



The true self as essentially morally good: An obstacle to virtue development?

Matt Stichter

To cite this article: Matt Stichter (2022) The true self as essentially morally good: An obstacle to virtue development?, Journal of Moral Education, 51:2, 261-275, DOI: [10.1080/03057240.2021.1887830](https://doi.org/10.1080/03057240.2021.1887830)

To link to this article: <https://doi.org/10.1080/03057240.2021.1887830>



Published online: 11 Mar 2021.



Submit your article to this journal [↗](#)



Article views: 1044



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

ARTICLE



The true self as essentially morally good: An obstacle to virtue development?

Matt Stichter 

School of Politics, Philosophy, and Public Affairs, Washington State University, Pullman, WA, USA

ABSTRACT

Psychological research has revealed that there is a strong tendency for people to believe that they have a ‘true self’, and to believe that this true self is inherently morally good. This would seemingly be very good news for virtue theorists, since this may help to promote virtue development. While there are some obvious benefits to people having morality intrinsically tied to their sense of self, in this paper I want to suggest instead that there may also be some significant drawbacks, especially when it comes to motivating virtue development and moral improvement. In part this stems from people’s belief in their own inherent moral goodness being merely assumed (as part of one’s core identity), rather than earned (say through reliably good moral behavior). This disconnection between identity and behavior can result in attempts to reinforce one’s identity as morally good, at the expense of virtuous behavior or self-improvement.



KEYWORDS

True self; virtue; moral identity; moral development; mindsets; essentialism

Introduction

Psychological research has revealed that there is a strong tendency for people to believe that they have a ‘true self’, and to believe that this true self is inherently morally good. As Newman et al. (2014) put it, the belief is that: ‘deep inside every individual, there is a “true self” motivating him or her to behave in ways that are virtuous’ (p. 211). This would seemingly be very good news for virtue theorists, such as myself, who want to promote virtue development. While there are some obvious benefits to people having morality intrinsically tied to their sense of self, in this paper I want to instead suggest there may also be a significant drawback when it comes to motivating virtue development and moral improvement, especially in the wake of moral failure. In part this stems from people’s view of their own moral goodness as inherent (as part of one’s core identity), rather than earned (say through reliably good moral behavior). This disconnection between identity and behavior can result in attempts to reinforce one’s identity as morally good, at the expense of engaging in moral improvement.

The goals of this paper are twofold. Primarily, I attempt to integrate research in psychology on the true self, moral identity, and essentialist mindsets, as well as reconciling that research with a virtue theory perspective (in philosophy) on moral development. Secondly, because the paper is somewhat speculative in nature about the connections

CONTACT Matt Stichter  mstichter@wsu.edu  School of Politics, Philosophy, and Public Affairs, Washington State University, Pullman, WA, USA

between these psychological and philosophical constructs, and is based on the connections I think the research is suggestive of (rather than conclusive of), a further goal of this paper is to generate fruitful hypothesis for further testing.

This article is divided into six sections. First, I will briefly review the literature on the true self and moral identity. Second, I will bring to bear a virtue theory perspective on moral identity and the belief in the true self as essentially morally good. Third, I will present mindset theory as a helpful framework for understanding why an essentialist view of the self may encourage setting problematic moral goals, as well as leading to maladaptive responses to failure. Fourth, I show how the phenomena of moral ‘credentialing’ and ‘cleansing’ could be indicators that people are adopting problematic moral goals, aimed at merely bolstering their identity as being good, which inhibits attempts at actual virtue development. Fifth, I discuss how our moral identities might generate multiple, and potentially conflicting, responses to moral failure. Finally, I offer some concluding thoughts on how to deal with the potential tensions between views of the true self as essentially good and virtue theory’s emphasis on continuing moral development.

The true self as essentially morally good

Recent work in psychology suggests that while there are many characteristics that make up people’s self-concept, some characteristics are seen as more fundamental than others. The concept of the true self picks out these most essential characteristics. Importantly, moral traits are seen as the most essential traits of the self. For example, Strohminger and Nichols (2014) studied people’s reactions to cases where people undergo a dramatic change, in order to see under what conditions people think someone changes to the point they are no longer the same person. They found ‘strong and unequivocal support for the essential moral self hypothesis. Moral traits are considered more important to personal identity than any other part of the mind’ (Strohminger & Nichols, 2014, p. 168). So, although our moral self constitutes only part of our overall identity, it is viewed as the most essential part of our identity compared to other mental faculties.

Furthermore, the true self is viewed not only in moral terms, but also as fundamentally morally good (Strohminger et al., 2017). This is revealed by studies, such as those done by Christy et al. (2016), which show that when people do good things they feel they are in touch with their true self, whereas actions that are morally problematic are viewed as a departure from one’s true self. Along these lines, Bench et al. (2015) found that the true self lends itself to metaphors of self-discovery, such that when people undergo positive changes, including positive moral changes, they are viewed as discovering who one really is. So, while people can see that their behavior has changed for the better, it is not viewed as also a change in one’s true self, but rather that their morally good ‘true self’ is somehow motivating this change.

A few other aspects of this view of the true self are worth noting. First, people attribute a view of the true self as morally good both to themselves and others (Strohminger et al., 2017). Of course, this is consistent with still thinking that other people are behaving poorly, but in such cases these people are seen as acting in a way that departs from their true self. Second, the above results seem to hold even when examined cross-culturally. For example, De Freitas et al. (2018) have shown that this connection between the true self and moral goodness is seen across different cultures, including societies that are seen as independent (e.g., United States) and interdependent (e.g., Russia, Singapore, and Columbia). Third, these

cross-cultural results might be due to the belief in the ‘true self’ being a product of psychological essentialism. As Strohminger et al. (2017) point out, ‘recent studies show that beliefs about the true self are characterized by telltale features of essentialist reasoning, such as immutability, informativeness, and inherence’ (pp. 556–557). This could account for why the belief in the true self is found cross-culturally, for psychological essentialism is itself a cross-cultural tendency. As Strohminger et al. (2017) also point out, ‘Positive, desirable personality traits are more essentialized than negative, undesirable traits, and essentialized traits are in turn seen as more central to defining who someone is’ (p. 553). This helps to explain why people tend to view the true self both in moral terms and as essentially good.

Further evidence for belief in the true self being a form of essentialism appears in work by De Freitas et al. (2017). Though, as they note, the previous research on psychological essentialism was regarding essentialism about categories (such as gender), whereas the belief in the true self focuses instead on an individual entity (such as a specific person). They speculate that individual and category essential features are tied by what are perceived as causally central features. Also, they note that another tie between these two lines of research is that essential features in general are seen as normatively good, even for non-human entities like an organization. Christy et al. (2019) also address the fact that previous research on essentialism deals with categories, while the true self belief is about the essence of individuals. They claim that the results of their research show that:

essentialist reasoning also guides how people understand the identity of individual persons. That is, the same processes that lead people to believe that category members possess a shared essence that accounts for their similarities and common identity may also lead people to believe that each person possesses a personal essence (a true self) that explains the regularities in their psychology and behavior across time and contexts and that makes them an individual with a distinct identity. (p. 402)

In this sense, for essentialism regarding individuals, there is a continuity across time for a particular individual, rather than the continuity that essential features provide between individuals of a particular category.

Finally, it is important to note that this connection between identity and morality is different from what is usually thought of as ‘moral identity’ in the psychological literature. As Lefebvre and Krettenauer (2020) explain:

The sense in which identity is typically used in the true self literature refers to numerical identity, or the continuity of the identity of an entity over time. This sense of identity should not be conflated with the term as it is used in developmental literature relating back to Erikson’s writings (e.g., Erikson, 1959). In this second sense, identity is conceived of as the accretion of personal commitments to particular ways of being in the world that provide the individual with a personal sense of unity and coherence. (p. 3)

One aspect of this difference is that regarding the connection between the true self and moral goodness, the emphasis seems to be on one’s capacity for moral behavior (rather than one’s specific moral commitments).¹ Strohminger et al. (2017) highlight that ‘people report the greatest identity discontinuity when *moral* capacities have been altered or removed’ (p. 552, their emphasis). Similarly, Strohminger and Nichols (2015) reported that ‘As long as core moral capacities are preserved, perceived identity will remain largely intact’ (p. 1477). So, the belief in the true self as essentially morally good more specifically relates to the having and preserving of capacities for moral behavior, and such capacities

are consistent with variation in one's actual moral behavior.² In which case, what remains underexplored is whether the belief in the essential moral goodness of the true self might influence (for better or worse) moral behavior in other less direct ways, including the development and exercise of virtue.

A virtue theory perspective on the true self

Regarding a belief in the true self, in some obvious respects it is good news if most people have an intrinsic concern about morality. However, what is unusual in this case is that the belief in the essential moral goodness of the true self is basically disconnected from one's actual behavior. That is to say, the idea that one's identity is rooted in morality is also being tied to a default assumption that one is in fact morally good, rather than this necessarily being based on a record of morally good behavior. Inherent moral goodness seems merely assumed, rather than earned. In other words, you would think that if you view yourself as morally good, that this perspective should be based on you being somewhat reliably good at behaving morally. To take a similar example from skills, if you're going to claim that you are a good tennis player, then presumably you're going to base that claim on a record of past behavior in reliably playing tennis well. But that does not appear to be what is going on with views of the true self as essentially morally good, where the ascription of goodness is independent of patterns of actual moral or immoral behavior.

This separation between views of the true self and actual behavior is supported by further psychological research, such as one study that found that even misanthropes still endorse the idea that people have an essentially good true self. As De Freitas et al. (2018) noted:

What is surprising about these results is that the very same participants who say that most human beings are awful also appear to hold the belief that human beings are fundamentally good deep down in their true selves. This result suggests that whatever cognitive processes are at work in people's true self judgments, these processes are remarkably unaffected by individual differences in judgments about other aspects of the self. (p. 12)

This study reinforces the view that the goodness of the true self does not arise from the behavior of the person in question, since one can both view a person's true self as good while at the same time expecting morally bad behavior from them. This is presumably explained in part by the focus of the true self belief on one's mere possession of capacities for moral behavior, rather than how such capacities might have been exercised over time. But there should be limits to how much immoral behavior one could engage in, consistent with still believing that one's true self is essentially morally good, as too much immorality could signal the loss of the moral capacities that were seen as central to one's continued identity.

In which case, such a view of the self as morally good stands in quite a bit of contrast to what would count as having a morally good character according to virtue theory. For a virtue theorist, being morally good involves developing and exercising virtues (at least to some degree). Virtues are acquired excellences, which take experience and practice to acquire, and thus one cannot merely assume that one is virtuous to begin with. Furthermore, to possess a virtue is to be reliable in displaying virtue relevant behavior. For example, a kind person is

one who reliably acts kindly when that is what the situation requires. The virtues are developed through sustained moral action, and so one cannot make claims about the possession of moral virtues independent of one's actual behavior.

Insofar as the true self is not something one acquires, yet is viewed as essentially morally good, then it does not seem to require any experience or practice in order to have achieved that goodness, contrary to the expectations of virtue theory. However, if we view the true self as essentially good in the more limited sense of motivating us to display virtuous behavior, then that could be consistent with what virtue theory requires, as we would still need to spend time and effort in acquiring virtue in order to reliably behave morally well. Or, in other words, we might need to exercise virtue in order to remain in 'touch' with our true selves. However, it is unclear whether there is any good evidence supporting the idea that a belief in the goodness of the true self has this kind of motivational influence on virtuous behavior.

Although judgments about the goodness of the true self are mostly unaffected by judgments about people's moral behavior, there may be evidence that while morally good behaviors are not the foundation for a belief in the true self as morally good, immoral behavior can pose a significant challenge to that belief. That is, insofar as a person believes that their true self is morally good, immoral behavior on the part of that person should challenge that belief (at least, assuming one does not morally disengage in response, as that would result in denying that the behavior was immoral). Some evidence comes from Lefebvre and Krettenauer's (2020) research on the true self, which indicated that a 'change from a good moral state to a bad one is more disruptive to perceptions of identity because it violates the assumption that the person was good deep down', and that 'positive moral changes are uniquely seen as uncovering some essential quality that already existed within (resulting in less identity disruption); by contrast, negative moral changes are perceived as impinging on that existing essence (resulting in greater identity disruption)' (p. 2). So, while immoral behavior is consistent with the belief that one still has an essentially good true self, it nevertheless poses some challenge to maintaining that belief.

That challenge might be met with defensiveness in order to maintain a positive self-view, or it might in turn motivate one to seek out opportunities to do some good deeds, in order to re-establish some certainty about one's moral goodness. Some support for this later response comes from research that reveals that acting immorally makes one feel less certain about who they are as a person. Christy et al. (2016) found that 'when people feel they have committed a moral transgression, they feel uncertain of who they truly are, and, conversely, when they perceive their behavior as moral, they experience feelings of self-understanding' (p. 9). One's own immoral behavior will be personally distressing in a way that the immoral behavior of others is not, and something must be done to manage that distress. But there are adaptive and maladaptive responses to that kind of distress, and in the next sections I provide reasons to be concerned about the belief in the true self contributing to maladaptive responses, given how characteristics that are seen as immutable make failure highly distressing.

Mindsets about morality

Insofar as people view the true self as a part of themselves that does not change, this bears a striking resemblance to having a 'fixed' mindset. I will first detail the distinction between 'fixed' and 'growth' mindsets, along with the consequences of those mindsets, before detailing

further parallels between true self beliefs and fixed mindsets. Dweck and Leggett's (1988) work on intelligence, amongst other abilities, shows that people fall along a continuum with respect to how they view abilities—it is more of a fixed entity that you cannot do much to change, or it is more malleable such that you can train incrementally to improve it. Regarding fixed mindsets, Bastian and Haslam (2006) provide evidence that a fixed view of an ability is also a type of essentialist notion, since the hallmark of a fixed view is its association with immutability (as is true of essentialism in general), and so this also provides a parallel between the belief in the goodness of the true self and having a fixed mindset. Furthermore, these two mindsets, 'fixed' versus 'growth', have different consequences for how we react to failure, as a fixed mindset leads to maladaptive responses because of feelings of helplessness to do better; whereas a growth mindset leads to adaptive responses where a person puts forth effort to learn how to act better the next time (e.g., recognizing that they made a mistake, but also that they can put effort into improving).

In general, a malleable approach is associated with more adaptive responses to failure, and a motivation to strive for improvement, both of which would be important for moral development. Malleable views also encourage 'learning' goals, where one is attempting to increase one's abilities, along with an acknowledgement that we often learn from our errors. Fixed mindsets, by contrast, encourage the adoption of 'performance' goals, in the sense that the goal is to demonstrate that one has a certain level of ability to oneself and/or others (Dweck, 2000; Dweck & Leggett, 1988). That is, the ultimate goal is that one wants to be judged positively, and avoid being judged negatively, such that errors are inherently threatening to that goal. In which case, one tends to avoid tasks that pose a significant risk of failure. Dweck and Leggett (1988) realized that one of the implications of their view was that 'the same conceptualization may be applied to the moral domain to illuminate the reasons or purposes for which individuals (at any stage of moral development) engage in moral actions' (p. 266). So, in the case of morality, this means that with a fixed mindset the goal is merely to demonstrate morality to others to avoid punishment, perhaps along with being motivated to behave in ways that are uncritical and conformist. The difference in these goals and the associated reactions to failure are important, because moral failures will be a part of anyone's life, and we will all have to learn in part from experience. So, from the standpoint of virtue development, it is important that moral goals are adopted as learning goals, both to encourage active virtue development and to allow moral failures to be perceived as opportunities to learn how to act better (Stichter, 2020).

Furthermore, there are implications for how those with a fixed moral mindset would react to seeing themselves as having made a moral mistake. On a fixed view, failures are so distressing because they are taken as evidence that one has low abilities, and if such abilities are fixed, then there isn't much that can be done to improve in response. In other words, failure has an additional distressing element to it for those with a fixed mindset, insofar as the failure is attributed to having low abilities or capabilities in general. In which case, people are likely to respond instead with defensive reactions (or with moral disengagement), to avoid that negative appraisal of their ability. So, in the moral case with a fixed mindset, failure can be taken as revealing that one is not as morally capable as one originally thought. This kind of implication of failure appears in some of the research on the true self, as Christy et al. (2016) note that 'the commission of immoral acts challenges people's fundamental tendency to view themselves as morally good, and, as a result, acting immorally has negative consequences for

how people view and feel about themselves' (p. 1). In this sense, your identity as a morally good person is being called into question, and this can produce a lot of distress.

To be clear, one typically ought to be feeling some distress in response to one's own moral failure, but it is important for moral development that one has adaptive responses to that failure (i.e., reparations and self-change). But the worry here is that insofar as an immoral act challenges the belief of essential moral goodness, and that belief is tied to a part of the self that is viewed as unchanging (i.e., fixed), then it might lead the person to have a maladaptive reaction in response (Christy et al., 2019). Though it should be noted that these mindsets are domain specific—people can have fixed mindsets about some abilities, and malleable mindsets about others. As a result, there are important individual differences amongst people regarding which abilities are seen as fixed or malleable. This appears to be a point of contrast with the belief in the true self, which is more of a universal tendency with little in the way of individual differences. Though for my purposes here, I'm not concerned with the individual variation that is typically found with mindsets, but rather with what implications might follow from the more universal tendency to view the true self in fixed terms. So, we know that people don't react well to failures associated with abilities that are viewed as fixed. Given the similarities between having a fixed mindset and the view of the true self as essentially good, this should give us reason to worry that moral failures which challenge one's identity as fundamentally good could be contributing to maladaptive responses.³ In the next section, I review evidence that people do seem to exhibit the kind of maladaptive responses to moral failure that you would predict from having adopted moral goals as mere performance goals.

Moral credentialing and cleansing—Evidence of 'performance' goals

Further evidence that suggests people may have fixed mindsets regarding some aspect of their moral identity can be found in Daniel Lapsley's (2016) discussion of two behaviors people seem to engage in after recognizing a moral failure. One is moral 'credentialing', such that 'when there are threats to moral identity individuals are more likely to over-estimate their moral credentials, as if to reassure themselves that their moral identity is secure' (Lapsley, 2016, p. 50). He cites work indicating that this credentialing process is basically one of over-estimating one's past moral behavior (Effron, 2014), in order to defuse the threat a potential moral failure would pose to one's overall identity. That is, one points to past good behavior to maintain one's positive moral self-image in the face of potential criticism of current behavior. This is basically a form of defensiveness, where one is seeking to deflect criticism of a recent action by directing attention to supposedly good past behavior.

Lapsley (2016) then contrasts this behavior with moral 'cleansing', where people engage in moral actions as 'a way to prop up or restore moral self-concept when one has engaged in (or merely recalled or contemplated) unethical behavior' (p. 50). He cites the work of Jordan et al. (2011) as showing that when 'moral identity is threatened by unethical conduct we are motivated to restore it by taking compensatory action' (Lapsley, 2016, p. 50). One important difference between 'credentialing' and 'cleansing' is that 'cleansing' at least motivates actively engaging in moral behavior, rather than merely recalling past good behavior. Lapsley suggests that this has implications for virtue, such that virtue is compatible with cleansing behavior but not credentialing. More important for my purposes here, he suggests that:

the distinction between moral performance (credentialing) and moral improvement (cleansing) tracks the dual mindsets ... On this reading I would suggest that there is both a fixed and incremental approach to moral self-identity; and that moral identity mindsets encourage individuals to pursue either performance goals that encourage the demonstration of moral credentials; or else learning or development goals that encourage behavior associated with moral cleansing (Lapsley, 2016, pp. 55–56).

Credentialing does appear to reflect the adoption of performance goals, as one is not grappling with actual moral failure, but rather looking to past behavior as evidence that one must still be morally good. The need to look to past behavior reflects a worry that the current behavior may show that one lacks the relevant moral abilities, such that the person needs to ‘cherry-pick’ from one’s past behavior, along with a bit of exaggeration, to try to protect oneself from criticism (see also Mullen & Monin, 2016 for a distinction between two forms this might take—‘moral credits’ and ‘moral credentialing’). This credentialing behavior then does not lead to any attempts to make up for one’s current wrongdoing, or to improve oneself.

However, it should be noted that while the cleansing behavior is better in the sense of motivating some form of moral action, it does not necessarily lead to moral self-improvement. In other words, it may not necessarily signal the presence of learning or development goals (as the contrast with credentialing behavior might suggest). It depends on what kind of action was motivated by the cleansing behavior. From the perspective of virtue theory, in the wake of moral failure, one ought to try to make amends to anyone harmed by one’s failure (e.g., if you were dishonest or cruel to someone in particular), and to strive to figure out how to avoid committing that mistake again in the future (i.e., moral improvement). In regards to the latter point, for example, if the failure is one of dishonesty, then you should try to figure out why you acted dishonestly, and what steps you could take to act more honestly in the future. But if the kind of moral action that is motivated in response (to the dishonesty example) is merely some ‘random act of kindness’ the next day (e.g., volunteering at a soup kitchen), then one is avoiding the work of moral improvement that we would hope to see (even if it would still count as prosocial or moral).

It might help here to use a skill example to make this point. If a basketball player misses a game-winning free throw, you would expect that they ought to be working on improving their free throws in the next practice session, and not trying to make up for it by practicing their 3-pointers instead (however useful that might also be). So to better test for the presence of virtuous motives, in the sense of someone having adopted a moral goal as a learning or developmental goal (and not merely a performance goal), one would need to examine whether people are also taking steps in the wake of moral failure to try to prevent such failure from reoccurring in the future.

Another reason to be cautious about what moral cleansing behavior signals comes from Daniel Batson’s (2017a) discussion of a more cynical view of such behavior, that of moral hypocrisy, which in his view is the ‘*motivation to appear moral while, if possible, avoiding the cost of actually being moral*’ (p. 21, his emphasis). With hypocrisy, one is focused on maintaining a morally good social image, rather than with a sincere concern with one’s self-identity as moral. Batson points out in relation to studies that show a prosocial behavior occurring after recognition of a moral failure (like with moral cleansing), especially for those people that score high on a moral identity scale, that:

it seems clear that participants knew of their moral lapse at Time 1 before being faced with the chance to be/appear moral at Time 2. If so, increased moral behavior at Time 2, which is what was found for high scorers on the Importance of Moral Identity scale, is what the hypocrisy motive should produce. (Batson, 2017b, p. 70)

This provides further reasons to be careful in what conclusions we draw from moral cleansing behavior, as it might be motivated by merely an egoistic concern to appear moral to others, as with moral hypocrisy, or by a more genuine commitment to morality. So, we might lack sufficient information regarding the behavior of the participants to know whether it was consistent with virtue or not.

In response to Batson, Karl Aquino (2017) suggests something like the latter, noting that ‘moral hypocrisy is one possible outcome of the self’s attempt to achieve two equally adaptive goals: (1) maintaining a belief in its essential goodness . . . and (2) striving for consistency’ (p. 54). While Aquino does not directly connect this to the research on the true self as essentially morally good, this would account for why striving for consistency could lead one to do morally good deeds in response to a moral failure. Though rather than ‘consistency’ per se, what could be motivating some cleansing behavior is the need to reestablish certainty about oneself after moral failure, as mentioned previously. Recall that Christy et al.’s (2016) research related to the true self and immorality provided evidence that immoral actions reduce people’s self-knowledge, and they go on to claim that ‘a perceived lack of self-knowledge activates motivations to re-establish certainty about oneself’ (p. 10). So, if one is trying to reestablish certainty about oneself as morally good following a moral failure, then presumably doing a good deed helps you to regain this certainty about your own moral goodness (or rather, capacity for moral goodness). In which case, this would also motivate the pattern of behavior described in the previous Batson quote (though not for reasons of hypocrisy), where an immoral action at Time 1 could motivate re-establishing certainty about one’s moral self, and the chance to be moral at Time 2 could fulfill this motivation. This is at least suggestive of one way in which cleansing behavior might be motivated by the belief in an essentially good true self.

Implications of moral failure for the true self

In reaction to a moral failure, uncertainty about oneself might prompt doing a good deed, in order to confirm to oneself that one is capable of moral goodness. On the one hand, this seems like a good reaction, insofar as it motivates doing something morally good. However, when this happens, it may result only in some temporary good deeds, followed by going back on ‘autopilot’ once certainty about the self is re-established. At some point, we likely end up feeling certain again about who we are, and the motivation to engage in certainty-establishing (and in this case moral) actions then presumably wanes (even if you don’t know in advance what it will take to feel that way). If so, then we are ‘free’ to go back to our normal routine, which then no longer includes those moral actions which we took only to get back to our authentic sense of self. In which case, such moral cleansing behavior would still reflect having merely moral performance goals, if the goal seems predominantly about demonstrating one’s capacity for moral goodness to oneself, rather than the kind of learning goals and self-improvement that would reflect a malleable mindset about morality. So before we can know that a particular behavior taken in response to a moral failure is consistent with virtue (or having a moral learning goal), we need to know how that behavior connects up to the

moral failure (i.e., are you making restitution to someone you harmed or working on moral self-improvement).

This is why the impact of the true self belief on moral improvement might reveal itself most in responses to moral failure, when there's a clear need for improving oneself (and not merely reassuring oneself that one is capable of moral goodness), but where that failure carries with it a potential threat to one's identity. Recall that having performance goals (and a fixed mindset) make failure threatening, as it questions assumptions about one's abilities, and can lead to defensive reactions. As such, moral failures may threaten one's sense of self, and thus may involve defensiveness in denying the moral failure in the first place (such as via moral disengagement), or recognizing it but only temporarily engaging in good deeds (i.e., easily achieved performance goals) in order to reassure oneself that one is good. It is for this reason that I raise concerns about whether the view of the true self as essentially morally good may be prompting reactions that inhibit self-change, in terms of the true self being viewed as a fixed part of one's moral identity.

Furthermore, to the extent that we have this fixed view of some aspect of our moral identity, then it could be a separate trigger for distress from a perceived threat of moral failure—that is, separate from the distress that comes from having violated one's own moral commitments (in terms of the developmental account of moral identity). In which case, a moral failure might have negative implications for both forms of moral identity, which also gives rise to the possibility of two different (and therefore potentially conflicting) responses.⁴ The reason for this would stem from the difference in responses to failure found in the two mindsets, given that moral behavior is more malleable (or at least relative to the fixed view of the true self as essentially good).

This possibility seems to have some evidence in support of it coming from Krettenauer's work on moral identity maintenance. On the relationship of moral identity to moral action, Krettenauer (2020) claims that 'moral actions are instrumental for moral identity maintenance' (p. 3). Now this will initially sound puzzling to virtue theorists, as moral (or virtuous) actions are supposed to be of intrinsic value (or constitutive of living well), and not of merely instrumental value as a means to furthering some other goal. However, Krettenauer's (2020) view is rather that:

any honest, caring or fair behavior can be motivated by the desire to do what is considered morally right or good and by the goal to maintain one's moral identity. One goal does not come at the expense of the other. Instead both can support each other. Moral identity as a goal adds another motive for moral action to the desire to do what is good or right. (p. 3)

Thus, in this sense, moral identity maintenance is supplying a further goal (and thus further motivation) to do what is morally right. While a virtue theorist would likely claim that a desire to do what is virtuous ought to be the primary motive, additional motivations to be virtuous can help people act morally well when they might be tempted to do otherwise (perhaps because the desire to be virtuous is not strong, or there are strong situational pressures to do otherwise).

However, moral identity maintenance as a goal does not necessarily promote moral action, as Krettenauer rightly notes. In fact, it can promote actions at odds with morality. The reason is that insofar as:

action is instrumental for achieving the goal of moral identity maintenance, there might be other means for achieving this goal that are equally effective. Individuals may deny the moral

significance of an action by using various *strategies of moral disengagement* (Bandura, 2016). They may, for instance, minimize the negative consequences of an action for others or deny their own responsibility to act. As a consequence, one's moral identity remains unaffected by whatever course of action is taken in a given situation (Krettenauer, 2020, p. 3, emphasis his).

Here we can see a conflict arising between a motive to take a moral action and being motivated to take the easier and/or less distressing route of disengagement (since it can be an effective means of moral identity maintenance). A similar conflict could arise between engaging in an act of moral cleansing behavior (i.e., an easily achievable performance goal that allows one to re-establish certainty in oneself as essentially morally good), and engaging in the more effortful task of working on self-improvement (i.e., in the sense of having a learning goal that motivates you to put effort into doing better next time). Given the similarities between an essentialized view of the true self and fixed mindsets, there are reasons to be worried that maintaining the moral identity of the true self as essentially good could provide motivation for moral disengagement or merely short-term moral cleansing behavior (at the expense of long-term moral improvement).

Conclusion

If the view of the true self as morally good could prompt maladaptive responses to moral failure, such that one avoids moral self-improvement, what might be done about it? I suspect it will be unlikely to prove helpful to try to undermine the fixed belief in one's own essential moral goodness. I offer here at least two practical considerations. First, given how common the tendency appears for people to believe in this view of the true self, it is probably futile to try to get rid of it. Second, people will probably not appreciate it if you try to convince them they're not essentially morally good and will respond with further distress and resistance (in contrast to how someone might appreciate being able to let go of a fixed view of low intelligence).

Instead, I suggest working with the idea that people have some essential drive to be morally good, while emphasizing that work needs to be done to realize that with some reliability in practice. How might this be accomplished? One speculative possibility arises from an observation that one way in which the belief in the 'true self' as essentially morally good differs from other fixed abilities, like intelligence, is that it doesn't imply a limitation to one's abilities. By contrast, a fixed view of intelligence implies that one's level of intelligence is a limit on what one can do. For goodness, though, it seems more of a fixed (i.e., unchanging) capacity for goodness, and this would not necessarily admit of a limit on behavior (as one can always do more good). A related point is that people recognize a distinction between one's overt behavior and the essential goodness of the true self. When people act wrongly, they are thought to be departing from their true self. Even cynics, who might expect the worst behavior from people, still consider people's true self to be good (i.e., they retain the essential capacity for moral goodness). So, one can try to be more in touch with their true self by doing good deeds (though there remains the worry that such deeds are carried out merely as performance goals, rather than as genuine attempts to improve). By contrast, there is no similar story for a fixed view of intelligence—that is, people don't tend to see poor intellectual performances as merely someone departing from their inner genius.

In this sense, while the belief in the true self remains a fixed mindset about one's moral capacities, it goes along with a perspective that people's overt behavior is not fixed. Insofar as

the connection between the true self and moral goodness focuses on one's essential capacity for morally good behavior, this is consistent with the idea that it can also take work to express that goodness and 'stay true to oneself' (at least on a consistent basis). This latter idea is hopeful for virtue theorists, as it's an avenue for promoting moral development in tandem with an essentialized view of moral goodness. The belief in essential goodness might then be harnessed by reminding people about the practice that is involved in consistently expressing that goodness (Stichter, 2018). This approach would not seek to replace the fixed mindset people have about their true self, but rather switch the focus to whether their behavior is expressing that moral goodness. For the virtue theorist, this could be done by focusing on particular virtues that people could work to improve in order to stay more true to themselves. This way moral failures might be interpreted as signaling that only part of one's overall moral behavior needs improvement, rather than the more distressing prospect of having called into question one's whole identity as essentially morally good (Stichter, 2020).

In this case, though, the 'true self' wouldn't function exactly as an aspirational ideal in the way in which the virtuous person does for virtue theory, because there's still an assumed goodness to begin with (else it would not be 'essential' moral goodness). Furthermore, this assumed goodness likely retains its potential for promoting defensiveness, especially for those who don't have a good idea of how to better express that goodness. While people may share a similar belief in their own essential moral goodness, it's clear that not everyone reacts the same way to moral failure. So while there appears to be universality in the belief of the true self as essentially morally good, what may be predictive in regards to how individuals respond differently to moral failure are their strategies for getting in 'touch' with their essential goodness (and there may be further differences depending on whether these strategies are aiming at performance or mastery goals). But more research is needed here to account for individual differences, and to find moderators between failure and responses (whether resulting in defensiveness, temporary good deeds, or long term improvement).

Notes

1. While these two forms of identity are distinct, Lefebvre and Krettenauer (2020) go on to note: 'It is plausible that our early-forming intuitions about true selves may be an ontogenetic precursor of moral identity in the Eriksonian sense' (p. 13). Though in terms of motivating virtuous behavior, while a belief in the true self as essentially morally good might motivate one to take on moral commitments (including acquiring virtues), it would then be moral identity in the developmental (or Eriksonian) sense that has the more direct motivational impact in carrying out moral behavior on a regular basis. In virtue theory, the acquisition of virtue would more directly relate to this developmental sense of identity, in terms of having 'particular commitments to particular ways of being' that provide a 'sense of unity and coherence'.
2. While it is beyond the scope of this paper to engage with the larger moral identity literature, see Krettenauer (2020) where he discusses multiple accounts of moral identity in the psychological literature, and advances a view of moral identity 'as a context-specific adaptation and goal-orientation' (p. 2). This is based on his own work with self-determination theory (SDT), for 'the intermediate level of context-specific adaptations and goal-orientations is most akin to SDT as context-dependent personal goals play a pivotal role in this theory' (Krettenauer, 2020, p. 3). This also happens to match well with my preferred account of virtue, which is based on goal constructs and self-regulation theory (Stichter 2018). Further parallels can be seen in Lefebvre and Krettenauer's (2020) claim that this

developmental sense of moral identity relates to ‘the integration of self-concerns with moral commitments’ (p. 13). This kind of integration would connect to the importance virtue theory places on exercising practical wisdom (i.e., *phronesis*) and integrating one’s activities with a view to what it is to live well (i.e., *Eudaimonia*).

3. Krettenauer (2019) also suggests an overlap between the two forms of moral identity with the two distinctions found in Dweck’s work, and that the true self as essentially good seems to map on to Dweck’s ‘fixed’ mindset or ‘entity’ implicit theory. Though it’s less clear that moral identity in the developmental (or Eriksonian) sense maps onto a malleable mindset, as personal commitments are not abilities per se. But virtues as acquired excellences could bridge that gap, being both a developed ability and a moral commitment, especially since virtue possession (like skill) is a matter of degree (Stichter 2018).
4. I agree with Krettenauer (2019) in thinking that there are likely two different goals associated with these two forms of moral identity. Moral identity in the developmental sense gives us moral goals to aspire to, whereas the goal associated with moral identity in the sense of the true self seems primarily to be to maintain the view of oneself as essentially morally good (or as he puts it, not being immoral). This would be another reason why the influence of the true self on moral behavior might reveal itself most in response to failure, insofar as the goal of preserving one’s moral identity as essentially good is easy to maintain except when one looks to have acted immorally.

Acknowledgments

My thanks to Matthew Vess, Rebecca Schlegel, Joshua Hicks, Joe Maffly-Kipp, and Patricia Flanagan for helpful comments on earlier drafts. I also thank Tobias Krettenauer for making valuable editorial suggestions.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Matt Stichter is an Associate Professor of Philosophy in the School of Politics, Philosophy, & Public Affairs at Washington State University. He pursues research at the intersection of moral psychology, virtue ethics, and the philosophy of expertise. He has published extensively on the ‘virtue as skill’ thesis, arguing that the development of virtue should be understood as a process of skill acquisition, and he draws on the psychological research on self-regulation and expertise to formulate this thesis. He recently published a book on this topic, *The Skillfulness of Virtue: Improving our Moral and Epistemic Lives*, with Cambridge University Press (2018).

ORCID

Matt Stichter  <http://orcid.org/0000-0002-0342-5587>

References

- Aquino, K. (2017). In defense of (a little) moral hypocrisy. In C. B. Miller & W. Sinnott-Armstrong (Eds.), *Moral psychology, Volume 5: Virtue and character* (pp. 53–62). MIT Press.
- Bandura, A. (2016). *Moral disengagement*. Worth Publishers.

- Bastian, B., & Haslam, N. (2006). Psychological essentialism and stereotype endorsement. *Journal of Experimental Social Psychology, 42*(2), 228–235. <https://doi.org/10.1016/j.jesp.2005.03.003>
- Batson, C. D. (2017a). Getting cynical about character: A social-psychological perspective. In C. B. Miller & W. Sinnott-Armstrong (Eds.), *Moral psychology, Volume 5: Virtue and character* (pp. 11–44). MIT Press.
- Batson, C. D. (2017b). Help thou my unbelief: A reply to May and Aquino. In C. B. Miller & W. Sinnott-Armstrong (Eds.), *Moral psychology, Volume 5: Virtue and character* (pp. 63–74). MIT Press.
- Bench, S. W., Schlegel, R. J., Davis, W. E., & Vess, M. (2015). Thinking about change in the self and others: The role of self-discovery metaphors and the true self. *Social Cognition, 33*(3), 169–185. <https://doi.org/10.1521/soco.2015.33.3.2>
- Christy, A. G., Schlegel, R. J., & Cimpian, A. (2019). Why do people believe in a “true self”? The role of essentialist reasoning about personal identity and the self. *Journal of Personality and Social Psychology, 117*(2), 386–416. <https://doi.org/10.1037/pspp0000254>
- Christy, A. G., Seto, E., Schlegel, R. J., Vess, M., & Hicks, J. A. (2016). Straying from the righteous path and from ourselves. *Personality & Social Psychology Bulletin, 42*(11), 1–13. <https://doi.org/10.1177/0146167216665095>
- De Freitas, J., Sarkissian, H., Newman, G. E., Grossmann, I., De Brigard, F., Luco, A., & Knobe, J. (2018). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive Science, 42*(S1), 134–160. <https://doi.org/10.1111/cogs.12505>
- De Freitas, J., Tobia, K. P., Newman, G. E., & Knobe, J. (2017). Normative judgments and individual essence. *Cognitive Science, 41*(S3), 382–402. <https://doi.org/10.1111/cogs.12364>
- Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development*. Psychology Press.
- Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*(2), 256–273. <https://doi.org/10.1037/0033-295X.95.2.256>
- Effron, D. A. (2014). Making mountains of morality from molehills of virtue: Threat causes people to overestimate their moral credentials. *Personality & Social Psychology Bulletin, 40*(8), 972–985. <https://doi.org/10.1177/0146167214533131>
- Erikson, E. H. (1959). *Identity and the life cycle: Selected papers*. International Universities Press.
- Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the moral self: The effects of recalling past moral actions on future moral behavior. *Personality & Social Psychology Bulletin, 37*(5), 701–713. <https://doi.org/10.1177/0146167211400208>
- Krettenauer, T. (2019, November 7–9). Moral identity as a dual goal-orientation. *45th Annual Meeting of the Association for Moral Education*, Seattle.
- Krettenauer, T. (2020). Moral identity as a goal of moral action: A self-determination theory perspective. *Journal of Moral Education, 49*(3), 330–345. <https://doi.org/10.1080/03057240.2019.1698414>
- Lapsley, D. (2016). Moral self-identity and the social-cognitive theory of virtue. In J. Annas, D. Narvaez, & N. Snow (Eds.), *Developing the virtues: Integrating perspectives* (pp. 36–68). Oxford University Press.
- Lefebvre, J. P., & Krettenauer, T. (2020). Is the true self truly moral? Identity intuitions across domains of sociomoral reasoning and age. *Journal of Experimental Child Psychology, 192*(April 2020), 104769. <https://doi.org/10.1016/j.jecp.2019.104769>
- Mullen, E., & Monin, B. (2016). Consistency versus licensing effects of past moral behavior. *Annual Review of Psychology, 67*(1), 363–385. <https://doi.org/10.1146/annurev-psych-010213-115120>
- Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgments and the true self. *Personality & Social Psychology Bulletin, 40*(2), 203–216. <https://doi.org/10.1177/0146167213508791>
- Stichter, M. (2018). *The skillfulness of virtue: Improving our moral and epistemic lives*. Cambridge University Press.
- Stichter, M. (2020). Learning from failure: Shame and emotion regulation in virtue as skill. *Ethical Theory and Moral Practice, 23*(2), Special Issue on Virtue Ethics and Moral Psychology, 341–354. <https://doi.org/10.1007/s10677-020-10079-y>

- Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*, 12(4), 551–560. <https://doi.org/10.1177/1745691616689495>
- Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, 131(1), 159–171. <https://doi.org/10.1016/j.cognition.2013.12.005>
- Strohminger, N., & Nichols, S. (2015). Neurodegeneration and Identity. *Psychological Science*, 26(9), 1469–1479. <https://doi.org/10.1177/0956797615592381>