Moral Reasoning and Social Norms

It is a vexed question whether emotions are the consequence of a rational process or the source of the premises on which rationality operates. It seems that the moral emotions are a form of self-regulation, and possibly constitute a learning process whereby we anticipate and respond to social feedback. It seems further that while the emotional foundations of morality may be consistent across cultures, these moral foundations will find various expressions across different societies.

Jonathan Haidt argues that intuitive moral emotions drive morality, and that moral reasoning is a form of post-hoc justification that we use to convince others (but, crucially, not ourselves). He begins with the observation that human beings are uniquely sensitive to social events that do not directly affect the self, noting our capacity for keeping track of the reputations of hundreds of distinct actors, and he defines moral emotions as those emotions that are linked to the interests and welfare of 1) society as a whole or 2) other people. In general, moral emotions are triggered by "disinterested elicitors," resulting in prosocial action tendencies, and these emotions lead people to care about the world and to support, enforce and improve its integrity. Haidt emphasizes that all of the moral emotions offer indirect or attenuated benefits to the self, and he explores the idea that moral emotions act as commitment devices forcing individuals to follow strategies that maximize long-term benefit. The moral emotions, on this understanding, would be a mechanism to prevent individuals from realizing short-term benefits inimical to long-term success. Haidt contrasts the *homo sapiens* pursuing this strategy with the short-term interests of *homo economicus*, whom he characterizes as a self-interested psychopath.

Haidt discerns four families of moral emotions, which he labels the "other-condemning" family, the "self-conscious" family, the "other-suffering" family and the "other-praising" family. He argues that the first two loom largest in our moral lives. The "other-condemning" family is said to include anger, disgust and contempt. Anger is a response to unjustified insults, combining themes of frustration and goal blockage (simple anger, found in animals and infants) with moral concerns about being betrayed, insulted and treated unfairly. The associated action tendency is revenge, where transgressors are punished proportionately, impartially and in public. Disgust likewise blends a mammalian distaste response (core disgust) with an expanded "guardian[ship] of the temple of the body", triggered by the violation of cultural rules for the use of the body, including mere contact with those occupying a socially lower station. Haidt argues that this "sociomoral" disgust patrols the lower boundary of what it means to be human, triggered by degradation and blurring of the distinction between humans and other animals. The associated action tendency is a motivation to expel or end contact, coupled with a motivation to purify oneself.

Anger and disgust jointly instantiate a reward-punishment structure to deter those involved in culturally inappropriate behaviors, motivating us to change our relationship with violators. Similarly, contempt maintains distinctions of rank and prestige by inducing feelings of moral superiority.

Haidt argues that the "self-conscious" family of moral emotions is designed to help individuals fit into groups without triggering the "other-condemning" emotions in others. Significantly, these moral emotions are culturally-specific, taking the familiar forms of shame, embarrassment and guilt in "Western" settings, but collapsing into a single emotion (combining shame, embarrassment shyness, modesty and social fear) in Asian cultures. Other cultures are not discussed. Haidt explains this distinction on the basis of "construal of the self" – if the self is construed as independent and the social structure is egalitarian, shame, embarrassment and guilt will be differentiated, but if the self is construed as interdependent and the culture is hierarchical, then self-conscious emotions will merge. Haidt makes no mention of interdependent-egalitarian or independent-hierarchical situations.

Shame is based on a feeling caused by being in the presence of one's social superiors ("protoshame"), but expands into a more complex form triggered by norm violation of which others are aware. In "Western" cultures, shame is an indication that one has failed in the project of bringing about a strong, competent and virtuous self (pride is the reverse). The related moral emotion of embarrassment is elicited within the context of a single interaction, and emerges most commonly from the violation of socio-conventional rules. The associated action tendencies involve reduction of social presence. By contrast with these hierarchical relationships, guilt grows out of communal relationships and the attachment system, in cases where one believes one has caused harm, loss or distress in a relationship partner. Guilt is far more common in close relationships, and is further distinguished from shame by its specificity – guilt touches particular actions, while shame concerns self-image. The associated action tendency is increased care in relationships, treating partners as they would like to be treated. All the "self-conscious" moral emotions lead to prosocial behavior by inducing conformity to social rules and the upholding of the social order.

The "other-suffering" family of moral emotions is limited to compassion. Haidt rules out the distinct construct of "distress at another's distress" because it is said to be an "affective precursor" of compassion. Compassion grows out of the mammalian attachment system, and is elicited by the suffering or sorrow of others, particularly close kin. Its action tendencies are impulses to help, comfort or alleviate the suffering of other people. Intriguingly, those who are most prone to feel compassion are among the least prone to feel shame (though feeling normal amounts of guilt). Finally, the "other-praising" family involves such moral emotions as pride, gratitude and elevation. As we saw, pride is simply the opposite of shame.

Gratitude is part of the mechanism of reciprocal altruism, encouraging beneficiaries to repay benefactors in the same way that anger motivates the punishment of nonreciprocators, and is triggered by the perception that others have done us a good turn, with an associated tendency towards prosocial action. Elevation (awe) is elicited by exposure to certain kinds of beauty and perfection, particularly manifestations of humanity's "higher" or "better" nature such as charity, kindness, loyalty and self-sacrifice. It seems to be the opposite of disgust, patrolling the upper boundaries of human nature in a similar fashion. Elevation seems to make people more open to new experiences and new ideas, directly motivating prosocial behavior. The associated action tendency is a desire to follow the example of the moral exemplar and become a better person oneself. Haidt concludes that moral reasoning is an epiphenomenon, and he proposes a "social intuitionist" model whereby moral emotions are primary but individuals deploy moral reasoning to persuade others.

Some experimental evidence seems to bear out Haidt's argument. Daniel Fessler (2004) conducted studies in Bengkulu (Indonesia) and California (USA), determining that self-reported shame was more common in Indonesia and that self-reported guilt was more common in California. He also found that shame was associated with guilt-like accounts in California but not in Bengkulu, and that subordinance was prominent in Bengkulu but not in California. On this basis, Fisher hypothesizes that shame evolved from a rank-related emotion and was originally displayed in the presence of those of higher rank. Despite the fact that shame has expanded to motivate additional behavior such as prestige competition, cooperation and conformity, it continues to play this rank-related role in contemporary humans, though this subordinance is said to be less prevalent ("culturally hypocognized") in California.

Haidt's assertion that our moral emotions are hypocritically self-justifying sits uneasily with his view that the same moral emotions prod us towards prosocial actions that redound to our long-term benefit. Surely if moral emotions enable group cooperation they must consist of something more than mere ratiocinated self-interest. His distinction between "Western" and non-"Western" cultures also seems spurious. As Kwame Anthony Appiah aptly puts it, "[t]reating international difference between…"the West" and the "non-West," as an especially profound kind of something called "cultural difference" is, in my view, a characteristically modern mistake." (Appiah, 2005, p. 254). A theory of morality ought not to contain culturally specific epicycles, and should explain a multiplicity of moral behaviors on the basis of universal human moral impulses.

Elliot Turiel vigorously disputes Haidt's conclusion that moral reasoning has an exclusively external function. He argues that moral judgments begin at a very young age, and that they are distinct from social and personal judgments. On this account, emotion and reason are intimately intertwined and analytically

inseparable. Turiel rejects a relativistic account of morality across cultures, arguing that the distinction between so-called individualistic and collectivist cultures is spurious and unfounded (p.487), particularly because many people occupying subordinate status in so-called collectivist cultures are deeply invested in concepts like autonomy, independence and resist cultural practices promoting inequality.

He also rebuts Haidt's proposed theory of moral intuitions, invoking research that questioned response speed as a measure of reason. According to Turiel, "a quantitative criterion of response speed is inadequate as a means of determining whether reason is at work" and he argues that response speeds are simply a function of certainty coupled with a low incentive to investigate details. Turiel thinks this analysis vitiates the concept of a moral emotion, and he bemoans psychologists' tendency to "propose that people function in fundamentally different ways from their own." He further criticizes Haidt's five alleged "moral foundations," arguing that on historical examination, Haidt's argument precludes progressive claims of social justice and universal rights, and affirms the morality of particular epochs as foundational and fundamental.

Turiel points out that if we assume that moral decisions are unconscious and given, an empiricist can simply ignore their definition, but if we assume that people are consciously engaged in moral reasoning, it becomes important to know the definition to which they are appealing. He takes issue with moral-psychological research along the lines of the infamous "trolley problem", arguing that in fact such scenarios are "highly unusual…complex and emotionally-laden," and are interesting precisely because they feature the rare conflict of fundamental values. As a result, they produce conclusions that are not generalizable and are a poor starting point for moral research. Turiel also argues that universal moral sentiments interact with beliefs about the world (embodied in culture) to produce differentiated effects in diverse societies. Thus, alternative acts seemingly different from each other can be motivated by similar moral concepts.

Finally, Turiel disputes accounts that emphasize the primacy of emotion over reason. He believes that emotional appraisals are vital constituent parts of reasoning that take into account the reactions of others. Children as young as three are able to differentiate easily between issues of convention and morality, and (contra Piaget) they readily label harmful acts as wrong even when told by authority figures that the act in question is permitted. Morality appears not to be conditioned by existing social arrangements. Turiel takes this to indicate that "authority is not a moral orientation." He emphasizes that young children think hard about their moral experiences, and that emotions involve both evaluative appraisals and moral reasoning.

Turiel's critique of Haidt seems decisive, and his reminder to not reify concepts like "the West" is instructive. However, the precise definition of reason at

work in the arguments of Turiel and Haidt seems inconsistent. Turiel quotes Martha Nussbaum approvingly to the effect that humans are fundamentally reasoning beings, but his "reasoning" seems to embrace a great deal that Haidt would call moral intuition. However, Turiel's proposed authority-independent foundation of morality, if true, implies that cultural difference occludes significant commonality regarding the most freighted moral questions. On this point, Piazza and Sousa (2016) reexamined a prior cross-cultural study (Fessler et al. 2015) that had purported to show moral parochialism. After reanalyzing the results, Piazza and Sousa find that in cases involving harm or injustice, the parochialism effect disappears and judgments are highly correlated across cultures. Respondents in all samples rejected attempts to substitute authority for moral judgment.

Turiel's framework has been questioned in two ways.[1] First, there is evidence that some conventional transgressions are seen as authority-independent, and second, some harmful transgressions are *not* seen as authority-independent. Sousa and Piazza (2014) extend Turiel's framework by proposing that harmful transgressions are seen as authority-independent and general in scope if the causation of harm is interpreted as involving "basic-rights violation and injustice." It is this injustice, rather than the harm itself, that makes the transgression a *moral* transgression. This would not seem to be a significant enhancement of Turiel's formulation, because Sousa and Piazza acknowledge that their proposed pathway (basic rights violation) is not the only factor that can make a transgression a moral violation and thus authority-independent (p.125). Though they present impressive symbolic formulations of their straightforward propositions, the authors do not appear to make an analytically useful distinction. Resolution of the anomalies noted at the outset of this paragraph might be better achieved by questioning the prevailing definitions of harm and authority.

In a related study, Finger and her coauthors (2006) find that brain regions associated with moral reasoning were activated in different ways by prompts involving moral transgressions and social transgressions. Prompts involving moral transgressions activated the relevant brain regions whether or not an audience was present. By contrast, prompts involving social transgressions activated the same brain regions, but only in the presence of an audience. These brain regions may modify behavioral responses in reaction to social cues. The authors suggest that even verbal descriptions of observation may be sufficient to activate differential processing.[2]

---

[1] Along the lines of Type I and Type II errors.

[2] Work by these authors on the effects of anticipated gaze were anticipated by Julian Jaynes in his idiosyncratic magnum opus, "*The Origin of Consciousness in the Breakdown of the Bicameral Mind*" (1976). Jaynes studied, *inter alia*, the curious prevalence across ancient cultures of god-statues with grossly oversized eyes, reaching similar conclusions to Finger et al. concerning the social utility of known observation (i.e. effects on behavior).

It seems that we may have one system for processing conventional norms, and another for reaching moral conclusions. White et al. (2017) investigate whether social norms are processed by a single unitary system or diverse brain systems, finding that there is neural differentiation between harm/welfare-based and conventional transgressions. They speculate that processing the intent of social norm transgressors was vital in early hominid societies, prompting the evolution of a separate pathway. In addition, even minor moral transgressions were seen by study participants as more harmful than major social transgressions. The findings of this study are neurologically subtle, and seem to indicate that while a common process underlies judgments regarding social norms, norm judgments take on relative degrees of affect associated with the transgression, recruiting different brain processes as a result.

The role of anticipation in norm construction and maintenance appears to be underemphasized. Mackie et al. (2015) argue that social norms can be maintained by approval or disapproval within a reference group, and that this approval or disapproval is often conveyed by facial expressions. Anticipation of (positive or negative) sanction can lead us to change our behavior. Compliance, on this account, follows mostly from anticipated sanction rather than actual sanction. The norm is thus maintained by beliefs about what would happen if we did not comply. Social-normative regulation is generally subtle, indicating light approval or disapproval that drives anticipatory behavioral change. Even in the absence of explicit sanction, mere (internal) attitudes of approval or disapproval can induce similar effects. Norm compliance is not merely instrumental, however – intrinsic valuation of approval or disapproval can operate even in contexts where there is no possibility of sanction. Strictly moral obligations appear to operate differently, requiring individual compliance without regard to external sanction.

The face itself appears to be crucial in generating these effects. Liu et al. (2019) note that face-to-face interactions are more effective in inducing compliance than other forms of interaction, and they propose the explanation that this "face effect" is largely attributable to anticipated facial feedback. In addition, individuals can be primed to increase the face effect by sensitizing them to human faces generally. Effects are strongest when subjects have been so sensitized and when the experimenter's face is relatively expressive. Crucially, anticipated facial feedback is sufficient to drive the face effect. The authors speculate that we are concerned to elicit positive facial feedback, even if we must take costly actions to do so. The facial feedback from the counterparty must be both expressive and situationally appropriate for the face effect to operate.

These studies help us to appreciate the vast amount of virtually seamless emotional anticipation that occurs in any human society. The fact that at least some of this esteem is not merely instrumental suggests that some sort of primary goods exchange is at work. Geoffrey Brennan and Philip Pettit (2004) propose a

widespread market mechanism for the exchange of esteem. The authors seek to rehabilitate the concepts of honor and esteem as motivations for human behavior. They argue that esteem is an evaluative, comparative and directive attitude, which is to say that esteem is given or withheld on the basis of specific actions taken and the success or failure of those actions relative to the performance of others. Esteem, in other words, includes a core element of interpersonal competition. Self-regulation and self-control are character traits likely to induce appraisals of esteem. People will be concerned to increase the number of those who esteem them and reduce the number who disesteem them (compare with Mackie et al.'s "sanction"), though these efforts will be concentrated among qualified reference groups with the competence to properly evaluate performance.

The authors begin with the presumption that "among the things that may be expected to move people most reliably and forcibly is the desire to be thought well of by their fellows and the aversion to being regarded badly." So far, this is simply variations on a theme by David Hume. Their original contribution comes in the application of market terminology to the human search for esteem. They contend that the drive for esteem increases inclusive fitness and is therefore adaptive, and they argue that esteem is a Rawlsian primary good in the sense that it cannot be reduced to consumption goods.

The authors seem to have had the (correct) insight that human beings operate in a market for esteem, but their exposition is extremely disappointing. In the first instance, they might have looked to past attempts to understand this "economy of esteem". For instance, Aristotle's account of virtue in the Nicomachean Ethics suggests that ancient Greek society was best understood as an *agon* – a competition for the good opinion of one's fellow citizens – and that this agonistic understanding of life was baked into every aspect of society. We already possess a well-developed vocabulary for talking about an agonistic society, and its terminology has been applied to societies as diverse as feudal Japan and the American South. Further, leaving these missed antecedents aside, the authors' prose is labored and repetitive, answering few of the interesting questions raised by their proposed model. If we exist in a market for virtue, what are the implications of living in an anonymous modern society? The authors' contention that modern societies are not so anonymous after all is true but facile – interesting analysis might have involved, for instance, consideration of the multiple opportunities for self-reinvention unavailable to those living in traditional societies or the impact of widespread value pluralism on decisions to grant or withhold esteem.

Probing the relationship between esteem and social norms more deeply, Richard McAdams argues that as long as people seek esteem as an end in itself, then norm formation is inevitable. His proposed mechanism is that particular behaviors will cause many people to grant or withhold esteem, and that this coordination is well-known. If these conditions hold, even a weak concern for esteem

can create significant costs for acting against the consensus. "When the private costs exceed the private benefits of violating the consensus, a norm emerges" (p.433). Over time, these norms become entrenched and the costs of norm violation increase. Secondary enforcement norms will rise up around the initial norm, and norms can be said to be "internalized" once enforcement of the primary norm occurs internally for quasi-moral reasons. In this way, concrete esteem-based norms can define the meaning of abstract internalized norms. However, I am not persuaded that the concept of norm internalization is analytically useful, because as Mackie et al. discuss, anticipatory concerns are far more salient than actual sanction in norm enforcement. This could of course mean that most norms are internalized, but it seems difficult to distinguish between anticipatory sanction and acceptance of the principles on which the norm is based. Finally, McAdams makes the cogent point that law can manipulate norms by manipulating the information environment, by such means as publicizing an emerging consensus.

We now have some of the tools to better understand moral reasoning. The anticipatory role of emotion may actually be its most important feature. Baumeister et al. (2007) propose understanding emotion as a feedback system with indirect influence on behavior. Rather than directly influencing behavior (which the authors argue would be maladaptive), emotion retrospectively associates strong affect with past experience, thereby making particular patterns of behavior either more or less likely. On this account, people will prospectively choose behavior likely to result in pleasant emotional states. The authors distinguish between conscious emotions (which they call "full-blown emotions") and mere automatic responses, which they label "affect." Affective responses are typically strongly positive or negative, without significant intermediate detail. Affect is said to arise instantaneously, whereas emotions arise only on reflection, typically after an incipient crisis has passed.

The authors theorize that people make split-second decisions entirely on the basis of affect, and only then experience emotions related to the event. These emotions are said to aid in reflection, and to make the pursuit of positive affect and the avoidance of negative affect more likely in the future.[3] The authors ask whether it might be possible to reap the advantages of these different types of thinking without bearing the concomitant disadvantages. To me, this puts the question the wrong way around. Surely the two types of thinking are solutions to an evolutionary problem, rather than random processes that early humans discovered a way to optimize.[4] In any event, the authors make a persuasive case that emotion is not a behavior-causing mechanism, but rather exists to prompt reflection and elicit optimal long-run behavior by prompting anticipation of future emotional outcomes.

[3] This way of understanding behavior emphasizes the similarities between human affective response and the aversion/approach mechanisms of even the simplest forms of life, such as paramecia.

[4] It is perhaps a drawback of training in politics that one begins to see legislative processes everywhere, but I am struck by the analogy between rapid executive action in a short-term crisis and the sort of deep legislative deliberation required to prosper in the long run.

On this understanding, a great deal of behavior can be understood as prompted by the desire to regulate future emotion. Emotion serves as a stimulus to cognitive processing, and anticipated emotion may be more important than actual emotion. Limited experimental evidence bears out this view. In a meta-analysis of studies of emotion, DeWall et al. found that direct causation of behavior was only significant in 22% of tests, while the emotion-as-feedback perspective received support in 87% of tests. They concluded that empirical evidence was weak for direct emotional causation of behavior, but the proposition that anticipated emotion reliably impacts social behavior received much stronger support.

We can now give a much more satisfying account of the emotional basis of norm accretion. Emotions seem to exist to guide future behavior, providing anticipatory guidance and moving people towards outcomes associated with positive affect. Mutual anticipation in strategic contexts can result in the emergence of behavior-guiding norms. Our moral emotions appear to rely on distinct brain systems from those underpinning social conventions, but these consistent moral emotions can be refracted by a multiplicity of arbitrary conventional expectation and situational context, giving rise to considerable diversity in social norms.

3988 words.

Works Cited

Baumeister et al. "How Emotion Shapes Behavior: Feedback, Anticipation, and Reflection, Rather Than Direct Causation." Pers Soc Psychol Rev 2007 11: 167

Brennan, Geoffrey and Pettit, Philip. "The Economy of Esteem – An Essay on Civil and Political Society." Oxford University Press (2004).

DeWall et al. "How Often Does Currently Felt Emotion Predict Social Behavior and Judgment? A Meta-Analytic Test of Two Theories." Emotion Review, Vol. 8, No. 2 (April 2016) 136 –143.

Elliot Turiel. "Moral Development." Ch. 13 in The Handbook of Child Psychology and Developmental Science, Seventh Edition (2015).

Elster, Jon. "Emotional Choice and Rational Choice" in The Oxford Handbook of Philosophy and Emotion.

Fessler, Daniel. "Shame in Two Cultures: Implications for Evolutionary Approaches." Journal of Cognition and Culture 4.2 (2004).

Finger et al. "Caught in the act: The impact of audience on the neural response to morally and socially inappropriate behavior." NeuroImage 33 (2006) 414–421

Haidt, Jonathan. "The Moral Emotions" in Davidson et al. "Handbook of affective sciences". Oxford University Press (2003) pp.852-870.

Liu et al. "The Role of the Face Itself in the Face Effect: Sensitivity, Expressiveness, and Anticipated Feedback in Individual Compliance." Front. Psychol. 9:2499. (2019)

Mackie, Gerald, Moneti, Francesca, Shakya, Holly and Denny, Elaine. "What Are Social Norms? How Are They Measured?" UNICEF / University of California Center on Global Justice. July 27, 2015.

McAdams, Richard. "The Origin, Development and Regulation of Norms" *Michigan Law Review*, Vol. 96, No. 2. (Nov., 1997), pp. 338-433.

Paulo Sousa & Jared Piazza (2014) Harmful transgressions qua moral transgressions: A deflationary view, Thinking & Reasoning, 20:1, 99-128

Piazza J, Sousa P. 2016. "When injustice is at stake, moral judgements are not parochial." Proc. R. Soc. B 283:20152037.

White et al. "Neural correlates of conventional and harm/welfare-based moral decision-making." Cogn Affect Behav Neurosci (2017) 17:1114–1128.