's words. Winnicott had talked of the "true self" as that most alive, authentic, spontaneous aspect of identity. His notion of a true self is separate from a moral sense, which encapsulates the capacity to recognize and separate out our desires and intentions, to direct our passionate dispositions and control our whims. I would argue that this very capacity partakes of the old, indeed cross-cultural acknowledgement of the need to curb self-serving impulses and direct our emotions through reflection, transcending selfish and ego-bound passion. And as Ophelia Deroy points out in her comment, our capacity to take such a distance from our current selves, and perceive ourselves as capable of doing and being "better" does not obviously require us to make the self/true self split.

In a more complex picture of an interplay of the moral sense with the sense of self, without recourse to a "true" self, the self would be defined by the ways in which the conscious "I" relates to immediate emotions, navigates the interplay of desires, wishes, duties, roles, dreams, the fluctuations between these, and deliberates in order to accommodate these fluctuations. This would unpack somewhat the "folk" picture Strohminger presents of people understanding "their emotional state as an expression of their true selves". For many people do take anti-depressants that change moods and sensory perceptions without fear of changing the sense they have of their core identity. It is unclear how the notion of a "true self" will help identify the changes wrought by the pharmaceutical substance, or how such essentialist talk, where the "true self" as folk construct seems the modern equivalent of the soul of old, helps understand what happened to someone whose cerebral damage has annihilated the person one once loved. How someone perceives the "true self" of their spouse will also differ from how the spouse experiences the "I", labile as such an entity must be. We do not know exactly how the sense of self is the outcome of cerebral processes, but we do know that it changes devastatingly when certain processes are damaged, that these changes are discrete, and that the explanatory gap between these processes and overall experience remains. Strohminger concludes that "What counts as part of the true self is subjective, and strongly tied to what each individual person herself most prizes". It seems a philosophically separate matter to argue that what we prize most must be seen as "morally good" in order to be valued.

Paulo Sousa : "The intuition of a moral true self and the evolution of cooperation and morality"

Here is a brief speculation. If the intuition of a morally positive true self is a default intuition about persons in general (about all selves and others as long as they are considered as persons), this intuition may have evolved in the context of the evolution of cooperation and morality as described in the work of Baumard and Sperber on the topic. In other words, it is a default intuition that facilitates cooperation. This speculation could also potentially explain the boundary conditions of this default intuition: when entities are considered to be outside of the moral circle, as in the extreme dehumanisation of outgroups, the default intuition would be cancelled or would loose its strength.

Dan Sperber: "The evidence raises interesting questions calling for more evidence"

In her reply, Nina writes :

“Many of the commentators point to qualitative, cultural, or anecdotal experiences as potential counterpoints to arguments [about the universality of the “true self”] in the paper. I think these sources of evidence are fertile sources for idea generation. …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.”

Quantitative evidence giving support to empirical claims on the one hand, and qualitative experiences as sources of idea generation? I have attended too many debates polarized in similar terms to hope that framing our discussion in such a way would really be helpful.

I see the social psychological literature on the self (and now on the true self) with its valuable experimental data as being itself a “fertile source for idea generation” rather than as having delivered a well-understood and robust theory in the matter. One somewhat vague but truly interesting idea that this literature suggests and for which it provides some initial evidence is that people in the West have an understanding of the true self distinct from their understanding of the self, that they conceive of the true self as positive and moral, and that some such conception is, if not universal, at least com-mon across cultures. To turn such an idea into a more precise hypothesis, some basic questions should be answered.

1. Are we talking of an evolved trait, or of a trait culturally acquired in cognitive development, or of the environment-sensitive cognitive development of a specific evolved disposition? I am not asking for an immediate answer, but for a discussion of what kind of evidence experimental or otherwise would help answer such a question.
2. A second related but different question: How much is such a conception of the true self part of a semi-explicit doctrine of selfhood and morality, which might vary across cultures and be absent in some of them, as opposed to being part of the spontaneous and intuitive way in which people

understand themselves and others (which would probably leave less room for important cultural variation). The evidence to answer such a question couldn't be, I take it, just experimental – and the experiments would be different from those now common in the field. Ethnographic and historical research would have to be called upon (as it is for instance in Garfield, J. L., Nichols, S., Rai, A. K., and Strohminger, N. (2015) discussion of buddhist idea in the matter).

3. What role do conceptions of the true self play in individual and in social life? Do they play a role mainly in the way people conceive of themselves and of others, in their individual attempts at explanation and prediction, as suggested for instance by Simon Cullen or Ophelia Deroy, and, in a comment in the general discussion, by Noga Arikha? Or do such conceptions play their main role in interactions with others, for instance in building and maintaining one's reputation, as suggested for instance by Gloria Origgi and, in a comment in the general discussion, by Paulo Sousa? Again, the evidence that would help us answer this question should, I take it, involve both novel experiment and ethnographic observations.

In the absence of well-developed answers to such questions, we don't really know what the available experiments tell us about. A sceptic would argue that they show that the kind of people who have been asked this kind of questions in these experiments tend to give the kind of answers the participants gave. I am not such a sceptic, not because I think the evidence is that much stronger, but because I think the available evidence, even if limited, suggests much more relevant questions and hypotheses some of which (that I might come back to later in this discussion) are well developed in the work of Nina and her collaborators.

Radu Umbres: "Morally familiar and morally-unfamiliar selves"

Following Nina's and Paulo's response, I would also push more on the issue of the role played true moral self in social cooperation. Perhaps there is a two-step way in reaching judgements about true selves, with social affiliation a discriminating factor in the way people evaluate others, and the moral true self works only when a boundary condition has been satisfied.

Here is a related question: how would people judge the true morality of other selves when the information involves moral duties entirely unfamiliar to the respondent? Such as strange cultural taboos : do not eat pumpkins on Fridays. What would a radically different – and incomprehensible in own moral terms – case study make respondents think about moral true selves? Would abiding to weird but contextually-strong moral duties make us perceive something about the self of a (culturally-) other?

And again following Paulo and Dan's ideas, would there not be strong selective pressures against a mental mechanisms biased towards seeing the real self of others as true and moral, if that could lead to social interaction with non-cooperators? Or is there an evolutionary advantage in being so socially-optimistic? Linking this with the earlier ideas about boundary conditions, perhaps there is a positive re-enforcing process when dealing with some people, but not with others.

Nina Strohminger: "Why morality?"

Noga writes:

“Strohminger concludes that “What counts as part of the true self is subjective, and strongly tied to what each individual person herself most prizes”. It seems a philosophically separate matter to argue that what we prize most must be seen as “morally good” in order to be valued.”

It seems too strong to say a trait needs to be seen as morally good *in order* to be valued. But the evidence does suggest that moral goodness is generally valued *more* than other traits, such as those related to warmth and competence (for an example of this phenomenon, see Goodwin et al., 2014). And this is a mystery in want of an explanation.

In earlier work I have suggested that the reason morality is to so deeply linked with identity is because keeping track of individuals is a necessary part of the fundaments of morality, such as cooperation and personal responsibility. But Paulo’s suggestion takes this a step farther: maybe morality is more valued in the selfhood conception because it facilitates cooperation. And indeed some work on individual differences in moral identity suggests just that (Aquino et al., 2009). I find this view entirely plausible.

I would like to add that we need not be committed to the view that there is just one reason why the true self takes the form that it does. It could be due to some mixture of the evolutionary origins of morality, cultural influence, and psychological essentialism. The essentialism point should not be forgotten I think. There is a positivity bias for the essences of persons, but also for the essences of non-human concepts (see De Freitas et al., in press). This suggests there may be domain-general cognitive mechanisms contributing to the true self’s features.

References

Aquino, K., Freeman, D., Reed II, A., Lim, V. K., and Felps, W. (2009). Testing a social-cognitive model of moral behavior: The interactive influence of situations and moral identity centrality. Journal of Personality and Social Psychology, 97(1):123–141.

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

Goodwin, G. P., Piazza, J., and Rozin, P. (2014). Moral character predominates in person perception and evaluation. Journal of Personality and Social Psychology, 106(1):148–168.

Dan Sperber: "The relevance of hypocrisy "

Let's consider the idea, suggested in Gloria's commentary and, in this discussion, by Paulo and Radu, that the main function of the image of people's true self might be to help identify reliable partners in cooperation and to be so identified by others (and idea to which I cannot be be partial – see Baumard, André & Sperber 2013; Sperber & Baumard 2012).

By advertising my own true self, I give others a fundamental and compact argument to the effect that they should see me as a basically moral person disposed to act for the common good. To the extent that such is the function of producing an image of one's own true self, it should of course be super-positive. But why should others accept this image at face value? Why should the true self of others be viewed as positively as one's own, as Nina et al suggest it generally is? Shouldn't one be vigilant regarding the fundamental moral dispositions of others? If everybody had a good moral true self, it would be irrelevant to differentiating people's reliability. This quasi-universally good true self might serve as little more than an automatic, lame, and superficial excuse for one's and others' failings. It may inspire feel-good Christmas stories. Still, what about the bad true selves discussed by Victoria and Radu?

At this point, I see two ways to go: (1) No, wrong idea: true self is not about trying to be positively valued by others and trying to value them at their real worth and in particular it is not aimed at informing trust. Maybe its main function is in one's relation with oneself rather than with others. (2) Yes, this defending one's own reputation and evaluating that of other may well be the main function of images of the true self but then the fact that the experimental evidence suggests that others' true self is also quite generally morally good should be viewed with suspicion: it might be an artefact of the experimental situation and design. After all, unlike what people do in real life, participants in experiments are not evaluating and choosing partners, let alone doing so in conditions where being too trusting might be quite costly.

I have no knock-down argument for (1) or for (2), but here is a type of (so far) non experimental, non-quantified consideration ("anecdotal"?) that might be relevant. Hypocrites! We are aware that there are hypocrites in the world. I am sure you know some personally. They make interesting character in stories, from Molière's Tartuffe to Sinclair Lewis' Elmer Gantry (not to mention the wolf in Little Red Riding Hood). They display an excellent self but it turns out that their deeper true self is evil. Forget novels and tales: people are on the look-out for hypocrites lest they become their victims. Imagine a study – experimental, why not? – aimed at investigating not, or at least not directly, the true self but people's understanding of hypocrisy. It might well throw an indirect but strong light on an aspect of common ideas of the true self and in particular on common ideas about bad true selves that the evidence discussed by Nina and others bypasses.

References

Baumard, N., André, J.B. & Sperber, D., 2013. A mutualistic approach to morality: the evolution of fairness by partner choice. Behavioral and Brain Sciences, 36(1), pp.59–78.

Sperber, D. & Baumard, N., 2012. Moral reputation: An evolutionary and cognitive perspective. Mind & Language, 27(5), pp.495–518.

Gloria Origgi: "True Self, Reputation and Character"

I still have some doubts about the whole thing:

1) It seems to me implausible that the "True Self module" might "fire" in the same way if the target is ourselves or other people. One of the authors of the article, Joshua Knobe, showed in previous works that people have a tendency to pay attention to "negative" moral traits, and this is confirmed by a huge literature on the very idea of responsibility (cf. Arendt 1987: "Collective Responsibility"). As Noga points out (and it is one of te major points of Derek Parfit's book, Reasons and Persons) identity and responsibility are deeply related: we have a "self" because people can track our actions and find us responsible of an evil action. The negative bias in responsibility attribution is well known. So the interest of a module to morally overrate your fellows seems odd.

But, as Dan says in his last comment (and I'm inclined to agree), the "TS-module" is more for self assessment, then, I really see it very close to the notion of character, self signalling and reputation (see J. Elster (2013) "Reputation and Character"), that is a form of self-signalling that can also motivate action in order to tell to ourselves how morally good we are.

And what about self – deception ? How reliable are we about our true selves? Don't we all have the tendency to morally overrate ourselves (cit. evidence/evidence/evidence)?

Also I share Larry's doubts about the universality of the concept. Cultural history and philosophy (cf. Bernard Williams 1993, Shame and Necessity) show that some conceptions of ourselves may emerge at a certain period of time (like feeling ashamed, so typical of the classic culture and feeling guilty, so typical of modernity, in a way that can be related to the emergence of the concept of "true self" and authenticity). Not only the geography, but also the history of the concept, might be much more local than the authors assume.

Victoria Fomina: "Is the true self effect limited to the domain of reasoning?"

Paulo, Radu, Nina and Dan's discussion of the relationship between the true self images and social cooperation raises a set of intriguing empirical questions regarding the role the true self images might play in social interactions. As Radu and Dan (in his latest comment) point out, taking the true self of others at face value makes for a poor strategy of identifying reliable partners for cooperation. Indeed, a scenario, in which people automatically attribute a morally excellent self to complete strangers and determine their attitude towards these strangers based on such an attribution, does not seem very plausible. Nor does such a radical scenario necessarily naturally follows from the experimental evidence laid out in Nina et al.'s paper. The evidence for the true self presented in the paper (if my understanding of it is correct) mostly stems from documenting the patterns in reasoning about the self of others based on the available pieces of information about these others' behavior. As such, it can hence provide little insight into how people in a real-life setting perceive the moral dispositions of strangers in a condition of complete vacuum of information regarding these strangers' past actions.

Perhaps, the true self is then better construed not as a default intuition about the character of others that affects every day social encounters, but as a bias in reasoning and interpretation of available information about the behavior of others that privileges the significance of positive acts and dispositions over the negative ones. In other words, it is in the domain of making a practical judgment about concrete acts of specific individuals, which involves reflexivity, deliberation, and reasoning, that the true self effect is manifest. While such a bias might not be the ultimate factor informing the decision about the others' attractiveness as potential cooperators, it might in some contexts function as a social mechanism of reputation calibration over time. After all, reputation building and evaluation is a long-term (even life-long in some cases) process. Although such a mechanism would not necessarily work in all the contexts (an inveterate criminal's overnight transformation into a model citizen is likely to be taken by the public with a healthy dosage of skepticism), it could function well in the cases of status transition. One example that comes to mind is a child's transition to adulthood: the change in the behavior that accompanies such a transition – suppression of selfish impulses and improving display of cooperative dispositions – would, one can argue, be perceived by the community as a manifestation of a young adult's true self. The ethnographic data on forms of moral pedagogy and the role narratives of moral transformation (including the feel-good Christmas stories mentioned by Dan) play in it can offer a fruitful terrain for further investigating the issue.

Samuel Veissière: "True Humeans, Intuitive Lockeans"

If there is a True Self module, I am inclined to think it is more likely a relatively stable, but culturally variable cognitive illusion that serves a cooperative function on the one hand, and a self-monitoring, self-rehearsal function for the adaptive purpose of self-deception (as suggested by Gloria) on the other. Against Jaynes' bicameral mind hypothesis, the illusion that we are the author of our own thoughts may be phylogenetically older than ancient Greece…but likely not much older than the evolution of explicit, declarative memory systems (which have also been shown to serve a self-deceiving prospective function). Could it be that the True Self illusion is still dependent on language in ontogeny, and many have started to spread epidemiologically at beginning of the broad spectrum revolution 100 or so K ago?

To make clear sense of the muddle, I like to ground these questions in philosophical debates on personal identity. First, we should ask whether the Lockean view of the "true Self" as a psychologically continuous bundle of episodic memories passes a cross-cultural and historical checklist.

I would say yes and no. On the one hand, a person must remember having had experiences they can attribute to themselves in order to feel like they are a Self, and these experience/memories must be validated intersubjectively by other people's memories and reports.

On the other hand, classical work in psychology (e.g. Elizabeth Loftus) and the current predictive-processing view of the brain clearly show that most memories are confabulated post-hoc and serve a rehearsal function for future action (see also Marh & Csibra's excellent piece in BBS on the communicative function of episodic memory), and that, the marvels of distributed cognition notwithstanding, it is very difficult to get two people to agree to agree on the veracity of each other's memories.

So ontologically, Hume's proto-Buddhist view of the Self as a bundle of disconnected mind moments (a bundle of whatever we feel and reconfabulate from moment to moment) might be more correct. Epistemologically, this is a more difficult question.

We could posit that like intuitive mind-body dualism (I defer to Paul Bloom on this question), humans are also intuitive Lockeans. Descartes and Locke may be wrong ontologically. As a cognitive scientist, I leave my ontology to physicists and biologists.

The Lockean view of the self as psychologically continuous, thus, may be a stable cognitive illusion that may or may not precede language phylogenetically, but got solidified with the advent of declarative memory systems, which themselves co-evolved with language.

Ontogenetically however, different narrative and moral environments [mutually expected standards of behaviour and how they are internalized, narrativized, with looping effects] may give rise to very different personal and inter-subjective experiences of "true selfhood". Or do they?

Here, Maurice Bloch's comments on Galen Strawson's "against narrativity" hypothesis may be helpful. Strawson proposes that the idea that we are the story we tell ourselves is overblown in the recent western Canon (Bruner, etc.). He argues that the difference may be one of personality traits. On this view, some people are clearly diachronic and feel a strong sense of psychological continuity from their childhood self to their current self.. Others — which he terms "episodics" – do not. Maurice Bloch hypothesizes that cross-cultural differences in Selfhood might be found at the level of diachronicity and episodicity. Western cultures may be more diachronic, and diachronic ideologies might loop back to further enhance the illusion (see Veissiere, 2016 for a discussion an a scale to test for diachronicity and episodicity).

But even then, the argument is not fully satisfactory. Clearly, there cannot be cultures where people are so episodic as to not remember themselves and each other. We are back at square one, so let's seek simpler clues from the phenomenologists. In Dan Zahavi's Merleau-Pontian reading of the question, the "core" Self ("minimal" in other versions of the same story) is what is universally found among humans and most sentient beings: the "true" self is simply that which experiences itself as sentient; that which is aware of itself as aware; that which experiences pains and feelings as its own: the so-called first-personal quality of selfhood.

We can now return to the dualist/monist question in the epistemological sense. Do all cultures primarily think of their and other people's selves as mental over physical? Like Paul Bloom and Galen Strawson, I would argue that all humans do. This is an argument that still doesn't appeal to many anthropologists, and for which we likely need more experimental evidence.

Once we have laid out all these basic problems and anthropologized them a little, I feel more comfortable returning to the fascinating question posed by Simon when he asks whether, where, and how people attribute some actions to the "True self", and to agencies other than the Self; .

It wouldn't be difficult to find cultures that de-emphasize agents' intentions in the assessment of a wrongdoing. Methodologically, running focus groups in the field would be more advantageous than working with Amazon Turk, where participants might already be a little too WEIRDed.

It will be more difficult to find ways of determining how children and adult across cultures conceptualize the seat of true personhood and agency.

Here are some ideas:

Querying people on how they conceptualize the continuity of personhood in old age after dementia. In WEIRD cultures, we tend to think of our elders as "no longer there" once they don't remember us or no longer behave according to our very rigid expectations on standards of behaviour. Would that pass a cross-cultural checklist? I don't think so!

The many licenses to legal and moral responsibility we WEIRDs grant to people in dissociative or psychopathological states is also very interesting. If someone was in a state of psychosis, intense inebriation, or even rage, we might agree that their True Self "wasn't there". Our folk ontologies regarding parasomnia (sleep) disorders is fascinating in that regard: sleepwalkers and sexsomniacs are usually not held responsible for their actions because the general consensus is that "they" were not there — yet, someone had these experiences and conducted those actions, even if their waking self doesn't remember any of it!. This is also very likely not to pass a cross-cultural checklist. Lo and behold, of course, current neuroscience tells us that the assumption that consciousness is absent from dreamless sleep was also misguided!

Conclusion: good psychology and neuroscience, like the Samaritan experiment that show how vulnerable to context our attention and actions are, tend to support the Humean, "Buddhist" view of the Self — and indeed of volition itself — as a cognitive illusion. If the Self is nothing but a pseudovolitional bundle of disconnected mind moments that arises from anticipation of experience primed by prior learning and responses to exogenous and interoceptive cues, then we WEIRDS may have an even weirder folk psychology than previously assumed.

So let's get to work with the non-WEIRDS?

References

Bloch, M. (2011). The blob. Anthropology of this Century, (1).

Chudek, M., MacNamara, R., Birch, S. A. J., Bloom, P., & Henrich, J. (2013). Developmental and cross-cultural evidence for intuitive dualism. Psychological Science.

Jaynes, J. (2000). The origin of consciousness in the breakdown of the bicameral mind. Houghton Mifflin Harcourt.

Martin, R., & Barresi, J. (2013). The rise and fall of soul and self: An intellectual history of personal identity. Columbia University Press.

Mahr, J., & Csibra, G. (January 19, 2017). Why do we remember? The communicative function of episodic memory. Behavioral and Brain Sciences, 1-93.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: the prospective brain. Nature Reviews Neuroscience, 8(9), 657-661.

Strawson, G. (2004). Against narrativity. Ratio, 17(4), 428-452.

Veissière, S. (2016). Varieties of Tulpa experiences: the hypnotic nature of human sociality, personhood, and interphenomenality. In. Raz, A, & Lifshitz, M. Hypnosis and meditation: Towards an integrative science of conscious planes, 55-76.

Von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. Behavioral and Brain Sciences, 34(01), 1-16.

Windt, J. M., Nielsen, T., & Thompson, E. (2016). Does Consciousness Disappear in Dreamless Sleep?. Trends in Cognitive Sciences, 20(12), 871-882.

Zahavi, D. (2009). Is the self a social construct?. Inquiry, 52(6), 551-573.

Brent Strickland: "Do "true self" representations have a function?"

I've found the questions about the potential evolved function of a "true self" representation to be interesting. Perhaps for example, as Victoria suggests, the possession of a bias to place a higher weight on positive acts than negative acts for the purposes of evaluating the potential for successful cooperation could be useful in some specific contexts. These are cool ideas, and I'm definitely not opposed to them as a matter of principle.

However, I'm not yet convinced that we have good reasons to be assuming evolved functions here at all. In particular, I still wonder if the cross-culturally similar notions of the "true-self" are due to some form of environmental sensitivity. As I was suggesting in my original reply to Nina, it seems quite possible to me that (i) humans have multiple competing desires (ii) some of those desires count much more heavily than others towards creating a sense of meaning (as can potentially be measured by deathbed regret) and (iii) morality plays a particularly important role in determining this sense of meaning. If these are generally true of members of the human species (and I think it's likely that they are), then there may be little need to talk about the evolved function of a true-self concept (or TS module or related cognitive biases along the lines of what Victoria suggests). The true self concept could simply be the product of sensitivity to an environmental regularity in much the same way that "lightening" and "thunder" concepts are. Both are presumably universal but have no evolved function per se (though the learning mechanisms that underlie general concept formation do of course). Whatever general mechanisms of concept formation underlie the acquisition and development of concepts like "thunder" and "lightening" may also be responsible for the formation of "true self" concept.

The view I'm laying out here predicts that you should find a fairly strong correlation between what cultures consider to be the most meaningful or important elements in their life (e.g. time with friends and family, pursuing one's dreams, helping others) and the traits that members of that culture would assign to the true self. How strong are the reasons to reject this "null hypothesis"?

The case that Nina mentioned in her main reply bears on this point somewhat. Between warring factions within the self, liberals attribute homosexual urges of others to the true self while conservatives attribute the desire to resist such urges to the true self. One might argue that they can't both be right. So roughly half the population is employing a representation of the true self that doesn't map onto an environmental regularity, thus falsifying my environmental theory. However, this interpretation could be complicated by lack of information (e.g. not knowing what other people will find deeply meaningful) or by true differences in local environments. Perhaps homosexuals who came out of the closet in the Bible belt (where conservatives are quite common) really do regret doing so on their deathbed, while latent homosexuals in NYC generally do regret not being more honest with themselves. If there's a difference in exposure between liberals and conservatives to these types of cases, differences in guesses about what belongs to the "true" (i.e. meaningful) self may be perfectly rational. So in summary, I think finding exactly the right type of information to falsify the "environmental regularity" hypothesis isn't straight forward, but I think it's necessary to do this work before moving on to the more interesting idea that representations of the true-self have an evolved function (in a way that differs with mental representations of "lightening" or "thunder").

Ophelia Deroy: "Moral optimism generalised?"

I am jumping on the bandwagon of the function of the ‘true self’, which Gloria, Dan, Brent, Victoria and others have already raised. More specifically, in my initial comment, i related to the ‘true self’ assumption to the general optimism bias which has been well-documented (e.g. Sharot, 2011 for review) – although i am sure that there are more cross-cultural studies to be done here. Basically, we tend to believe that we are less at risk of experiencing negative events (such as dying of lung cancer if we smoke, seeing one’s house burned down, etc.) compared to others. This self-directed optimism eventually makes us, among other things, more motivated and less risk-averse – it’s a good illusion to have about oneself, in a nutshell, and something which also has little costs by contrast with realism or pessimism.

The optimism bias, as defined in the literature, is about the likelihood of facing risks or tragedies more or less than others. Take now the likelihood of doing something morally good in the near future or if given a chance. Perhaps we all, optimistically, think that we would act in such a moral way.

In her response to my comparison, Nina is right to point that the optimism bias is only directed toward oneself, while the ‘true moral self’ bias (if this is indeed a bias) is, according to the evidence she quotes, equally applied to oneself and others. The optimism bias is studied in a comparative way, and is also contrastive between self and others by essence. Perhaps one does not need to be so contrastive when it comes to moral outcomes, and be generally optimistic that humanity as a whole, would be morally good if offered the chance. Your moral deeds do not diminish the value of mine, where as my likelihood of not being one of the 70% of smokers who have lung cancer partly depend on others being the unlucky ones.

My main point here is not to identify the ‘true moral self’ to the optimism bias of course. It is mostly to suggest that studying it as a bias instead of a folk concept or theory could be a fertile route – and this seems to be in line with what was suggested by others. This bias in turn could serve both a social function and an individual one (as the optimism bias does – it is good for our sense of self-worth, it is motivating, etc.).

Writing this comment and seeing Gloria’s comments, i also think that we should look more into the ‘equality’ assumption that is granted by the true self. I could be generally optimistic, and think that both you and I, if given a chance, would show our great moral dispositions or characters. But I could still think, that, given equal chances, I would show a greater moral disposition than yours.

Reference

Sharot, T. (2011). The optimism bias. Current biology, 21(23), R941-R945.

Nina Stromingher:

Thank you to everyone for such a fun and intellectually engaging discussion. True self research is in its infancy, and many of the points raised in this webinar lay out important directions as the field matures. I am excited to see where this new research takes us.