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.