In press: *Becoming Someone New: Essays on Transformative Experience, Choice, and Change*

# Learning from moral failure

Matthew Cashman[1] & Fiery Cushman[2]

[2] *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology*

[1]*Department of Psychology, Harvard University*

*Pedagogical environments are often designed to minimize the chance of people acting wrongly; surely this is a sensible approach. But could it ever be useful to design pedagogical environments to permit, or even encourage, moral failure? If so, what are the circumstances where moral failure can be beneficial? What types of moral failure are helpful for learning, and by what mechanisms? We consider the possibility that moral failure can be an especially effective tool in fostering learning. We also consider the obvious costs and potential risks of allowing or fostering moral failure. We conclude by suggesting research directions that would help to establish whether, when and how moral pedagogy might be facilitated by letting students learn from moral failure.*

## 1. Introduction

Military basic training is organized around a simple goal: Within a few months, train an ordinary 18-year-old to follow orders that include risking their own life and ending others'. To accomplish so much change in so little time would seem laughably ambitious if it didn't work so well.

The recruit must learn new motor skills (e.g., how to fire and maintain a rifle), new semantic knowledge (e.g., rules of engagement), and new social orders (e.g., the chain of command). The most remarkable transformations, however, are moral. Recruits are asked to subordinate personal wellbeing to that of the group; to supplant individuality agency with collective allegiance and hierarchy; to regard harm not as intrinsically wrong but as instrumentally justified. How is this done?

We focus on one small but exquisitely counterintuitive piece of the puzzle: Sometimes, students are set up to fail. That is, instructors occasionally create situations where recruits are unlikely to succeed at a stated objective, and where their failure has moral consequences subject to corrective action by the instructor (Heckathorn, 1988; Gilham, 1982). The recruits' failures are often on simple tasks: uniform not in order, bed not made correctly, late to parade. These are tasks that, individually, are well within the capacity of any recruit. Taken collectively—in the context of little food, reduced sleep, copious exercise, and stress—the likelihood of somebody in the squad failing increases. Because these tasks are simple, recruits are prone to attribute failure to some personal defect or deliberate choice; after all, who can't make a bed? Moreover, what would otherwise be non-moral failures become moral failures when punishment is applied to the recruit's entire group. This invites other-oriented emotions such as guilt and shame—a feeling of having "let down my squad".

In other words, basic training seeks to change the moral character of recruits in part by designing situations in which moral failure is likely. An implicit assumption is that the lessons learned from experiencing failure, or witnessing it, are especially powerful and enduring. Our goal is to scrutinize this assumption. We ask whether, and when, moral failure can be a productive element of moral education.

In its most essential form, moral failure has two parts: it is a (1) *direct experience* of (2) *subjectively failing* to meet moral standards. Forgetting your child's birthday, for instance, would likely be experienced by most people as a moral failure. Of course, much (perhaps most) moral learning does not involve moral failure. First, not all forms of moral learning involve direct experience; an alternative form of moral learning involves exposure to abstract rules (e.g., exposure to the ten commandments). Second, not all forms of learning from direct experience involve failure; an alternative form of learning involves practicing and experiencing success (e.g., cultivating the habit of empathy through meditation).

There are also useful and important ways of relaxing this strict definition of moral failure. If a person reads a novel in which the protagonist experiences moral failure, this may prompt a similar psychological response to personally experiencing moral failure even though it is not "direct experience" of the reader's. Or, if a person performs an action they sincerely believe to be right but notices that everybody else considers it wrong, their subjective experience may be a moral failure of some kind, even if not of the purest kind. We will therefore consider not only prototypical moral failures, but also their many kin.

Here, we build a case for the value of moral failure by drawing on several circumstantial lines of evidence: Direct experience is a privileged form of learning; Morality requires representations of what *not* to do (not just what *to* do); Guilt is an adaptive response to moral failure that facilitates reparation and learning; and, child development appears to involve periods designed to "test limits" in a way that will reliably lead to moral failure. Yet there is little direct research on the practical value of moral failure, and several reasons to doubt whether it is especially effective, or whether its obvious costs could appropriately outweigh its potential benefits. We therefore conclude by considering opportunities for further research.

## 2. Direct experience is a privileged form of learning

There are some things you have to try in order to learn. You can't read a book to learn how to water ski, and you can't pick up the guitar without touching one. Of course, many other things can be learned perfectly well from books, lectures and the like—things like who won the World Series in 1927. Semantic knowledge of this form rarely depends on experience: To know who won the World Series, it isn't at all important that you attended the game. This reflects the fact that human memory and learning comprises multiple distinct systems. Procedural memories of how to execute the small, finely-tuned motions necessary to stay upright on water skis are gleaned from practice, episodic memories supply the broad strokes (such as how to attach the tow rope and put on the skis), and semantic memory may supply as-yet unused hand-signals (such as for an emergency stop).

At first blush, morality would seem to depend principally upon semantic memory. To understand that it is wrong to kill, isn't it sufficient just to be told so? Surely it isn't necessary to actually kill a person oneself?

We propose that moral knowledge occupies a point somewhere in between waterskiing and knowing who won the world series: Although having a direct, personal experience of the consequences of an action are certainly not necessary to represent that it is right the wrong, the *manner* in which you know it and the *strength* of your conviction may depend upon experience. Perhaps a person who has killed can experience its wrongness in a way that is difficult—even impossible—for others to fully understand.

The key concept that connects procedural learning to moral behavior is *value*. Several lines of evidence indicate that human moral behavior depends upon representations of value (Crockett 2013; Cushman

2013; Ruff 2014). The way that people make moral tradeoffs exhibits signatures of general value-guided decision-making mechanisms, such as diminishing marginal returns (the difference between saving 0 and 1 lives feels more profound than the difference between saving 100 and 101 (Shenhav & Greene 2010). Neural systems implicated in learning and representing value are reliably recruited during moral judgment and decision-making (reviewed in Ruff 2014). The disruption of these systems by injury can lead to disorders of moral judgment and behavior (Koenigs, Young et al 2007; Damasio 1994; Lough et al 2006; Darby et al 2017), and the same systems appear to be dysfunctional in psychopaths (Buckholtz et al 2010; Finger et al 2008).

Meanwhile, much procedural learning depends upon value representation. Computational models of value-guided learning and decision-making provide an excellent fit to behavioral and neural evidence on forms of procedural learning, including habitual action and thought (Dolan & Dayan 2013; Daw & Shohamy 2008; Glimcher 2011). Experience is the canonical path by which value is learned. Although it is certainly possible to learn value in other ways—for instance, by observing other people, being instructed by other people, or imagining alternative possibilities (Gershman et al 2014; Olsson & Phelps 2007; Ho et al 2015; Ho et al 2016)—direct personal experience appears to be a privileged form of value learning (Paul 2014; Olsson & Phelps 2007). Indeed, there are many domains where learning without experience, as by reading or instruction, will be ineffective. Similarly, there are situations where reading a book or attending class will result in some useful amount of learning—but this is strengthened by practice (Kolb 2014; Boud 1993), and many instances where simulations are used in pedagogy to access experiential learning that would otherwise be impractical (Ruben 1999).

Motivated by this pair of observations—morality depends on value representation, and direct experience is a privileged form of learning value—several recent theories propose that moral behavior is strongly influenced by implicit value representations, often learned through direct experience (e.g., Crockett 2013; Cushman 2013; Rand et al 2014). Although recent years have seen renewed interest in this idea, its historical roots lie at least as deep as Aristotle (Ethics 1166b5–29). When applied to the concept of moral failure, the implication is that sometimes *actually* failing will engage an especially powerful form of learning.

## 3. You have to learn the "don'ts": Proscriptive morality is unique

If moral behavior depends upon value representations that can be acquired through experience and practice (something roughly like learning to water ski) then the most obvious implication for moral education would seem to be to practice doing the *right* thing—precisely the opposite of moral failure. Typically, we assume the hard work of acquiring a new skill consists in discovering what *to do*, rather than discovering what *not* to do. It seems as if learning what *not* to do should come for free—just learn to do what's right, and you'll never even think about doing wrong.

Yet, in contrast to this picture, people seem to have dissociable systems for learning what to do and what not to do. These are sometime described as "appetitive" versus "aversive" learning systems, and sometimes in other terms (Carver 2001; Frank et al 2004). Several lines of evidence indicate the significance of this distinction. They can be broadly, if not perfectly distinguished neuroanatomically. For instance, the striatum plays a more essential role in appetitive responding and the amygdala in aversive responding (LeDoux 2003; but see Seymour et al 2005), although both structures clearly participate in both, and may be best characterized in terms of different learning rules (Li et al 2011). Likewise, neurochemically, dopamine seems to play a more essential role in appetitive learning and serotonin in aversive learning (Crockett et al 2009; Cools et al 2011; Boureau & Dayan 2011), although again the functions of each neurotransmitter are varied and overlapping. These findings indicate that the human mind does not respond to all motivational influences alike; rather, it respects a rough divide between systems that regulate behavior around "promotion" (i.e., appetite) and "prevention" (i.e., aversion).

This division is also clearly reflected in moral judgment and behavior (Janoff-Bulman et al 2009; Crockett et al 2015). That is, people appear to have distinct representations of what is morally required (or at least laudable) and what is morally prohibited (or at least blameworthy). In theory, then, a person could adequately learn to do what is morally good and yet not have learned to avoid what is morally bad.

This implies that even if a person has experienced the rewards doing the right thing, there is a potential for additional learning by experiencing the consequences of doing wrong. For now, we set aside the vital question of how such pedagogical "value" could ever outweigh the obvious cost of having acted wrongly. Instead, our next goal is to understand more clearly the specialized psychological mechanisms that help us to learn from moral failure.

## 4. Guilt facilitates learning from failure

Humans are designed for corrective learning from moral failure, and a linchpin of this design is guilt. Put simply, if you feel guilty about something you did, you're less likely to do it again (Monteith et al 1993; Mosher 1965; Monteith et al 2002; Baumeister et al 1995). This does not imply that guilt is the best path to moral learning, and it is certainly not the only path. But, insofar as guilt is triggered by episodes of self-perceived moral failure, it tends to improve future behavior.

Unfortunately, adaptive moral learning via guilt is not the only possible outcome of perceived moral failure. In opposition to this desired pathway, there is a parallel undesired pathway from subjective moral failure to social withdrawal and "externalization"—the attribution of blame for the moral failure to external circumstances, rather than the acknowledgment of personal responsibility (Figure 1; Leach 2015). (Incidentally, guilt establishes the motive to compensate harms and repair social relationships (reviewed in Bybee, Merisca & Velasco 1998); although this is not the goal of the desired pathway from moral failure to moral learning, neither is it detrimental to that goal).

The effort to delineate adaptive and maladaptive responses to moral failure centers on the distinction between "guilt" and "shame". Guilt is typically regarded as a more adaptive response (characterized by learning and reparation), and shame as a less adaptive response (characterized by externalization and withdrawal). We do not mean to imply that either is less biologically or culturally adaptive; likely, both of these mechanisms are "adaptive" in those senses. Rather, we mean adaptive from a social and pedagogical perspective: if you were responsible for somebody's moral education, you would probably wish them to respond to failure with feelings of guilt rather than shame.

It is also important to clarify our use of the very terms "guilt" and "shame", and the status of the distinction we are drawing between them. We do not intend to analyze the folk concept of guilt versus shame, or an attempt to understand what these words mean in ordinary usage. In fact, there is very little difference in the way that ordinary people understand or apply the words "guilt" and "shame" (Leach 2015). Rather, these have become terms of art in a literature demonstrating that self-perceived moral failures can lead to a series of psychological and behavioral reactions that roughly follow two rival paths. This literature seeks to understand how these paths differ: What determines which path is followed, and where each is likely to lead, both personally and socially. From this perspective, the distinction might as well be described as "Type 1" versus "Type 2" reactions, as opposed to "guilt" and "shame".

Guilt and shame may be widely agreed to be moral emotions and the principal emotions arising from moral failure, but the exact distinction between the two is the subject of some debate (e.g. Tangney 2002, Tangney et. al. 2007, Gilbert 1994, Lindsay-Hartz 1995). There is disagreement about what constitutes guilt vs. shame, their properties, and their relative usefulness. Most commonly guilt is viewed as the "better" emotion because it results in approach behaviors helpful to the agent and the aggrieved, whereas shame is the "worse" emotion that leads to unhelpful withdrawal behaviors (Nelissen 2013). On some views, guilt and shame are distinguished primarily by their sources (guilt being internal norm violation, shame external), while other, more recent work differentiates based on the object of the emotion (guilt focuses on what has been done, whereas shame focuses on who has done it; Nelissen 2013).

Here, we use "guilt" to describe an emotion elicited by moral failure that focuses on the act rather than qualities of the person who may have caused it, and which generally results in approach behaviors such as reparation. We use "shame" to describe an emotion that is elicited by moral failures attributed to a defect in the self, which focuses on self-image, and which can lead to either approach or withdrawal behaviors. We therefore treat the emotions that arise from moral failure as organized along a spectrum from those concerned with the act (focused on what has been done: guilt) to those concerned with the agent (focused on who has done it: shame).

How, then, does moral failure lead either to useful motivations (e.g., pro-social or self-improvement motivations), or instead to undesirable ones (e.g., self-defensive motivations)?

Moral failures that are focused on the act rather than the agent ("I can't believe I did that, throwing my recyclables in the trash"), and which are not indicative of some sort of self-defect, can elicit adaptive guilt and therefore prosocial behaviors aimed at self-improvement. Moral failures that are focused on the agent ("I can't believe I was late again, I'm always late") and which are attributed to a specific (likely mutable) self-defect, generally motivate self-improvement via shame. Moral failures that are focused on the agent, and which lead to appraisal of global (seemingly immutable) self-defect ("I caused that crash because I am an alcoholic") engender a type of shame which leads to feelings of inferiority and self-defensive motivations to hide, avoid and externalize (Gausel & Leach, 2011).

Given this framework, we can start to identify a set of moral failures expected to be useful in pedagogy: those generally resulting in focus on the act itself or those which are attributed to a specific self-defect. Before turning to the practical implications of this view, however, we consider one final piece of evidence that humans learn especially well from the experience of moral failure.

## 5. Children may be designed to fail

Not only are humans designed to learn from failure, there is good reason to believe that they are actually designed to fail. Colloquially, we speak of children "testing limits". This behavior is especially pronounced during the toddler and preschool years, and then again during adolescence. To describe child misbehavior as "testing limits" implies that the principle aim of the misbehavior is moral learning. For instance, consider a child who takes a cookie from a jar and begins to eat it in view of her father. Some potential explanations for this behavior are (1) she has not considered the possibility that this behavior is disallowed, or (2) she knows it is disallowed, but doesn't care because she really wants a cookie. But if she is truly testing limits, then an additional explanation is (3) she is unsure whether the behavior is disallowed, or exactly what is consequences will be, and an effective way to learn is to try it. An additional and likely possibility is that (1) or (2) may characterize her proximate psychological motives, while (3) characterizes the ultimate adaptive rationale that explains them.

Of course, children are capable of learning a new rule without first violating it. A child uncertain of the rule could just ask an authority ("Daddy, would it be OK for me to take a cookie right now?", on the questionable assumption that her father has any relevant authority). Or, she could observe others' behavior (Bandura & Walters 1977). Either of these strategies would presumably avoid likely costs of limit-testing, such as punishment or reputational harm. Why don't children rely exclusively on these less fraught methods? Possibly because personal experience of moral failure is an especially effective or powerful form of learning.

Precisely this argument has been made to explain adolescent limit-testing. It is argued that adolescents test the limits set around them in order to acquire a visceral understanding of the consequences of violating those limits—despite already having an abstract understanding of those consequences. For instance, they may know that insulting a friend could do lasting harm to a relationship, and yet not have the "feel" for it. Once this feel is acquired, risk-taking is reduced (Baird 2008; Rivers 2009).

This rationale for limit-testing in early childhood is less discussed in the literature. Nevertheless, this period of development is characterized by frequent conflicts between parents and children over misbehavior. As reviewed by Dahl and Killen (2017), "Naturalistic studies have found that conflicts about prohibited behaviors can occur 10 or more times *per hour* in the second year (Dahl, 2016b; Kuczynski, Kochanska, Radke- Yarrow, & Girnius-Brown, 1987; Power & Parke, 1986)" (emphasis added). Although it is possible that this is a remarkable design failure in the moral lives of young children, a more plausible conclusion is that the failure is, in fact, part of the design.

Finally, a distinct benefit of limit-testing behavior may be the value of learning to recover from moral failures. While learning the rules that govern the world and the costs of breaking them (punishment and guilt/shame) is probably the largest portion of the benefit, there is additional value to be had from an ability to minimize costs once the violation has occurred.

## 6. Can we engineer adaptive failure?

When moral failure happens, it can promote moral learning. But this alone does not imply that we should *create* opportunities for moral failure. There are obvious and weighty risks associated with nudging someone towards immoral actions. Still, our analysis offers some potential strategies to mitigate those risks. We next describe several necessary conditions for achieving adaptive moral—ingredients that could, in theory, transform lemons into lemonade.

### *6.1 Highlight failure*

For the experience of moral failure to promote future moral behavior, it is necessary that the failure is personally acknowledged as such. And, indeed, people often do feel guilty without any prompting. In such cases they may be reinforcing existing moral attitudes. Yet, a variety of self-protective motivations bias people away from acknowledging their own moral failures, even to themselves (Kunda 1990). Thus, one key ingredient to produce useful moral learning from failure is to help prompt individuals to recognize that they have failed.

This may be especially for children, who are still building a basic set of internalized moral norms: Some form of external feedback could help to ensure that they encode their behavior as a moral failure, when appropriate. This could take several different forms, ranging from explicit censure to a more subtle, public reminder of the relevant rule.

### *6.2 Attach failure to acts, not people*

Guilt tends to produce more desirable learning outcomes than shame. And, among the varieties of shame, a focus on specific self-defect is associated with more desirable learning outcomes than a focus on general self-defect. Thus, moral learning from failure will work best when the source of failure is not perceived by wrongdoers as an essential property of their selves. Situations where the wrongdoer does not perceive the failure to stem from an immutable property of themselves are *recoverable*: the agent believes subsequent changes to behavior based on learning from the failure are possible, and that reasonable observers would update their expectations—would be willing to forgive—based on a change in behavior.

### *6.3 Aim for moderate failures*

Failures can be small or large, and ideally should be neither. It is obvious that a failure could be so minor that it prompts no learning. At the opposite extreme, however, a major moral failure may motivate withdrawal behavior, even if the failure is only weakly indicative of the agent's character. For instance, being late frequently may only be moderately indicative of an agent's character, but being late to one's own wedding may be unrecoverable.

### *6.4 Pedagogy in practice*

How could moral failure be useful in pedagogical contexts? Our discussion is speculative, and certainly not prescriptive. As we have already said, there is tremendous uncertainty regarding how, when and whether the potential benefits of moral failure could ever outweigh its costs. To be clear: Although are discussing the possibility of creating opportunities for moral failure, we do not endorse it.

The most extreme possibility we consider is that teachers could deliberately create a situation in which they hope to prompt a pedagogically productive form of actual moral failure. This, for instance, appears to be the approach adopted in at least some elements of military basic training. An active intervention could take many different forms, such as exercises that guarantee failure in one or more pupils or changes that merely provide the opportunity for pupils to fail morally.

Passive methods offer a milder approach. Many social environments are engineered to prevent any possibility of moral failure as far as can be achieved. Schools have zero-tolerance policies, parents maximize safety and minimize hurt to others, and companies harshly punish even small deviations from safety protocols. Thus, moral failures could be prompted not by intervening to cause moral failure, but instead by relaxing current restrictions currently in order to allow it to arise naturally. For instance, passive intervention in a classroom of young children might involve eliminating the prohibition of "roughhousing" styles of play. Of course, when failures in such a context do occur, it may often require active intervention by the teacher to help a child draw the appropriate lessons, as described above.

A still milder use of moral failure is to draw lessons from observational learning. This method depends on the assumption that some of the same learning processes engaged when actually failing oneself can also be engaged by observing and considering the failure of somebody else. This might occur when, for instance, observing another person who has failed triggers an act of vivid simulation or imagination of what it would be like to be that person. At a practical level, it implies a curriculum of moral education that complements abstract reasoning with concrete narratives, and narratives of moral exemplars with narratives of moral "counter-exemplars". Additionally, there is the prospect of group shame and guilt, which might lead to positive outcomes. Brown et. Al (2008), Mazziotta et al (2014) and others point out that collective guilt can lead to increased pro-social behavior which suggests that collective guilt may also have a role to play in the active use of moral failure in pedagogy.

Along similar lines, people might learn from exposure to fictional or hypothetical moral failure—for instance, by reading literature or being prompted to imagine how they would feel if they acted wrongly themselves. Here, again, the approach depends on the assumption that imagination furnishes a kind of simulated experience that engages learning mechanisms ordinarily reserved exclusively for direct experience.

### 6.5 *Can the ends justify the means?*

The potential costs of encouraging moral failure are obvious and perhaps insurmountable. To begin with, according to some moral theories, it is never permissible to contribute to one moral wrong in order to prevent others (reviewed in Fischer & Ravizza 1992). Even if such cost-benefit tradeoffs can be permissible, the question remains whether the benefits really *do* outweigh the costs. In this regard, three mechanisms that we have mentioned seem especially promising. The first is to ensure that, when moral failure happens to occur spontaneously, an effort is made to ensure that the subsequent teaching and learning processes that occur maximize the likelihood of productive moral growth: moral failures do happen, and we can design pedagogical environments to take advantage of failures rather than to merely deal with them. The second is to investigate the potential role of mere imagined moral failure (e.g. in the context of reading literature or history) to elicit some of the same adaptive mechanisms that likely accompany actual failure. Finally, there is the prospect of group shame and guilt (Brown et al 2008; Mazzotta et al 2014).

The value of moral failure will also depend greatly on whether lessons learned from it are enduring. There is some cause for optimism: Tangney et. al. (2014) report that prisoners who have feelings of guilt about actions are less likely to re-offend than prisoners who feel less guilt, implying an effect that may last months or years. However, they also report more mixed results with shame. In particular, they report little overall effect of shame on recidivism. This may reflect the grouping of both specific self-defects and global defects into one category of "shame", though there is a significant effect of shame on recidivism when it is accompanied by externalization of blame. Of course, these correlational findings cannot by themselves rule out the possibility that the effect is driven by shame- or guilt-proneness overall in an agent, but it is consistent with mediation by the extent to which an act is indicative character.

## 7. Conclusion

Errors are an important source of learning, and educators often exploit this fact. Failing helps to tune our sense of balance; Newtonian mechanics sticks better when we witness the failure of our folk physics. We consider the possibility that moral failure may also prompt especially strong or distinctive forms of learning. First, and with greatest certainty, humans are designed to learn from moral failure through the feeling of guilt. Second, and more speculatively, humans may be designed to experience moral failures by "testing limits" in a way that ultimately fosters an adaptive moral character. Third—and highly speculatively—there may be ways to harness learning by moral failure in pedagogical contexts. Minimally, this might occur by imagination, observational learning, or the exploitation of spontaneous wrongful acts as "teachable moments".

## References

Agerström, J., Björklund, F., & Carlsson, R. (2012). Emotions in Time: Moral Emotions Appear More Intense with Temporal Distance. Social Cognition, 30(2), 181–198. https://doi.org/http://dx.doi.org/101521soco2012302181

Bandura, A., & Walters, R. H. (1977). Social learning theory(Vol. 1). Englewood Cliffs, NJ: Prentice-hall.

Berger, J., & Jr, M. Z. (2002). New Directions in Contemporary Sociological Theory. Rowman & Littlefield Publishers.

Bolle, F., Breitmoser, Y., Heimel, J., & Vogel, C. (2011). Multiple motives of pro-social behavior: evidence from the solidarity game. Theory and Decision, 72(3), 303–321. https://doi.org/10.1007/s11238-011-9285-0

Boureau, Y. L., & Dayan, P. (2011). Opponency revisited: competition and cooperation between dopamine and serotonin. Neuropsychopharmacology, 36(1), 74.

Brown, R., González, R., Zagefka, H., Manzi, J., & Čehajić, S. (2008). Nuestra culpa: Collective guilt and shame as predictors of reparation for historical wrongdoing. Journal of Personality and Social Psychology, 94(1), 75–90. https://doi.org/10.1037/0022-3514.94.1.75

Buckholtz, J. W., Treadway, M. T., Cowan, R. L., Woodward, N. D., Benning, S. D., Li, R., ... & Smith, C. E. (2010). Mesolimbic dopamine reward system hypersensitivity in individuals with psychopathic traits. Nature neuroscience, 13(4), 419.

Bybee, J. (1997). Guilt and Children. Academic Press.

Chiu, C., Dweck, C. S., Tong, J. Y., & Fu, J. H. (1997). Implicit theories and conceptions of morality. Journal of Personality and Social Psychology, 73(5), 923–940. https://doi.org/10.1037/0022-3514.73.5.923

Cools, R., Nakamura, K., & Daw, N. D. (2011). Serotonin and dopamine: unifying affective, activational, and decision functions. Neuropsychopharmacology, 36(1), 98.

Crockett, M. J. (2013). Models of morality. Trends in cognitive sciences, 17(8), 363-366.

Crockett, M. J., Clark, L., & Robbins, T. W. (2009). Reconciling the role of serotonin in behavioral inhibition and aversion: acute tryptophan depletion abolishes punishment-induced inhibition in humans. Journal of Neuroscience, 29(38), 11993-11999.

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Ousdal, O. T., Story, G., Frieband, C., ... & Dolan, R. J. (2015). Dissociable effects of serotonin and dopamine on the valuation of harm in moral decision making. Current Biology, 25(14), 1852-1859.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. Personality and social psychology review, 17(3), 273-292.

Curtis, A. J. (2013). Tracing the School-to-Prison Pipeline from Zero-Tolerance Policies to Juvenile Justice Dispositions Note. Georgetown Law Journal, 102, 1251–1278.

Damasio, A. R. (1994). Descartes' error. Random House.

Darby, R. R., Horn, A., Cushman, F., & Fox, M. D. (2017). Lesion network localization of criminal behavior. Proceedings of the National Academy of Sciences, 201706587.

David, B., Ruth, C., & David, W. (1993). Using Experience For Learning. McGraw-Hill Education (UK).

Daw, N. D., & Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. Social Cognition, 26(5), 593-620.

de Hooge, I. E., A, M., Breugelmans, S. M., & Zeelenberg, M. (2011). What is moral about guilt? Acting "prosocially" at the disadvantage of others. Journal of Personality and Social Psychology, 100(3), 462–473. https://doi.org/10.1037/a0021459

Dienstbier, R. A., Hillman, D., Lehnhoff, J., Hillman, J., & Valkenaar, M. C. (1975). An emotion-attribution approach to moral behavior: Interfacing cognitive and avoidance theories of moral development. Psychological Review, 82(4), 299–315. https://doi.org/10.1037/h0076826

Dienstbier, R. A., & Munter, P. O. (1971). Cheating as a function of the labeling of natural arousal. Journal of Personality and Social Psychology, 17(2), 208–213. https://doi.org/10.1037/h0030384

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. Neuron, 80(2), 312-325.

Dweck, C. S. (2008). Can personality be changed? The role of beliefs in personality and change. Current Directions in Psychological Science, 17(6), 391–394.

Finger, E. C., Marsh, A. A., Mitchell, D. G., Reid, M. E., Sims, C., Budhani, S., ... & Pine, D. S. (2008). Abnormal ventromedial prefrontal cortex function in children with psychopathic traits during reversal learning. Archives of general psychiatry, 65(5), 586-594.

Fischer, J. M., & Ravizza, M. (1992). Ethics Problems and Principles.

Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science, 306(5703), 1940-1943.

Gausel, N., & Leach, C. W. (2011). Concern for self-image and social image in the management of moral failure: Rethinking shame. European Journal of Social Psychology, 41(4), 468–478. https://doi.org/10.1002/ejsp.803

Gausel, N., Leach, C. W., Vignoles, V. L., & Brown, R. (2012). Defend or repair? Explaining responses to in-group moral failure by disentangling feelings of shame, rejection, and inferiority. Journal of Personality and Social Psychology, 102(5), 941–960. https://doi.org/10.1037/a0027233

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. Journal of Experimental Psychology: General, 143(1), 182.

Gilbert, P., Pehl, J., & Allan, S. (1994). The phenomenology of shame and guilt: An empirical investigation. British Journal of Medical Psychology, 67(1), 23–36. https://doi.org/10.1111/j.2044-8341.1994.tb01768.x

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proceedings of the National Academy of Sciences, 108(Supplement 3), 15647-15654.

Haslam, N. (2004). Essentialist Beliefs about Personality and Their Implications. Personality and Social Psychology Bulletin, 30(12), 1661–1673. https://doi.org/10.1177/0146167204271182

Heckathorn, D. D. (1988). Collective sanctions and the creation of prisoner's dilemma norms. American Journal of Sociology, 535–562.

Hooge, I. de. (2014). The general sociometer shame: Positive interpersonal consequences of an ugly emotion. Retrieved from http://repub.eur.nl/pub/51671/

Ho, M. K., Littman, M. L., Cushman, F., & Austerweil, J. L. (2015). Teaching with rewards and punishments: Reinforcement or communication?. In CogSci.

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. In Advances In Neural Information Processing Systems (pp. 3027-3035).

Hooge, I. E. de, Zeelenberg, M., & Breugelmans, S. M. (2007). Moral sentiments and cooperation: Differential influences of shame and guilt. Cognition and Emotion, 21(5), 1025–1042. https://doi.org/10.1080/02699930600980874

Hooge, I. E. de, Zeelenberg, M., & Breugelmans, S. M. (2010). Restore and protect motivations following shame. Cognition and Emotion, 24(1), 111–127. https://doi.org/10.1080/02699930802584466

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: two faces of moral regulation. Journal of personality and social psychology, 96(3), 521.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. Nature, 446(7138), 908.

Kolb, D. A. (2014). Experiential Learning: Experience as the Source of Learning and Development. FT Press.

Kunda, Z. (1990). The case for motivated reasoning. Psychological bulletin, 108(3), 480.

Leach, C. W., & Cidam, A. (2015). When is shame linked to constructive approach orientation? A meta-analysis. Journal of Personality and Social Psychology, 109(6), 983–1002. https://doi.org/10.1037/pspa0000037

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. Nature neuroscience, 14(10), 1250.

Lickel, B., Kushlev, K., Savalei, V., Matta, S., & Schmader, T. (2014). Shame and the motivation to change the self. Emotion, 14(6), 1049–1061. https://doi.org/10.1037/a0038235

Lindsay-Hartz, J., de Rivera, J., & Mascolo, M. F. (1995). Differentiating guilt and shame and their effects on motivation. In J. P. Tangney & K. W. Fischer (Eds.), Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride (pp. 274–300). New York, NY, US: Guilford Press.

Lough, S., Kipps, C. M., Treise, C., Watson, P., Blair, J. R., & Hodges, J. R. (2006). Social reasoning, emotion and empathy in frontotemporal dementia. Neuropsychologia, 44(6), 950-958.

Lukes, S. (1965). Moral Weakness. The Philosophical Quarterly (1950-), 15(59), 104–114. https://doi.org/10.2307/2218210

Mazziotta, A., Feuchte, F., Gausel, N., & Nadler, A. (2014). Does remembering past ingroup harmdoing promote postwar cross-group contact? Insights from a field-experiment in Liberia. European Journal of Social Psychology, 44(1), 43–52. https://doi.org/10.1002/ejsp.1986

Merrill, J., & Gross, A. E. (1969). Some effects of guilt on compliance. Journal of Personality and Social Psychology, 11(3), 232–239. https://doi.org/10.1037/h0027039

Murphy, J. B. (2015). Does Habit Interference Explain Moral Failure? Review of Philosophy and Psychology, 6(2), 255–273. https://doi.org/10.1007/s13164-014-0220-5

Nelissen, R. M. A., Breugelmans, S. M., & Zeelenberg, M. (2013). Reappraising the Moral Nature of Emotions in Decision Making: The Case of Shame and Guilt. Social and Personality Psychology Compass, 7(6), 355–365. https://doi.org/10.1111/spc3.12030

Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: the neural systems of social fear transmission. Social Cognitive and Affective Neuroscience. https://doi.org/10.1093/scan/nsm00

Olsson, A., & Phelps, E. A. (2004). Learned Fear of “Unseen” Faces after Pavlovian, Observational, and Instructed Fear. Psychological Science, 15(12), 822–828. https://doi.org/10.1111/j.0956-7976.2004.00762.

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. Nature Neuroscience, 10(9), 1095–1102. https://doi.org/10.1038/nn1968

Pagliaro, S. (2012). On the relevance of morality in social psychology: An introduction to a virtual special issue. European Journal of Social Psychology, 42(4), 400–405. https://doi.org/10.1002/ejsp.1840

Paul, L. A. (2014). Transformative experience. OUP Oxford.

Prentice, D. A., & Miller, D. T. (2007). Psychological essentialism of human categories. Current Directions in Psychological Science, 16(4), 202–206.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. Nature communications, 5, 3677.

Rangel, U., & Keller, J. (2011). Essentialism goes social: Belief in social determinism as a component of psychological essentialism. Journal of Personality and Social Psychology, 100(6), 1056–1078. https://doi.org/10.1037/a0022401

Rees, J. H., Klug, S., & Bamberg, S. (2014). Guilty conscience: motivating pro-environmental behavior by inducing negative moral emotions. Climatic Change, 130(3), 439–452. https://doi.org/10.1007/s10584-014-1278-x

Regan, J. W. (1971). Guilt, perceived injustice, and altruistic behavior. Journal of Personality and Social Psychology, 18(1), 124–132. https://doi.org/10.1037/h0030712

Rivers, S. E., Reyna, V. F., & Mills, B. (2008). Risk taking under the influence: A fuzzy-trace theory of emotion in adolescence. Developmental Review, 28(1), 107–144.

Rosthal, R. (1967). Moral weakness and remorse. Mind, 76(304), 576–579.

Ruben, B. D. (1999). Simulations, Games, and Experience-Based Learning: The Quest for a New Paradigm for Teaching and Learning. Simulation & Gaming, 30(4), 498–505. https://doi.org/10.1177/104687819903000409

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. Nature Reviews Neuroscience, 15(8), 549.

Seymour, B., O'doherty, J. P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., & Dolan, R. (2005). Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. Nature neuroscience, 8(9), 1234.

Sinnott-Armstrong, W. (2005). You Ought to be Ashamed of Yourself  (When you Violate an Imperfect Moral Obligation). Philosophical Issues, 15(1), 193–208. https://doi.org/10.1111/j.1533-6077.2005.00061.x

Sinnott-Armstrong, W. (Ed.). (2008). Moral psychology. Cambridge, Mass: MIT Press.

Tangney, J. P. (1995). Recent Advances in the Empirical Study of Shame and Guilt. The American Behavioral Scientist, 38(8), 1132–1145.

Tangney, J. P., Stuewig, J., & Martinez, A. G. (2014). Two Faces of Shame The Roles of Shame and Guilt in Predicting Recidivism. Psychological Science, 25(3), 799–805. https://doi.org/10.1177/095679761350879

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral Emotions and Moral Behavior. Annual Review of Psychology, 58(1), 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145

Tannenbaum, J. (2006). Emotional expressions of moral value. Philosophical Studies, 132(1), 43–57. https://doi.org/10.1007/s11098-006-9056-x

Tannenbaum, J. (2015). Mere moral failure. Canadian Journal of Philosophy, 45(1), 58–84. https://doi.org/10.1080/00455091.2014.997334

Thero, D. P. (2006). Understanding Moral Weakness. Rodopi.

van der Toorn, J., Ellemers, N., & Doosje, B. (2015). The threat of moral transgression: The impact of group membership and moral opportunity. European Journal of Social Psychology, 45(5), 609–622. https://doi.org/10.1002/ejsp.2119