Moral Psychology

James Dungan & Liane Young

Department of Psychology, Boston College

One goal of moral psychologists is to understand the cognitive processes that support and influence human moral judgment. Perhaps unsurprisingly, this task has proven to be quite daunting. Moral psychology, as well as moral philosophy and anthropology have revealed great diversity in human moral judgment. Moral philosophers have debated ethics for centuries, yet to this day display a wide diversity of opinion on what a correct solution is. Anthropologists have also documented diversity in moral judgment, only across different cultures – what is moral in some cultures is seen as extremely impermissible in others. Still further, psychologists have used the tools of cognitive neuroscience (such as functional magnetic resonance imaging) to show that individual differences in cognitive processes greatly affect moral judgment.

Historically, psychologists have dealt with this diversity by focusing on the unifying aspects of morality, studying commonalities in moral judgment across individuals and cultures. Many moral judgments are robust to different demographic factors such as gender, age, ethnicity, and religion. For example, intent plays a consistent role in people of all ages' moral judgment (e.g. intending to harm someone is worse than accidentally harming someone). Also, in the trolley dilemma (a popular philosophical scenario), an overwhelming majority of participants judge turning a runaway trolley away from a track with five people on it to a track with one person on it to be permissible, but pushing a man off a bridge onto the tracks to stop the trolley to be impermissible (even though both cases trade one life to save five).

While these approaches have rendered understanding moral judgment a tractable problem, many complexities in moral judgment are left unresolved. No comprehensive model or

taxonomy of moral judgment thus far has accounted for its full diversity. Some models call for a division of the moral space based on the content or kind of moral violation. We judge those who harm others, those who cheat and steal, those who betray their family, friends, and country, those who are disrespectful and disobey authorities, and even those whose actions do not necessarily affect others but instead render themselves "impure," such as consuming taboo foods. Each of these acts may represent a distinct area of moral judgment. Other models carve up morality in terms of the nature or structure of the relationships affected by the violation. For example, how one should act toward another depends on whether the target is a friend, a stranger, an equal, a subordinate, or an authority.

How should we divide up the moral space? Settling on a good taxonomy represents a crucial step toward understanding moral psychology, allowing us to determine through experimentation how different kinds of moral judgment are influenced by psychological, emotional, social, and cultural factors. Here, we discuss the limitations of existing act-based and relationship-based divisions and offer a compromise between these existing divisions. We propose a two-type model of morality, wherein both the moral act and the relationship it affects are taken into account. We suggest this model reflects a real psychological distinction with evidence from emotional, behavioral, and cognitive processes.

Do different kinds of moral acts define different moral domains?

When faced with the substantive task of dividing up the space of morality, the specific content of moral actions emerges as an obvious starting point. That is, moral boundaries may serve to separate actions that cause harm to others, from actions that show disrespect, from actions that offend God, and so on. To its credit, this approach explains much of moral diversity

across cultures. For example, Shweder, Much, Mahapatra and Park (1997) surveyed Hindu Indians and compared their explanations of moral actions to explanations delivered by Westerners. By and large, Westerners presented a restricted conception of moral action, defining immoral actions as primarily those that violate the "ethic of autonomy". In this case, an action is wrong because it directly harms someone or violates his or her individual freedom and rights. By contrast, Indians, as well as many other eastern cultures, additionally moralize actions concerning disobedience and impurity, the "ethic of community" and "ethic of divinity", respectively. These ethics include additional moral concerns of doing one's duty to the community and respecting social hierarchy as well as respecting the sacredness of God and the sanctity of the human body. Shweder and his colleagues propose that variation in human moral psychology across cultures around the world can be explained by differing adherence to these three distinct ethics, community, autonomy, and divinity, nicknamed the CAD Triad hypothesis (Rozin, Lowery, Imada & Haidt, 1999).

This content-based approach also proves fruitful in explaining different emotional responses to different kinds of moral violations. In particular, Shweder, et al.'s three ethics map well onto three kinds of emotional reactions to moral violations. In one study (Rozin, Lowery, Imada & Haidt, 1999), students in Japan and the United States read descriptions of moral violations and indicated their emotional response by selecting an appropriate facial expression or emotion word. As this study found, violations of community evoke contempt, violations of autonomy evoke anger, and violations of divinity evoke disgust, supporting the CAD Triad hypothesis (Rozin, Lowery, Imada & Haidt, 1999), and more generally Shweder, et al.'s content-based division of morality.

A more recent content-based approach supports either further division in the moral space. Haidt and Joseph (2004) surveyed the anthropological literature for actions commonly governed by moral codes across cultures, but divided the moral space into five domains, not three. According to the latest version of their Moral Foundations Theory (MFT), these domains are harm/care, fairness/reciprocity, ingroup/loyalty, authority/respect and purity/sanctity (Haidt & Graham, 2007). In many respects, MFT represents a culmination of content-based divisions of morality, offering precise predictions about many levels of moral psychology. At a mechanistic level, the five moral domains correspond to specific evolved psychological mechanisms that explain the intuitive, emotional basis of many moral judgments (Greene, 2001; Haidt, Koller & Dias, 1993). For example, purity norms stem from evolutionary concerns of disgust (Rozin, Haidt & McCauley, 1993), thereby explaining the disgust response across many cultures to purity violations such as incest. At the level of social and cultural considerations, MFT explains cultural and political differences in moral judgment. Consistent with other content-based approaches, people from different cultures or of different political orientations vary in how they value different moral domains.

While MFT, like its content-based predecessors, has clear predictive power for moral psychology, it may suffer from being either too limited or infinitely divisible. For example, the MFT has been said to be limited in failing to account for the full range of moral values, including moral valuations of modesty and industriousness (Suhler & Churchland, in press). Content-based theories can of course be extended, to accommodate additional divisions. Indeed, even now the MFT is being extended to include a domain related to liberty/constraint and a domain related to wastefulness (Haidt & Joseph, in press). However, the flexibility of content-based approaches may also be a weakness, that is, infinite divisibility. Content can be divided (depending on the

individual or the culture) to fit any possible behavior, leading to three, five, eleven moral foundations or more. On what basis do we decide that one kind of behavior deserves its own domain? What principles that operate over moral content ought to guide moral psychologists to lump versus split? Though these questions aren't impossible to answer, they prompt careful consideration of content-based approaches to carving up moral psychology. Indeed, these questions center on a broader one: are content-based approaches capturing qualitative differences in moral psychology as opposed to differences in the mere content of different actions (e.g., kick versus hit versus cut)?

Do different moral relationships define different moral domains?

On March 16, 2008, seventeen-year-old Rand Abdel-Qader was beaten, strangled and stabbed to death by her father, Abdel-Qader Ali, in Basra, Iraq. What was the reason for this violent attack? Rand had been seen in public conversing with a Christian British soldier, considered, by her family, to be the enemy. Abdel-Qader was held for only two hours before being released without charge and reportedly congratulated by the local police for restoring honor to his family (Sarhan & Davies, posted 11 May 2008).

Though many people would emphatically denounce this action as unambiguously abhorrent, honor killings are permitted in the penal codes of a number of countries around the globe – Argentina, Venezuela, Israel, Jordan, Syria, Egypt and Iran, to name a few. For instance, according to article 460 of the Islamic Penal Code, a man who finds his wife committing adultery may kill her and the man she is with.

Suppose, however, that Abdel-Qader Ali did not murder his daughter, but a police officer. Surely the other officers would not have offered him praise. This hypothetical case and

the actual event present a problem for content-based divisions of morality, as surveyed in the previous section. Within the same culture, and for the same people, the same act (in this case, murder) may be judged moral or immoral, right or wrong, depending on its target (or victim). Observations about an action's content (e.g., whether the action is harmful, unfair, impure) appear to be insufficient to account for complex moral judgments. What theory of moral psychology can account for such moral judgments that a content-based approach fails to accommodate?

One solution is to divide up the space of morality not by content but rather by the nature of the social relationship that provides the context for the action. Rai and Fiske (2011) propose that morality consists of specific motives to preserve different kinds of relationships. This model of moral psychology is completely content-neutral – the kind of action does not matter; what matters is the relationship that is primarily affected by the action. This approach has the advantage of being able to explain how an action may be judged quite differently depending on the relationship context. Specifically, Rai and Fiske (2011) describe four distinct relationship schemas: Unity, Hierarchy, Equality and Proportionality. Unity relationships are close-knit in-groups that share a common fate, i.e. relationships between family members or close friends. In unity relationships, people can freely take from one another; indeed, active accounting to keep things objectively fair is counterproductive and often undermines the relationship. By contrast, in equality relationships, balance, fairness and reciprocity must be maintained. Equality emerges in justice systems where the punishment is equal to the crime (i.e. the death penalty for murder or more extreme cases in the Middle East, where a victim who was paralyzed by his attacker asked the Saudi Arabian courts to in turn medically sever his attacker's spinal cord; see CNN News, posted 20 August 2010). Hierarchy relationships maintain a linear ranking where people at the

top, the leaders, are entitled to more than people at the bottom. In exchange, people expect their leaders to provide protection and guidance. The balance differs further in proportionality relationships, where interactions are based on proportional cost and benefit, not equality (i.e. paying a fine that is proportional to the severity of a committed crime). On this theory, then, even fairness cannot be defined by the content of any single action – what is considered fair varies depending on the relationship of the interacting agents.

We suggest that while Rai and Fiske's model explains the variability of moral judgments of the same acts across different moral relationships, this model may go too far in abandoning considerations of content. A father murdering his daughter may be worse than murdering a stranger (or better, as in the view of Abdel-Qader Ali); however, a father murdering his daughter is still qualitatively different from, for example, lying to his daughter. Because Rai and Fiske's model is completely content neutral, it does not adequately account for differences between acts targeting the same relationship.

On a different account, the *kind* of relationship may not actually matter. Instead, what matters for morality is simply the presence of (at least) two interacting agents – independent of their relationship. Gray and Wegner (in press) propose that the moral dyad defines all of morality. More specifically, dyadic morality requires two different people – the moral agent who performs a moral action, and the moral patient whom the action affects. Critically, an act is perceived to be in the domain of morality whenever a moral agent helps or harms a moral patient.

Gray and colleagues provide the clever example of stealing to support the dyadic nature of morality – you can't steal from yourself. Yet, what about moral cases where there is no dyad? Self-harms, for instance, are often seen as immoral (e.g., eating taboo foods, committing incest

or suicide), but involve no clear dyad – the moral agent and moral patient may be one in the same. Furthermore, by restricting the realm of moral acts to a single positive/negative dimension of help/harm, dyadic morality may sacrifice the strength of content-based approaches – namely, explaining how different kinds of moral acts may be viewed so differently.

A Compromise: Content and Relationship Matter

Cultural views of self-harm vary substantially. Suicide is expressly forbidden in Abrahamic religions and often carries negative connotations in Hinduism and Buddhism, since it is seen as an affront to God's will or a desecration of the soul and its spiritual journey. Meanwhile, self-cutting is seen in Western cultures as desecrating the body, God's temple, or is otherwise highly stigmatized (Synnott, 1992). By contrast, more positive associations with self-harms can be found in other cultures. Seppuku, a form of Japanese ritual suicide, is performed to preserve honor or in response to committing a deeply shameful act. Many religious cults such as Peoples Temple or the Order of the Solar Temple endorse mass suicides as a purifying escape for the soul to a better world. Finally, throughout history, self-harms such as self-flagellation have been practiced in order to purify the flesh (see also, Bastian, Jetten & Fasoli, 2011).

Whether permissible or forbidden, a striking similarity across these self-harms is their association with purity. An important determinant of their association with purity seems to be that they affect the self. Performing precisely these acts – killing, cutting, whipping – on another person, is seen as obviously harmful, but not at all impure. This dichotomy lays bare the problem with theories of morality that separate content and relationship – content and relationship are not absolutely orthogonal. Both the content of an action and the relationship the action affects matter for how the moral violation is judged. Put concretely in terms of harm and purity, morals of purity are not simply concerned with preventing disgusting impure acts – but impure acts that

defile the self. Meanwhile, harm morals are not simply about preventing hitting, cutting or other harmful acts – but acts that *harm another person*.

To test this theory, we conducted a series of pilot studies aimed at testing people's willingness to commit purity and harm violations against themselves or others (Dungan, Chakroff & Young, in preparation). We presented people with a list of ten purity violations and ten harm violations varying in severity (e.g., picking up a dirty Kleenex, drinking cow's blood, being pinched on the arm, being whipped with a belt). Subjects ranked these violations 1 – 20 from what they would most 'prefer' to what they would least prefer. Critically, half the subjects ranked what they would prefer to happen to themselves while the other half ranked what they would prefer to happen to a close friend. We predicted that people would prefer the option most consistent with the adaptive function of the particular moral at stake – not hurting others, not contaminating the self. Consistent with this prediction, when participants chose for themselves, most preferred to be harmed rather than rendered impure; however, when choosing for a friend, the pattern reversed – most people preferred to render impure rather than to harm their friend.

Do people's preferences also track with their moral judgments? We conducted a second study where we presented subjects with a hypothetical situation in which they were standing in front of two buckets with a friend (Dungan, Chakroff & Young, in preparation). One bucket is filled with painfully hot water and the other with a stranger's sterile urine. In this scenario, subjects had to choose either for themselves or their friend to dunk their hand in one of the buckets for three seconds. Subjects were told to imagine choosing one of the options for themselves or a friend and answer three questions: how gross, how harmful, and how wrong is your choice. As predicted, subjects judged dunking a hand in the stranger's sterile urine as more gross, regardless of if it was their own hand or their friend's. Subjects also rated the painfully hot

water as more harmful for themselves and their friend. Importantly, though, both the act (purity versus harm) and its target (self versus friend) affected judgments of moral wrongness. Subjects rated choosing the urine for themselves as more morally wrong than choosing urine for their friend. Subjects made the opposite judgments about the hot water – choosing hot water for their friend versus themselves was more wrong. This overall pattern supports our theory, that harms are most immoral when directed toward another person, while purity violations are most immoral when directed toward the self.

Preliminary evidence is therefore consistent with the need for a theory that represents a compromise between content-based and relationship-based divisions of morality, a theory that takes both content and relationship into account. We propose that morality can be defined by two types of moral rules:

1) Morals that govern how one should treat the self. These morals dictate how we should and should not treat the self or highly similar others that are seen as connected to the self.

2) Morals that govern how one should treat other people. These morals dictate how we should and should not treat other people, including people we see as independent and unrelated to self-identity.

So far, we have only described how harm and purity fit into the model. Can a two-type model of morality account for other moral concerns such as loyalty, hierarchy and fairness? In the

remainder of this chapter, we argue that this two-type model of morality is a useful taxonomy of the full range of moral actions.

Beyond Harm and Purity

Purity norms dictate what acts that are appropriate or inappropriate when the self is the target. The same may be true for the domain of loyalty. Indeed, loyalty and devotion can be seen as a blending of self and other – the weaker the distinction between self and other, the stronger the ties of loyalty. This is quite literally true at the genetic level for family members, and at the cognitive level for strong relationship pairs, as in transactive memory – where couples operate using the same shared system of encoding and retrieving information (Wegner, Erber & Raymond, 1991). Even less extreme instances of loyalty, including loyalties to one's friends or country, often define who a person is. People experience loyalty and affinity to a group to the extent that a violation against the group threatens their own self-identity. This is seen in group members' reactions to an outside force that threatens the value of the group. Members who show low loyalty react by distancing themselves from the group to protect their individual identity, whereas members who display high loyalty aggress toward the threat, defending the group and by proxy their own identities (Ellemers, Spears & Doosje, 2002; Branscombe, Ellemers, Spears & Doosje, 1999). In this way, actions against one's own in-group often feel the same as actions against one's self.

Violations of hierarchy and respect are similar in that they also impact self-identity. Hierarchy constitutes the structure of a group, defining the specific duties and responsibilities of each member, importantly, in relation to the other members. The relationship between each member to the group, and all the members to each other, means that the entire group may

experience the impact of any individual member's action – a staff sergeant's transgression reflects poorly on the entire platoon he represents just as the transgression of a platoon reflects poorly on the staff sergeant who leads it. Disrespect undermines the cohesiveness of the group, and thus affects every member. To the extent that what's good for the self is tied to what's good for the group, it makes sense that another group member's disrespectful actions are seen and felt as impacting the self.

As in loyalty and respect, fairness also requires an entity besides the self (i.e. you cannot steal from yourself; see Gray & Wegner, in press). However, whereas loyalty and hierarchy necessarily entail special obligations and connections, fairness does not. In its purest sense, fairness means treating everyone equally, unbiased by relationship commitments. Family and friends are treated no different from strangers (and even enemies). Fairness, like harm, operates independently of groups or shared identities, honoring notions of autonomy, independence, and freedom. With the absence of any connection between people, unfairness directed toward another person does not necessarily affect the self. As such, fairness morals govern how one should treat other people – not one's own self.

The five domains of morality, as posited by Moral Foundations Theory (Haidt & Graham, 2007), seem to fit quite well into a two-type model of morality. Loyalty, hierarchy and purity morals govern acts aimed at the self (or others who are intimately tied to the self-identity, by group membership or relationship status), whereas harm and fairness morals govern acts aimed at others (strangers or people unrelated to the self-identity). Of course, these five "domains" might not constitute an exhaustive list of all moral concerns, but it is reasonable to believe that the two-type model can accommodate other concerns as well (modesty/boasting, industry, equity) (Suhler & Churchland, in press). Indeed, the unique strength of the two-type

model is that it is not based solely on content and therefore does not need require an amendment for every new moral concern that may arise (Suhler & Churchland, in press).

It is worth noting that grouping loyalty, hierarchy and purity together as making up one type of morality and harm and fairness as another matches Haidt's grouping of binding versus individualizing morals (Graham, Haidt & Nosek, 2009; Wright, in press). We emphasize that our model does not rely on the binding-individualizing dimension to distinguish between the two types of moral rules. Nevertheless, this dimension usefully describes the consequences of these morals on a society. For instance, since the function of purity norms is to keep a person's body and soul free from physical and perhaps spiritual contaminants, it may be assumed that one person shouldn't care about another's impurity (barring fear that another's impurity might spread to contaminate others). However, the way that one interacts with his or her environment (e.g., the avoiding or not avoiding of specific contaminants, body modification, or cleanliness) may serve as an overt signal as to who should be avoided and who would make for a compatible group member. Purity norms may thus reinforce group boundaries and bind groups together (or break them apart) (Sosis & Bressler, 2003) without specifically governing interaction among individuals.

Also worth mentioning is the factor analysis of Haidt's Moral Foundations Questionnaire – a scale of the extent to which people care about different moral concerns. Though the model that this questionnaire is designed to test (Moral Foundations Theory) is different from the proposed two-type model, it is significant that an exploratory factor analysis of the data, free from any imposed theoretical constraints, supported a single distinction between two types, i.e.

between binding and individualizing moral concerns (Graham, Nosek, Haidt, Iver & Ditto, in press).[1] It is significant that a factor analysis divides moral domains into two types.

The two-type model of morality succeeds in accommodating various types of moral transgressions. We argue that dividing morals into those that govern actions toward the self and those that govern actions toward others is the best taxonomy for moral judgment. Though we have shown how this model works in theory, it is important that the division into two types of morals is not simply an abstract categorization. Rather, this division should reflect a real difference in our moral psychology, each type giving rise to different emotions, different behaviors, and judged according to different cognitive rules. We devote the rest of the chapter to outlining three types of evidence that suggest it does.

Evidence from Emotions

Cicero famously said: "Study carefully, the character of the one you recommend, lest their misconduct bring you shame." This sentence highlights a crucial aspect of shame – it involves expectations, or more precisely the failure to meet them. In particular, shame arises when a person fails to live up to the expectations of others – when a person fails in his duties or responsibilities. Shame signals a threat to a person's social bonds. Importantly, a person's

[1]A subsequent confirmatory factor analyses, performed to see if the data conformed to their pre-established theory, supported a breakdown of the moral space into the five domains posited by Moral Foundations Theory (harm, fairness, loyalty, respect and purity); however, there are reasons to be skeptical of this support. It is trivial that dividing into more factors explains more variance in the data. As it is unclear how parsimony was weighed in the confirmatory factor analysis, it is possible that this analysis was subject to over-fitting. Furthermore, the Moral Foundations Questionnaire was designed specifically to emphasize five different domains. This could artificially bias the data toward favoring the five-domain breakdown. The confirmatory factor analysis should only be trusted if a less constrained survey of a broader range of moral values still supports a five-domain model over a two-domain model.

negative evaluation is directed at the self, taking into account others' opinions and impressions as well (Lewis, 1971; Niedenthal, Tangney & Gavinski, 1994). By contrast, in the case of guilt, a person's negative evaluation is directed not at the self, but rather at his or her specific immoral actions. Guilt arises when a person's behavior is out of line with his or her *own* conscience or moral standards. Consequently, feelings of guilt are found to rely relatively less on whether there happened to be an audience for the action – guilt can arise independently of external observers (Smith, Webster, Parrott & Eyre, 2002; Tangney, Miller, Flicker & Barlow, 1996).

The distinction between shame and guilt raises an opportunity for testing the two-type model of morality. Since shame represents a negative evaluation of the self, then violations of loyalty, hierarchy and purity – actions that violate the self, so to speak – should be more effective at evoking shame than violations of harm and fairness. Since guilt can arise regardless of whom immoral behavior targets (immoral behavior is immoral whether it affects the self or someone else), the difference between self versus other type violations should be diminished. If anything, one might expect actions that negatively affect another person to evoke a guiltier conscience than actions that affect one's self. Therefore, violations of harm and fairness – which impact autonomous agents, free from group concerns – may be more effective at evoking guilt than loyalty, hierarchy and purity violations. Evidence confirming the prediction that shame is associated with one type of morality while guilt is associated with the other would provide significant support for the two-type model.

Anthropological work provides preliminary evidence. For instance, many Eastern cultures such as Islamic fundamentalist cultures in the Middle East, India, and traditional Japan are often described as shame cultures. In shame cultures, shame is used as a primary deterrent of immoral behavior. Not by coincidence, these cultures are the very cultures that emphasize

loyalty, hierarchy and purity as important moral values. Shweder, et al. (1997) described these cultures as "holistic cultures", where the concept of a person is role-embedded or bound to context. In this view, an individual is conceptualized as a node in a network. Consequently, people are expected to do what others expect of them, as opposed to simply what is objectively right or wrong. By contrast, Western cultures conceptualize individuals as more independent (Nisbett, 2003). Of course, people are still connected to others, but these connections are less fundamental to self-identity. These cultures, including ours in the United States, are often described as guilt cultures. In tune with the proposed distinction between the two types of morals, these cultures overwhelmingly emphasize concerns of harm and fairness – freedom, autonomy and equality are valued highly (Haidt, 2007).

The difference between shame and guilt cultures leads to reliable, measurable differences in behavior such as punishment. In Japan, more people endorsed restitution as a sanction for moral transgressions – with the goal of repairing the damaged social bond. Other work shows that apology has a bigger impact on subsequent punishment in Japan than it does in America (Haley, 1986). Apology significantly reduces punishment in Japan, but plays little role in punishment in America. Furthermore, Americans are more likely to endorse retribution and punishment as sanctions for immoral behavior – reflecting a focus on the individual agent. Indeed, punishments in America often result in intentional isolation of the offender from society; reintegration appears to be secondary (Hamilton, et al., 1983, 1988). Future work should move from the level of cultures to the level of individuals to provide a more complete psychological characterization of the relationship between guilt and shame and the two types of morality.

These findings concerning shame and guilt offer indirect support for the two-type model. Evidence from other emotions is also consistent with the model. For instance, violations of

hierarchy and purity often evoke contempt and purity, respectively (Rozin, Lowery, Imada & Haidt, 1999). In the two-type model, hierarchy and purity violations represent violations of the self; thus, the function of these emotions is to distance the violator from the self. When the violator is a person, contempt and disgust sever ties, removing the person from the in-group. In contrast, emotions such as compassion, pity and empathy draw a person into the in-group. When a person meets a stranger in need, one of the strongest predictors of an empathetic response is whether the person perceives the stranger as being similar (Krebs, 1975; Stotland, 1969). Furthermore, when subjects responded to hypothetical vignettes, empathy increased helping behavior only when subjects felt connected to the victims (Cialdini, Brown, Lewis, Luce & Neuberg, 1997).

Of course, these emotions can be taken too far in either direction. Disgust and contempt can lead to pushing people completely outside the moral circle, dehumanizing them (Waytz, Epley & Cacioppo, 2010). Likewise, too much empathy and compassion for others comes at a cost to the self. The bi-directional aspect to these emotions fits well with the two-type model of morality. Enemies may be harmed, but not mere strangers. As emotional attachment forms for a stranger, so forms friendship. With friendship comes loyalty. Emotions impact moral judgment in many ways, but certain types of emotions may have a privileged role in determining what types of morals should apply.

Evidence from Behavioral Tensions

Army Warrant Officer Hugh Thompson Jr. ordered M60 machine guns to be turned on his fellow American troops. Though they were never fired, many congressmen were infuriated by Thompson's actions. Beyond congress, the general public sent him hate mail, death threats,

and even mutilated animals. On this description, Hugh Thompson Jr. appears to be violently disloyal and anti-American. And on this description, anyone would seem to be justified in judging him as criminally immoral. But from another perspective, the same man appears much more complicated. Hugh Thompson Jr. famously ordered his men to turn their machine guns on American soldiers who were mercilessly killing dozens of unarmed civilians, mostly women and children, in what would become known as the My Lai Massacre. In total, 347 – 504 unarmed civilians died in the massacre. Surely many more would have perished if it were not for Thompson Jr.'s brave actions, for which he would receive the Soldier's Medal, albeit, not until a full thirty years later.

Though this case seems extraordinary, many people find themselves in high stakes environments where they are forced to choose between following the orders of an authority and doing what is best for a group versus doing what is just and fair and in the interest of everyone, not simply for the good of one's own group or leadership. Whistleblowers – people who publically expose immoral or illegal acts – illustrate this tension dramatically, as in the case of Joseph Darby, who blew the whistle on the prisoner abuse at Abu Ghraib; and Cynthia Cooper, who exposed phony bookkeeping at WorldCom, which aimed to hide $3.8 billion in losses – at the time, the largest incident of accounting fraud in U.S. history. Unfortunately, just as common is the violent backlash from a community against those who speak out against wrongdoing. Consider the case of Michael Brewer, who reported a classmate to the police for stealing his father's bicycle. In response, the classmate and two others doused Brewer in alcohol and set him on fire for being a "snitch" (CNN Justive, posted 13 October 2009).

A two-type model of morality may explain this backlash against people who are nevertheless acting in the name of morality. Being devoted and loyal to one's group will by

definition require favoritism or bias – in other words, not treating everyone equally or being at least a little bit unfair. Similarly, moral action taken for one's own self-interest (or the interest of one's in-group) would naturally come into conflict with the proper action to take for unrelated others (or even the out-group).

With Adam Waytz, we conducted a pilot study to determine whether the type of morals people cared most about predicted how they would behave when given the chance to blow the whistle on immoral behavior. We first presented subjects with the Moral Foundations Questionnaire (Graham, Haidt & Nosek, 2009) to determine the extent to which considerations of loyalty and fairness mattered to them. Subjects then responded to a series of events in which someone they knew (varying in relationship closeness) committed a crime (varying in severity). Subjects were asked what they would do if they had the opportunity to report the crime to the police. The results show that the difference in how much people cared about loyalty versus fairness, not either loyalty or fairness alone, predicted their likelihood to blow the whistle and report the crime to authorities.

The tension between loyalty and fairness appears in children too. For instance, reactions to tattling change over development. Infants as young as ten months old possess a rudimentary sense of fairness and justice, distinguishing helpers from hinderers (Hamlin, Wynn & Bloom, 2007) and even preferring a character that punishes a hinderer as opposed to someone who rewards them, effectively endorsing third-party punishment and thereby showing an early sense of justice (Hamlin, Wynn, Bloom & Mahajan, under review). Toddlers' tattling behavior also reflects similar sensibilities, signaling cases of injustice, like one child taking another's toy (Ingram & Bering, 2010). Importantly, these tattles are almost always truthful and result in positive actions on the part of adults to fix problems. Yet in adolescence, when social groups,

group membership, and group loyalties become highly salient, tattling is seen extremely negatively. The more an adolescent is perceived as tattling, the less he or she is liked and the more he or she is socially rejected, as measured both by peer and caregiver ratings (Friman, et al., 2004).

Thus, the evidence suggests that both the young and the old experience a strong tension between acting loyally versus fairly. People seem to be forced to favor one versus the other, as indicated by their image of the ideal moral person. When people are asked to describe the prototypical moral exemplar, different conceptions emerge (Walker & Hennig, 2004). One conception is of a *caring* exemplar – the ultimate moral person is someone who is loving, empathetic and altruistic. The caring person is agreeable, generous and selfless to those around them. But another conception is that of the *just* exemplar – someone who is fair and objective, principled, rational and open-minded. These divergent conceptions of moral excellence are in line with the idea that people are guided by different sets of morals when acting on the behalf of their loved ones and acting in spite of relationships in the name of fairness; consequently, the two may conflict.

Perhaps the most obvious example of conflicting moral values can be seen in the so-called culture war between liberals and conservatives in the United States. (Graham, Haidt & Nosek, 2009). Anthony Giddens (1998) writes, "The left favors greater equality, while the right sees society as inevitably hierarchical" (p. 40). A meta-analysis of 88 studies conducted in 12 different countries confirms that a reliable difference between conservatives and liberals is conservatives' acceptance of inequality (Jost, Glaser, Kruglanski & Sulloway, 2003). In exchange, conservatives care more about social order and the familiar. This tension between liberals and conservatives is not limited to a tension between fairness and loyalty, but rather the

full spectrum of moral concerns, clustered in a way that is consistent with the two-type model. Liberals care more about morals of harm and fairness, while conservatives care more about loyalty, hierarchy and purity (Graham, Haidt & Nosek, 2009).

Evidence from Cognitive Processes

On the night of October 2, 1996 a piece of masking tape caused the death of seventy people. During a routine cleaning of a Boeing 757 airliner, Eleuterio Chacaliaza left a piece of tape over the static ports just prior to Aeuroperu Flight 603's departure. The static ports are pressure-sensitive sensors that send vital information to the pilot about an aircraft's airspeed and elevation. The tape interfered with these instruments, causing the plane to crash into the ocean, killing all nine crew members and sixty-one passengers. Eleuterio Chacaliaza was charged with negligent homicide and was sentenced to a two-year suspension from his job (Seattle Times, posted 21 January 1998). Compare this to the case of Angelica Ortiz who was also charged with negligent homicide but was sentenced to two years in prison for the death of a single person (KZTV10, posted 14 April 2011). Why did Ms. Ortiz receive such a harsher punishment than Mr. Chacaliaza when her crime, also accidental, resulted in the loss of only a single life? Perhaps, because the death she caused was that of her eleven-month-old daughter. The eleven-month-old was left in a hot car, where she died of heat stroke, for at least half an hour while Ms. Ortiz shopped for groceries.

Relationships between mother and child, husband and wife, are special in that they come with certain responsibilities. Parents have a duty to care and provide for their children. Friends have a duty to be devoted and faithful to each other. Leaders must ensure the wellbeing of their followers. These obligations, though they aren't necessarily defined or explicitly stated, are

intrinsic to the nature of close-knit or hierarchical relationships. Accordingly, qualitatively different standards may be applied to transgressions of loyalty or hierarchy, where people have a duty to protect, versus transgressions of harm and fairness, where 'extra' relationship-defined duties are absent.

One way to show this is by examining the difference between actions and omissions. People show a robust tendency to judge harmful and unfair actions to be worse than omissions or failures to act to prevent harm and injustice. This remains true even when the consequences of the action or omission and the intent of the actor are held constant (DeScioli, Bruening & Kurzban, 2011; DeScioli, Christner & Kurzban, 2011). However, this is not the case when people are bound by certain relationships, as is the case of mother and daughter. Subjects were less sensitive to the moral difference between actions and omissions when the perpetrator was an authority figure or in a relationship with the victim compared to when the perpetrator was a subordinate or in an anonymous relationship with the victim (Haidt & Baron, 1996). In other words, a mother failing to protect her child is judged just as harshly as if she took positive action toward harming her child – since she should have foreseen the danger, and was responsible for preventing it.

This difference in moral judgment has important implications for divisions in the moral space. Further evidence for a distinction between loyalty, hierarchy and purity as one type of morality and harm and fairness as another type would be different cognitive inputs to these two different types of moral judgment. This is precisely what Haidt and Baron's (1996) study suggests. In tight-knit relationships where loyalty and hierarchy play a large role, people are judged based on the consequences of the actions rather than on whether they intended for the action to occur or not. By contrast, intent plays a significant role in judgments of harm and

fairness. Specifically, people who accidentally cause harm or injustice are judged more leniently than those who cause harm or injustice intentionally, as in the difference between murder and manslaughter.

To test this idea, we presented subjects with stories depicting harmful actions (e.g., physical and psychological harm) and purity violations (e.g., committing incest and eating taboo foods). Subjects' judgments of moral wrongness reflected a large difference between accidental and intentional harms and a significantly smaller difference between accidental and intentional purity violations (Young & Saxe, in press). In particular, accidental purity violations were judged especially harshly whereas accidental harms were judged relatively leniently. Purity violations, like loyalty and hierarchy violations, appear to depend less on intent than violations of harm and fairness. To ensure that this behavioral difference was not due to specific features of the intent information provided in the stimuli (e.g., the possibility that participants are simply unwilling to accept that people could truly unknowingly sleep with their own siblings), but due to fundamental differences in the cognitive processing for different moral types, we conducted a version of this experiment in the brain scanner. Subjects delivered their moral judgments while undergoing functional magnetic resonance imaging (fMRI). FMRI measures blood oxygenation levels, a proxy for brain activity, allowing us to determine if the neural response to different kinds of moral violations is consistent with the observed behavioral response. Indeed, consistent with the behavioral evidence, brain regions for reasoning about mental states like intentions (including the right and left temporo-parietal junction, precuneus and medial prefrontal cortex) showed reduced activity in response to purity violations compared to harm violations (Young, Chakroff, Dungan, Koster-Hale, & Saxe, submitted). Whether measured behaviorally or neurally, intent information matters less for moral judgments of purity versus harm.

The findings that cognitive inputs like intent matter to moral judgments concerning loyalty, hierarchy and purity less than judgments of harm and fairness provide further support for the two-type model of morality. Actions affecting strangers or other people rely more on mental states, since intention serves as a reliable indicator of a person's future actions, if he or she is friend or foe, to be trusted or avoided. For actions affecting the self, however, we are usually aware of our own mental states. The focus is simply on avoiding bad outcomes.

Research investigating the cognitive differences across moral domains is in its infancy. Nevertheless, the findings so far corroborate the two-type model of morality. For instance, a cross cultural survey has found that in Japan (a interconnected society with strong concerns of loyalty, hierarchy and purity) people use mental state information significantly less for judgments of wrongdoing than people in America do (an independent, autonomous society concerned primarily with harm and fairness) (Hamilton, et al., 1983). Furthermore, new research shows that increasing cognitive load, interfering with the ability to regulate moral responses, causes conservatives to de-prioritize loyalty, hierarchy and purity in favor of harm and fairness (Wright, in press). Accordingly, it is plausible that adherence to one type of morality, whereby everyone is treated as strangers, not to be harmed or treated unfairly, is the baseline. It may require extra cognitive resources to adhere to the second type of morality, whereby people are seen as relating to the self and connections are formed. Taken together, these emerging results provide further evidence that the divide suggested by the two-type model reflects a real difference in cognitive processing and not simply content. We suggest that two-type model of morality succeeds best at explaining these differences.

Conclusion

Morality is complex. It makes sense then that the space of morality can be divided up in many ways. Past theories have suggested divisions by content to explain cultural variability. Newer theories have suggested divisions by relationship, accounting for differences in judgments of the same act across different relationships. We suggest that neither approach suffices. A compromise is needed to fully explain human moral psychology – a theory that takes into account both the content of a violation and the relationship the violation affects. We have proposed how a two-type model of morality divides the space of morality into morals that govern how we treat our selves and morals that govern how we treat others. Based on three types of evidence (emotion, behavior, cognition), we suggest that the two-type model reflects a real psychological difference, rather than simply an intuitive taxonomy.

While we believe this taxonomy can guide future experiments in moral psychology, we emphasize the importance of openness to compromise as new findings emerge. Morality pervades many aspects of our lives. It is inevitable that a full understanding of moral judgment will require perspective and insight from multiple fields of research. Psychology, as well as philosophy, anthropology, and evolutionary biology all have relevant findings for researchers interested in morality. Just as the two-type model of morality represents a compromise between competing theories, so will compromise through dialogue between these fields lead to fruitful results.

References

Bastian, B., Jetten, J. & Fasoli, F. (2011). Cleansing the Soul by Hurting the Flesh: The Guilt-Reducing Effect of Pain. Psychological Science, 22(3), 334 – 335.

Branscombe, N.R., Ellemers, N., Spears, R. & Doosje, B. (1999). The Context and Content of Social Identity Threat. In Social Identity, Ellemers, N., Spears, R. & Doosje, B. (Eds.), Blackwell Publishers: Oxford, UK.

Cialdini, R.B., Brown, S.L., Lewis, B.P., Luce, C. & Neuberg, S.L. (1997). Reinterpreting the Empathy–Altruism Relationship: When One Into One Equals Oneness. Journal of Personality and Social Psychology, 73(3), 481 – 494.

CNN Justice (13 October 2009). Police: Juveniles laughed after setting 15-year-old on fire. Retrieved from http://articles.cnn.com/2009-10-13/justice/

CNN News (20 August 2010). Saudi Judge Mulls Spinal Paralysis Sentence. Cairo, Egypt. Retrieved from http://www.cbsnews.com/stories/2010/08/20/world/main6789603.shtml

DeScioli, P., Bruening, R. & Kurzban, R. (2011). The omission effect in moral cognition: toward a functional explaination. Evolution and Human Behavior, 32, 204 – 215.

DeScioloi, P., Christner, J. & Kurzban, R. The Omission Strategy. Psychological Science, 22(4), 442 – 446.

Dungan, J., Chakroff, A. & Young, L. (in preparation). Pain Versus Purity: Distinct moral concerns for self and other.

Ellemers, N., Spears, R. & Doosje, B. (2002). Self and Social Identity. Annual Review of Psychology, 53, 161 – 186.

Friman, P.C., Woods, D.W., Freeman, K.A., Gilman, R., Short, M., McGrath, A.M. & Handwerk, M.L. (2004). Relationships Between Tattling, Likeability, and Social Classification: A Preliminary Investigation of Adolescents in Residential Care. Behavior Modification 28(3), 331 – 348.

Giddens, A. (1998). *The third way: The renewal of social democracy*. Cambridge, England: Polity Press.

Graham, J., Haidt, J. & Nosek, B.A. (2009). Liberals and Conservatives Rely on Different Sets of Moral Foundations. Journal of Personality and Social Psychology, 96(5), 1029 - 1046.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (in press). Mapping the Moral Domain. *Journal of Personality and Social Psychology*.

Gray, K. & Wegner, D. M. (in press). Morality takes two: Dyadic morality and mind perception. In P. Shaver & M. Mikulincer (Eds.), Proceedings of the 2010 Herzliya Conference on Morality. APA Press.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M. & Cohen, J.D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. Science, 293(5537), 2105 – 2108.

Haidt, J . (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*. *108*, 814-834.

Haidt, J. & Baron, J. (1996). Social roles and the moral judgment of acts and omissions. European Journal of Social Psychology, 26, 201 - 218.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20,* 98-116.

Haidt, J., & Joseph, C. (2004). Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus*, pp. 55-66, Special issue on human nature.

Haidt, J. & Joseph, C. (in press). How Moral Foundations Theory Succeeded in Building on Sand: A Response to Suhler and Churchland. Journal of Cognitive Neuroscience.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture and morality, or Is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65,* 613-628.

Haley, J.O. (1986). Comment: The Implications of Apology. Law and Society Review, 20, 499.

Hamilton, V.L., Sanders, J., Hosoi, Y., Ishimura, Z., Matsubara, N., Nishimura, H., Tomita, N. & Tokoro, K. (1983). Universals in Judging Wrongdoing: Japanese and Americans Compared. American Sociological Review, 48(2), 199 - 211.

Hamilton, V.L., Sanders, J., Hosoi, Y., Ishimura, Z., Matsubara, N., Nishimura, H., Tomita, N. & Tokoro, K. (1988). Punishment and the Individual in the United States and Japan. Law & Society Review, 22(2), 301 - 328.

Hamlin, J.K., Wynn, K. & Bloom, P. (2007). Social evaluation by preverbal infants. Nature, 450, 557 – 559.

Hamlin, J.K., Wynn, K., Bloom, P, and Mahajan, N. (under review). Third-party reward and punishment in infants and toddlers.

Ingram, G.P.D. & Bering, J.M. (2010). Children's Tattling: The Reporting of Everyday Norm Violations in Preschool Settings. Child Development 81(3), 945 – 957.

Jost, J.T., Glaser, J., Kruglanski, A.W. & Sulloway, F.J. (2003). Political Conservatism as Motivated Social Cognition. Psychological Bulletin, 129(3), 339 – 375.

Krebs, D.L. (1975). Empathy and altruism. Journal of Personality and Social Psychology, 32, 1134 - 1146.

KZTV10 (14 April 2011). *Kingsville Mother Pleads Guilty To Negligent Homicide*. Kingsville, Texas. Retrieved from http://www.kztv10.com/news/

Lewis, H. B. (1971). *Shame and guilt in neurosis*. New York: International Universities Press.

Niedenthal, P., Tangney, J. P., & Gavinski, I. (1994). "If only I weren't" versus "If only I

hadn't": Distinguishing shame and guilt in counterfactual thinking. *Journal of Personality and Social Psychology, 67,* 585–595.

Nisbett, R.E. (2003). The Geography of Thought: How Asians and Westerners Think Differently…and Why. The Free Press, New York, New York.

Rai, T.S. & Fiske, A.P. (2011). Moral Psychology Is Relationship Regulation: Moral Motives for Unity, Hierarchy, Equality, and Proportionality. Psychological Review, 118(1), 57 - 75.

Rozin, P., Lowery, L., Imada, S. & Haidt, J. (1999). The CAD Triad Hypothesis: A Mapping Between Three Moral Emotions (Contempt, Anger, Disgust) and Three Moral Codes (Community, Autonomy, Divinity). Journal of Personality and Social Psychology, 76(4), 574 - 586.

Sarhan, A. & Davies, C. (Sunday 11 May 2008). *My daughter deserved to die for falling in love*. The Observer. Retrieved from http://www.guardian.co.uk/world/2008/may/11/iraq.humanrights

Seattle Times (21 January 1998). *Peru Orders $29 Million Payment For Crash Victims*. Lima, Peru. Retrieved from http://community.seattletimes.nwsource.com/archive

Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The "big three" of morality (autonomy, community, and divinity), and the "big three" explanations of suffering. In A. Brandt & P. Rozin (Eds.), *Morality and health* (pp. 119–169). New York: Routledge.

Smith, R.H., Webster, J.M., Parrott, W.G. & Eyre, H.L. (2002). The Role of Public Exposure in Moral and Nonmoral Shame and Guilt. Journal of Personality and Social Psychology, 83(1), 138 – 159.

Sosis, R. & Bressler, E.R. (2003). Cooperation and Commne Longevity: A Test of the Costly Signaling Theory of Religion. Cross-Cultural Research, 37, 211 – 239.

Stotland, E. (1969). Exploratory investigations of empathy. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 4, pp. 271 - 313). New York: Academic Press.

Suhler, C.L. & Churchland, P. (in press). Can Innate, Modular "Foundations" Explain Morality? Challenges for Haidt's Moral Foundations Theory. Journal of Cognitive Neuroscience.

Synnott, A. (1992). Tomb, temple, machine and self: the social construction of the body. The British Journal of Sociobiology, 43(1), 79 - 110.

Tangney, J. P., Miller, R. S., Flicker, L., & Barlow, D. H. (1996). Are shame, guilt, and embarrassment distinct emotions? *Journal of Personality and Social Psychology, 70,* 1256– 269.

Walker, L.J. & Hennig, K.H. (2004). Differing Conceptions of Moral Exemplarity: Just, Brave, and Caring. Journal of Personality and Social Psychology, 86(4), 629 - 647.

Waytz, A., Epley, N., & Cacioppo, J. T. (2010). Social cognition unbound: Psychological insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science, 19*, 58-62.

Wegner, D. M., Erber, R., & Raymond, P. (1991). Transactive memory in close relationships. *Journal of Personality and Social Psychology, 61,* 923-929.

Wright, J.C. (in press). The Role of Cognitive Resources in Determining Our Moral Intuitions: Are We All Liberals at Heart? Journal of Experimental Social Psychology

Young, L., Chakroff, A., Dungan, J., Koster-Hale, J. & Saxe, R. (submitted). When we are mindful of moral minds?

Young, L., Saxe, R. (in press). The Role of Intent for Distinct Moral Domains. Cognition.

Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia, 47,* 2065–2072.