4-18-2019

The Limits of Sociality

Johnna B. McGovern
johnna.mcgovern@gmail.com

The Limits of Sociality

Johnna B McGovern

B.A. Philosophy, Saint Joseph's University 2016

A Thesis Submitted to The Graduate School at the University of Missouri-St. Louis
in partial fulfillment of the requirements for the degree
Master of Art in Philosophy

May
2019

Advisory Committee

Lauren Olin, Ph.D.
Chairperson

Jon McGinnis, Ph.D.

Gualtiero Piccinini, Ph.D.

> We had a victory over fascism in Germany. It's time we had a victory over racism at home. I love this game. I love baseball. Given my whole life to it. Forty odd years ago, I was a player coach at Ohio Wesleyan University. We had a Negro catcher. Best hitter on the team. Charlie Thomas. Fine young man. I saw him laid low, broken because of the color of his skin, and I didn't do enough to help. Told myself I did, but I didn't. There was something unfair at the heart of the game I loved, and I ignored it. But a time came when I could no longer do that. You, you let me love baseball again. Thank you. (*42: The True Story of an American Legend*)

Branch Rickey is the lesser known name behind the integration of American baseball. On April 15th players around the League wear 42 to commemorate the triumphs of Jackie Robinson, but if it weren't for Rickey's role, on that day in 1947 Robinson never would have taken the field. It is a testament to Robinson's character, his skill, and his temperament that he was the player chosen to take this step in the civil rights movement. It is a testament to Rickey's that he was the manager to give him that chance.

42: The True Story of an American Legend serves as a fictionalized retelling of Jackie Robinson's rookie year as the first African American in baseball. The above quote comes from a scene where Rickey explains to Robinson why he signed him to the Brooklyn Dodgers, why he decided to change baseball. The explanation, like much of the movie and history, refers to the role African Americans played in World War II in undermining widespread racial prejudice. This undermining role has been investigated by experimental psychology, where implicit racism has been demonstrably weakened in research subjects exposed to pictures of reputable African Americans (Joy-Gaba and Nosek).

That phenomenon gets cast into the general category of implicit sociality by John Doris. In *Talking to Ourselves: Reflection, Ignorance, and Agency*, Doris attempts to shift the focus of theories about humans' characteristic nature from an emphasis on capacities for reflection to an emphasis on human sociality. Doris argues that sociality, both implicitly and in the form of collaborative reasoning, is what makes humans best equipped for moral improvement. This collaborativism possesses a defining role in his account of agency and responsibility. The aim of this paper is to gain an understanding of how sociality affects moral behavior and to determine whether it is conducive to agency in the way that Doris hypothesizes.

The paper advances in three stages. First, I will provide an exegesis of what I take to be the three foundational aspects of Doris' account of agency and responsibility: value-expressive behavior, collaborativism and currentism. The account begins by identifying agential behavior as self-directed and recognizing values as the internal feature of the self which sources this self-direction. The valuational account of agency is filled in by collaborativism and currentism. Collaborativism is a mechanism that facilitates our moral improvement by helping individuals to better determine what they should value and how to express those values in their behaviors. Currentism makes the focus of inquiry an individual's current valuational states, not the etiological history of those states.

Doris accepts that currentism is subject to skeptical challenges. If values, or any given agency-grounding inner state, are deeply historical and unshakeable, they fail to be expressive of self-direction. On Doris' account, currentism and collaborativism are importantly interconnected. Sociality is a means for revising one's beliefs to resist this skeptical challenge. While Branch Rickey and the experiments on implicit racism echo

this possibility, consider another scene from *42*. The movie fictionalizes a meeting between the general manager, Branch Rickey, and one of the players, Bobby Bragan, regarding his refusal to play with Robinson. Bragan's character refers to the fact that his friends back home in Mobile, Alabama would never forgive him for playing with an African American. Bragan would feel guilty playing with Robinson and that guilt motivates him to refuse to play and go so far as to request a trade. This guilt refers explicitly to the social context which triggers it, namely the Jim Crow South. Bragan is an example implicit sociality not improving an individual's values and value-expressive behavior. In fact the sociality further exacerbates his morally objectionable values and behavior.

We can ask: what explains the difference in outcomes in the cases of Branch Rickey and Bobby Bragan? I argue that the normative claims Doris maintains about sociality are made based on examples that fail to generalize. Particularly, I demonstrate that while it is possible that sociality facilitates improved moral behavior, it does not reliably do so in the vast majority of cases. I develop this argument in the second and third parts of the paper. First, I introduce the norms literature in order to argue that (1) sociality inculcates us with a highly consistent set of values through mechanisms for norm acquisition. This acquisition process is problematic for agency, on Doris' account, if it turns out that individuals are stuck with the historically-grounded sets of norms, and are unable to revise or adapt their values. Next I review empirical moral psychology to understand the maintenance of values after acquisition. I argue that (2) sociality does not have an easy route to revising the acquired set of values due to confirmation bias, the strength of our moral convictions and the difficulties these factors raise for individuals

recognizing and resolving moral dilemmas. I take these phenomena to show that Branch Rickey is an exemplar as an agent, and in the vast majority of cases individuals will fail to revise their values. Beyond the examples, the argument demonstrates that because (1) and (2) are the case, values are not self-directed in the way agency requires. Accordingly, Doris' currentist, collaborativist, valuational account of agency and responsibility is in need of substantial revision, or amendment.

I. Three Aspects of Doris' Account of Agency and Responsibility

In *Talking to Ourselves*, Doris has two main goals: (1) to argue that accounts of reflective agency are vulnerable to skeptical defeaters that arise from psychological incongruence and (2) to motivate a collaborativist, valuational account of agency.[1] This paper is focused on his latter goal. In this section, I elaborate on the three fundamental aspects of this account: value-expressive behavior, collaborativism, currentism.

i. Value-Expressive Behavior

Despite taking aim at the reflectivism at the heart of our more traditional accounts of agency, Doris' overall goal is largely conservative. In establishing his account, he considers our current practices for and instances of attributing agency and responsibility for guidance. Reactive attitudes are the foundation of his approach. Taking from P. F. Strawson's account, moral responsibility is attributed when someone is appropriately

[1] Reflectivism appears in many philosophical disciplines in Western philosophy. Doris addresses reflectivist accounts of agency whereby "the exercise of human agency consists in judgment and behavior ordered by self-conscious reflection about what to think and do" (Doris 19). The problem arises when he considers that this exercise requires accurate reflection, but that psychological research undermines the possibility of individuals reliably fulfilling this requirement. Having raised skeptical concerns for accounts of reflective agency— accounts that makes up a large portion of the literature— Doris lays the groundwork for his currentist, collaborativist valuational account of agency.

subject to reactive attitudes, such as anger and contempt (Doris 23). Doris argues that attributions of responsibility are appropriate in cases where individuals are acting agentially (159). An individual acts as an agent, on his account, when her actions are self-directed. However, because Doris has argued against using rational, reflective capacities as the source of agential behavior, he needs a different internal feature of the self to do the job. Doris finds that humans experience reactive attitudes towards the actions of another individual in cases where that individual appears to be directed by their values. Accordingly, on this account behavior is self-directed when it is expressive of an individual's values (25).

The first fundamental aspect of Doris' account is the source of agential behavior. When an individual's behavior expresses their values, then the individual is acting as an agent. More should be said of values, and to do so Doris takes an indirect path. He addresses an understanding of values according to the maxim *where there is smoke, there is fire*. If we can identify an individual's desires, we can understand their values. However, not just any desires will do the trick. Doris gives five criteria for the type of desires which speak to an individual's values: strength, duration, ultimacy, justificatory role, non-fungibility (Doris 27-28). Appropriate desires are not fleeting. And, importantly their objects are desired for their own sake, not instrumentally. Without necessarily being aware of its influence at the time an action is undertaken, the appropriate desires will always be referenced wherever such justifications are called for. The objects of these desires should also be irreplaceable, otherwise an individual can't be said to have incurred a loss in having the satisfaction of her desire frustrated.

A quick example will help to elaborate. My desire for a bag of Skittles while standing in the Wegmans' checkout line seems to be a poor candidate for indicating value. I typically feel the desire on a whim. I don't want the Skittles as much as I want a kick in my blood sugar after a long day of work and errands. I certainly didn't go to Wegmans to fulfill the desire, and if I did eat the Skittles, I'd be more likely to feel guilty for my action than be inclined to defend my behavior. Finally, I would not count it as a loss if I opted to eat one of the apples I purchased instead of the Skittles. Comparatively, my desire to maintain a healthy lifestyle is a more obvious candidate to fit the criteria for indicating value.[2]

ii. Collaborativism

Doris' arguments against reflectivism also leave room for him to shift the traditional focus of humans as reflective creatures towards a new target— human sociality (Doris 109). Doris made reflectivism the target of the first half of *Talking to Ourselves*. Particularly, he identifies psychological phenomena that undermine accounts of reflective agency that claim that individuals reliably, consciously and effectively reflect about the behavior they will undertake. Reflectivism and collaborativism are not diametrically opposed, rather the implication of this shift is simply that more emphasis should be placed on the social nature of humans. What is directly being challenged is individualism, particularly in regards to how humans reason best. Accordingly, the account of agency and responsibility is not only valuational, but importantly collaborative.

[2] The relation of behavior and value should be stressed. If I were a diabetic experiencing low blood sugar, eating a high sugar food like Skittles is a more likely candidate for indicating value, particularly a value like health.

When being collaborative, Doris contends, individuals will better be able to determine what they should value and how to express those values in their behaviors than they would by reasoning and reflecting alone. Collaborative reasoning is one way that sociality facilitates this moral improvement. Doris supposes that sociality can also work implicitly to foster such improvement and supports this supposition by reviewing empirical work on the social roles of moral emotions, which indicates that moral emotions motivate those experiencing them to behave in ways that facilitate moral improvement (Doris 122).

Doris is not simply arguing that moral emotions motivate behaviors of moral interest, but that it motivates better moral behavior. In this way he is making a normative claim. Examples are abundant, and Doris provides an interesting comparison class. Take guilt. A normal individual who hurts someone and feels guilty may be motivated to avoid such behaviors in the future. Compare this to the extreme example of a serial killer, Ted Bundy, who does not feel emotions like guilt and lauds his ability to avoid such "social control mechanisms" that might interfere with pursuing his desire to hurt others (122).

The other way sociality facilitates moral improvement is through collaborative reasoning. Doris argues that collaborative reasoning is optimal reasoning. To begin, he reviews literature on the aversive social conditions characteristic of the lives of clinical narcissists, psychopaths, and isolated incarcerated individuals to motivate the idea that the capacities required for good reasoning are both developed and sustained in overwhelmingly social contexts (Doris 111-3). These claims about practical and moral reasoning support the idea that a certain measure of sociality is to some extent necessary for reasoning at least insofar as it is required to maintain good cognitive health (114).

Doris goes on to defend the stronger claim that not just normal but optimal reasoning is likewise socially embedded (Doris 115). He turns to studies which compare reasoning in groups and individuals, but the results are inconclusive. Individuals often had more positive output both qualitatively and quantitatively, but groups were better at filtering ideas to correct for error as well as synthesizing information into generalizations like rules (Doris 117). The filtration processes are what Doris is concerned with highlighting as collaborative. Apart from the lab, he considers the success of group deliberation in academia. Research, he surmises, when subjected to inclusive examination, benefits from feedback diversity by challenging individuals to entertain more than the partialities of their discipline. Feedback diversity helps individuals to both recognize better directions to take research and increase the exposure of a claim to a broader range of objections (118).

The point extends from academia to Doris' target: moral reasoning. The strongest support he has for collaborative moral reasoning is deliberative polling. An activity of deliberative democracy, in deliberative polling, "participants are surveyed before and after a weekend of moderated small group discussion and plenary discussion with expert panelists" (Doris 120). The discussions are efficacious, evidenced by shifts in positions on specific policy questions. The idea is that the better-informed reasoning undertaken under this structure of discussion leads to better reasoning. At the very least, we can expect that these judgments "could better withstand critical scrutiny than unconsidered political opinions" (121). The very least seems to be what Doris needs to show in order to motivate the idea that our reasoning in groups is more optimal than in the context of individual reflection.

The aim of this argument is normative. He's not only making the claim that humans reason collaboratively, but that our reasoning is better when we do so. Doris takes the two examples he does because they are descriptive cases that support the normative claim. First, he recognizes that in academic research, results are better when work is subjected to inclusive examination. Second, he said that deliberative polling engenders discussions that lead to individuals reconsidering and better defending their political positions. These cases should generalize, such that in everyday life our dialog within our social groups leave us better able to determine (1) what we, as individuals, should value and (2) how to express those values in our behavior.

iii. Currentism

The final fundamental aspect of the account comes from an intuition standoff between deep historicism and currentism. In other words, how much, if at all, should the etiology of values be considered? Doris will favor a currentist approach, but to understand why we must understand the dilemma posed by the two possibilities. Consider the proverbial JoJo, the dictator:

> JoJo is the favorite son of Jo the First, an evil and sadistic dictator of a small, undeveloped country. Because of his father's special feelings for the boy, JoJo is given a special education and is allowed to accompany his father and observe his daily routine. In light of this treatment, it is not surprising that little JoJo takes his father as a role model and develops values very much like Dad's. As an adult he does many of the same sorts of things his father did, including sending people to prison or to death or to torture chambers on the basis of whim. He is not coerced to do these things, he acts according to his own desires. Moreover, these are desires he wholly wants to have. When he steps back and asks, "Do I really want to be this sort of person?" his answer is resoundingly "Yes," for this way of life expresses a crazy sort of power that forms part of his deepest ideal. (Wolf 462).

If we take a currentist approach, the focus is on the instance of action. JoJo on a currentist, collaborativist, valuational account would be considered a responsible agent.

His behavior is self-directed as it expresses his value for power. A value indicated, based on the vignette, by desires that meet the appropriate criteria discussed in an earlier section. This result, however, seems counterintuitive. The conditions of JoJo's upbringing should make us uncomfortable attributing him the degree of agency required for responsibility (Wolf 462; Doris 30). The problem is that values are indicative of agency because they are an internal feature of the self. JoJo seems to have been indoctrinated with a set of values that he can't help but to have acquired. Accordingly, rather than indicate self-direction, his value-expressive behavior seems to be sourced externally in his upbringing.

Doris attempts to largely conserve our current practices and instances of attributing agency and responsibility. To heed the intuition that JoJo's values are out of his control, and accordingly not indicative of the self-directed behavior necessary for agency, one needs to take seriously that his values are deeply historical. Such an approach, however, invites skepticism. Particularly, if JoJo's upbringing has situated him with a set of values beyond his control, why should we feel any more in control of ours? Accordingly, to accommodate our intuitions that JoJo is not responsible, we become vulnerable to retracting most attributions of agency and responsibility (Doris 32).

Doris claims that cases like JoJo seem psychologically implausible; real cases of fixed, indoctrination without the type of psychological impairment that would undermine agential attribution are few and far between (Doris 32).[3] If we grant the possibility, however, Doris is willing to bite the bullet. Through collaborativism, most individuals are

[3] Doris does not specifically engage the example of JoJo, but treats the objection through other examples.

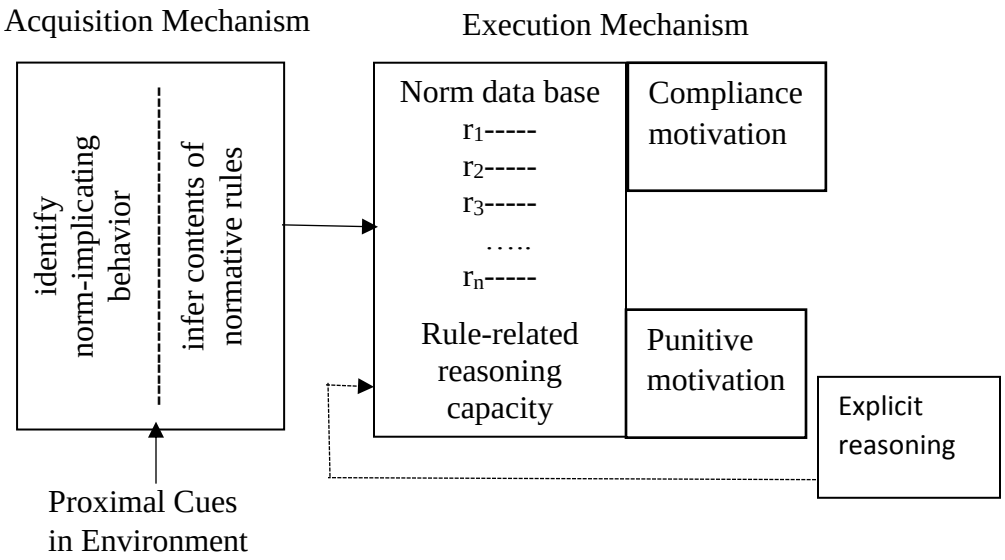
amenable to changing their values, and sociality is a reliable route toward doing so (29). An attribution of agency in particular cases need not consider the etiological history of how collaborativism has actually shaped the individual's values and value-expressive behavior. Rather, for Doris' account, collaborativism must be a reliable path for revision, such that regardless of *whether* an individual has taken that path, the opportunity was available. Accordingly, passively acquired values would not be unshakable, and current valuational states would be internal features of the self. In this sense, on a currentist, collaborativist account, value-expressive behavior maintains its status as self-directed behavior and is an appropriate candidate for attributions of agency and responsibility.

II. Norms and Sociality as Indoctrination

JoJo was a bullet worth biting in part because Doris maintains that the way his upbringing indoctrinated him with values resistant to change is not similar to the way most people acquire their values. I use the account from Chandra Sripada and Stephen Stich's "A Framework for the Psychology of Norms," as an example to establish that research supports the claim that in the wild sociality facilitates value acquisition in a similar way to the JoJo case. While most of us have cultural contexts that are much more mundane than totalitarian regimes, we share with JoJo a passive cultural inheritance of a set of local values. Further we complete this acquisition at a young age. I proceed in two stages. First, I argue that norm-guided behavior is indicative of value-expressive behavior. Second, I rely on their account of norm acquisition to demonstrate one way sociality operates in the wild. My conclusion is that sociality inculcates individuals with a highly consistent set of values by a young age.

Before beginning a review of the Sripada-Stich account, I wish to introduce their mechanism for norm acquisition and execution. This figure will be a useful reference in this discussion as well as in the later discussion of sociality's revisionary capacities.

Figure 1: The Sripada-Stich mechanism for the psychology of norms (Sripada and Stich 17).[4]



i. Where there is Smoke, there is Fire

On Sripada and Stich's account, norms are, broadly speaking, social rules, but the authors note three features that distinguish what norms are as a natural kind of rules (Sripada and Stich 2). First, they are independently normative. While there may be laws or regulations which explicitly enforce a norm, there need not be for compliance. Second and relatedly, norms are intrinsically motivating. People choose to follow them for their own sake

[4] I have excluded other boxes from the mechanism that are periphery to our discussion. Further the dotted line from explicit reasoning to the norm database indicates that this relation is a reputable speculation and should be treated with less certainty than relations depicted with solid lines.

rather than as a means to an end. Finally, transgressions against norms cause others to direct punitive attitudes, like anger and condemnation, toward the transgressor.

Consider an example that Doris uses (Doris 112). There is a norm in our society that people wait their turn in line. When waiting in the line at Starbucks, there aren't signs posted enforcing the rule that you not cut the line. However, unless someone absentmindedly misunderstands how the queue is formed, there is rarely an individual that takes it upon themselves to skip to the front of the line. Everyone simply follows the rule. When someone does violate the norm, the transgression is met with the rolling eyes, scoffs and possible snarky comments that communicate anger towards the transgressor.

There is something intrinsically motivating about norms and it works to manifest two of the features by which we recognize them. We are intrinsically motivated both to comply with norms as well as to feel certain punitive emotions when a norm has been violated (Sripada and Stich 11). Colloquially, this motivational force is explained in a self-evident way. Jerry Seinfeld may iterate to George that "It's just not done in polite society. It's not done in impolite society" (*Curb Your Enthusiasm*). In other words, we don't find a need to explain further the wrongness of the norm transgression. But there is still an interesting question. What aspect of the norm gives it this intrinsically motivating force?

To better understand the nature of the motivation, Sripada and Stich want to go beyond philosophical intuition and find an empirical basis for their claims. One of the most interesting reviews of compliance comes from the study of anonymous, one-off economic games (Sripada and Stich 11). In the games, players can choose either to cooperate for a lower payoff, or defect, where defecting will always maximize their own

gains. The fact that the players will play both anonymously and only one time removes the threat of retribution. However, studies show that players routinely choose to cooperate. The intrigue of this study is that its design makes it tough to interpret the seemingly altruistic results in an egoistic way (10).

There are parallels to these cases in the wild as well. A paradigmatic example is the fact that people uphold tipping norms while travelling. Tipping norms can be considered instrumentally motivating, at least insofar as they foster relationships between service staff and regulars and reinforce standards of good service for future visits. However, when travelling, individuals are relatively anonymous, and so do not have reasons to consider how their good behavior might be reciprocated in future interactions. Were norms not intrinsically motiving, it would be hard to explain why individuals uphold norms, like tipping, in anonymous, one-off situations.

The literature reviewed on the intrinsic motivation to punish norm transgressions follows the same theme of favoring cooperation over self-interest (Sripada and Stich13). In fact, the motivation to punish has been shown to persist even if it negatively affects the self-interest of the individual administering punishment (15). The authors stress that while this seems interesting in relation to a literature which has long stressed rational self-interest as a human motivator, they think that the in the wild experience of "third-party punishment" shows that the maintenance of this intrinsic motivation in the face of rational self-interest is "fairly obvious and unsurprising", if not necessary (15). On their view, the experience of moral emotions like anger or contempt, emotions indicative of punitive attitudes, are just enough to signal this intrinsic motivation (12).

With this empirical background, in place, it seems reasonable to suppose that *values* give norms their intrinsically motivating force. Consider, for instance, the following example:

> Michael is in a three-year relationship with Kelsey. Michael is in advertising and has been working on a project with his colleague Jenna. The two have really bonded and on multiple occasions have found themselves feeling strong attractions to one another. After a week of late nights at the office, Michael suggests the two go out for a drink. That night Michael commits infidelity with Jenna.

Upon reading the vignette, I assume most people would react with anger or disgust towards Michael's behavior. At the very least, we can imagine that Kelsey will react with some punitive attitudes when she discovers Michael's infidelity. These attitudes would be the expected results were we to consider the situation in light of either the Sripada-Stich account of norms or in light of Doris' account of agency.

First, consider where those emotional reactions come from on the norm-guided behavior account. There are a set of norms associated with long term monogamous relationships whose contents vary based on the cultural group to which individuals belong. While the specifics may vary, norms typically prohibit infidelity. We can assume for the sake of the imagined scenario that Michael and Kelsey are operating under a prohibitive norm. Accordingly, even if Michael were in a situation where Kelsey would not find out, we would still expect that Michael would not cheat on her. His behavior engenders the reaction from Kelsey and third-parties because transgressions against norms intrinsically motivate this type of reaction.

Next, we can note that Doris works via a different method than Sripada and Stich. He begins with the fact that we attribute responsibility via punitive attitudes, and from

there addresses the type of behaviors that elicit such reactions. For Doris, such attributions are appropriate when we expect that individuals are acting agentially, in ways that express their values. In a way, Michael's behavior can be said to fail to express his value for Kelsey or for his relationship with Kelsey. Kelsey experiences a punitive attitude, namely anger, toward Michael. His behavior expresses that he valued a sexual-experience with his colleague Jenna more than he valued his relationship with Kelsey. On Doris' account, this value-expression makes Michael an appropriate target for such punitive attitudes.

On Doris' account, desires serve as a marker for understanding what an individual values, but the values themselves are still the main psychological feature of interest. This example draws the obvious parallel that norms engender punitive responses in much the same way as Doris expects that value-expressive behavior should. Further, in the examples he gives, the object of the ultimate desire tend to be the values under consideration (Doris 27). In this regard it is at least plausible to say that the values themselves are intrinsically motivating.

How can intrinsically-motivating values relate to intrinsically-motivating norms? Refer back to the mechanism proposed for the acquisition and execution mechanisms for norms in Figure 1. One such question that the authors consider is a concern for how moral beliefs and principles relate to the norms stored in the database (26). Values fit in the same category with moral beliefs and principles. Based on the authors' claims, it is possible values are the entries, or a subset of the entries, in the databases. The relationships that maintain between beliefs, principles, values and norms will make more sense as research advances. I have argued that values and norms are intrinsically

motivating in the same ways. Values can be understood to partly constitute the content of the norm database, and thereby help to account for the intrinsically motivating force of norms.

We can elucidate this through our earlier examples. In regards to the norm of standing in line, the intrinsic motivation for compliance can be explained in a self-evident way. We stand in line because that is the proper behavior when waiting to order at coffee shop. When pressed to truly explain the behavior— or why we are upset when it is violated— my intuition is that we would defer to a value, like fairness. Similarly, in the example of infidelity, if Kelsey were pressed to explain why she is angry, she will of course give the explanation that Michael cheated on her. But, the same explanation can also include her saying something like, *I thought I meant something to you*.

More can be said to motivate the relation between understanding a norm and understanding the value related to it. Consider a norm uncommon to Western culture. For example, in Bhubaneswar there is a norm that prohibits widowed women from eating fish. This norm is not in the databases of most Westerners, and their reaction is likely one of confusion. That confusion is alleviated when the value it expresses is made evident. In this case, it is disrespectful to their deceased husbands since fish is believed to be an aphrodisiac (Haidt 615). Coming to understand the underlying value alleviates some confusion.[5] This is further evidence for the relationship between the content of norms and the nature of values.

[5] This example points in a good direction for research to explore the relation. Sripada and Stich recognize that within a cultural group norms are not stagnant. The relation between norms and values can be further elaborated by looking at the causal history of norm evolution. Namely, do proponents of change address values or the norms themselves?

I have argued that it is at least plausible that values are the intrinsically motivating content of social norms; further empirical research will reveal their exact relation. For now, the similarities drawn between value-expressive behavior and norm-guided behavior is enough for a weaker claim. As stated earlier, Doris recognizes values by the maxim *where there is smoke there is fire*. By his own lights, desires are good indicators of value; for the same reasons I maintain that norms are good indicators of value.

ii. A First Pass at Sociality in the Wild

Consider again JoJo's upbringing:

> Because of his father's special feelings for the boy, JoJo is given a special education and is allowed to accompany his father and observe his daily routine. In light of this treatment, it is not surprising that little JoJo takes his father as a role model and develops values very much like Dad's. (Wolf 462).

We anticipate from reading this that JoJo is indoctrinated by a young age with a set of values largely contingent on his environment. The vast majority of people learn and develop in a much more ordinary and tolerable environment than JoJo. However, based on the Sripada-Stich account of norm acquisition, it is reasonable to believe that the process results in the acquisition of highly consistent network of norms and values in an individual by a young age. With this account, we begin to approximate how Doris' valuational account of agency and responsibility fails. I argue that the norms literature demonstrates that sociality does facilitate the acquisition of norms and values, and that individuals are inculcated with sets of norms and values in a way that is beyond their control.

To begin, Sripada and Stich provide some social-level facts about norms. Norms are culturally universal. All human societies have norms and have had norms in a deeply historical sense (Sripada and Stich 3). Those norms often share thematic similarities. For example, nearly all societies will have norms that prohibit killing and promote equality (6). However, while cultures share thematic commonalities between types of norms, the contents of those norms are culturally variant (4). The cultural variance is explained by their pattern of ontogenesis. Individuals will reliably acquire the norms of their cultural heritage (7).

These social-level facts give us the grounds to understand the acquisition mechanism in Figure 1. The acquisition mechanism is based on research surrounding the cultural ontogenesis of the norm. Most humans acquire the norms of their local cultural group (Sripada and Stich 7). More particularly, in researching the diversity of norms between cultures, differences are established to an equal degree in children as adults (8). This means not only do most individuals culturally inherit their norm database, but they do so by a young age. The acquisition proceeds through a filtration process of absorbing the cues of one's environment, recognizing the at-issue behaviors and inferring the contents of the norms associated with them (16).

The above process is presumed to be automatic and involuntary, but that does not diminish the obvious role of sociality (Sripada and Stich 16). Children learn norms by observing and interacting within their social groups. The transmission won't always follow perfectly from cultural parents to child, though. One possibility is a copying error, in which the child acquires the wrong content but retains the norm mutation because it is a more favorable rule (32). This possibility is less obviously related to sociality. A second

possibility acknowledges that social group won't be completely homogenous. Conformity bias describes the case when individuals conform to norms held by the majority of the cultural group; this may cause the acquisition of the predominant norm in the group even if it isn't shared by the individual's direct cultural parents. Prestige bias may cause the acquisition to model off prestigious members of the group (34). Further research would be required to explain how the biases relate to the acquisition of norms. For now, the Sripada-Stich account shows that the process of norm acquisition is socially embedded, determined by one's local cultural group and complete by a young age. The characterization of this process demonstrates that, like JoJo, the acquisition of sets of norms and values (even in ordinarily tolerable environments) are beyond any individual's control. Further, I take this as an example of how sociality in the wild operates in the passive cultural inheritance of sets of norms and values whose members demonstrate a high degree of consistency.

III. The Limits of Sociality's Revisionary Capacity

This paper began with two stories from the integration of baseball. Branch Rickey was the manager who signed Jackie Robinson and explained that decision with reference to his reevaluation of racial prejudice after serving with African Americans in World War II. Bobby Bragan was a player on the Dodgers who refused to play with Robinson and explained that decision with reference to the guilt he felt thinking about how his friends back home in Mobile would feel. Both are examples of the operation of implicit sociality in maintenance of values. We can now return to the initial question that I raised: What makes sociality facilitate moral improvement in the case of Rickey but exacerbates morally objectionable values and behavior in the case of Bragan? In answering this

question, I will demonstrate that sociality fails to reliably revise highly consistent sets of values acquired by passive cultural inheritance in the vast majority of cases, even if moral exemplars like Branch Rickey exist. Such a charge is detrimental to the combination of the currentist and collaborativist features of Doris' valuational account.

i. The Implicit Role of Sociality Reconsidered

Doris claims that moral emotions motivate those experiencing them to behave in ways that improve moral behavior (Doris 122). But Bragan— at least his fictional representation— refuses to play with Jackie Robinson and requests that Branch Rickey trade him. Bragan says his friends back home in Mobile, Alabama would never forgive him for playing with an African American. Bragan would feel guilty playing with Robinson and that guilt motivates him to refuse to play and go so far as to request a trade. This guilt refers explicitly to the social context which triggers it, namely the Jim Crow South. The guilt of Bragan is not an example of moral improvement. His experience of guilt exacerbates his bad morality and bad moral behavior.

In his analysis of this implicit role of sociality, Doris does not focus on how socially-contextualized emotions motivate improvement of moral behavior. The connections I have drawn between his valuational account and the norms literature can make this clearer. Cases involving guilt have an obvious connection because it is a moral emotion which serves as a punitive attitude.[6] A transgression against a recognized norm

[6] Cases of other moral emotions, like empathy and sympathy, also have a role in how norms compared to conventional rules are understood as more serious, less permissible and authority independent which is explored by Shaun Nichols. These criteria plausibly lend themselves to an explanation of the intrinsically motivating features of norms (Nichols 2004).

should intrinsically motivate a feeling of guilt according to the Sripada-Stich account of norms. This guilt should further motivate us to do a better job of conforming to our accepted norms in the future.

Bobby Bragan's guilt operates psychologically appropriately then. Bragan has norms and values which recognize a white superiority to black Americans and enforce racial segregation. Related norms are recognized explicitly in the fictionalized account of the Jackie Robinson story, including a scene where he is kicked off the field in a Southern state by law enforcement because "No [African American] is going to play with white boys" (*42*). According to his own body of norms, it is wrong for Bragan to play with Robinson. The prospect of transgressing against this norm engenders guilt in Bragan. That guilt motivates him to refuse to play and request the trade.

On the Sripada-Stich account, individuals have a norm database which intrinsically motivates (1) compliance behaviors and (2) punitive attitudes towards transgressions. When an individual experiences guilt, the case may be that they are transgressing against one of the norms in their database. The motivated behavior following the transgression, however, will only be judged as an improvement if those judging agree with the individual's norm database. Insofar as the individual's norm database represents a bad morality, we will not judge the moral emotions serving as punitive attitudes to be improving their moral behavior.

The implicit role of sociality in the case of moral emotions fails to be a reliable means for facilitating moral improvement. Recall that it is important for Doris' account that sociality can reliably revise an individual's set of values. If not, we are stuck with a set of values that seem out of our control and, accordingly, inappropriate candidates for

the sorts of inner features the expression of which facilitates agency. Based on this fact, the problem posed by Bobby Bragan is serious; it shows that the efficacy of sociality reliably facilitating moral improvement, at least at times, is contingent on the acquired set of values an individual already holds. Such a mechanism of sociality should not make us anymore confident that our value-expressive behavior is self-directed.

ii. Collaborative Reasoning and Moral Dilemmas

Doris' account does not live and die by the implicit social role of moral emotions. If we refer back to the mechanistic representation of the psychological framework of norms that Sripada and Stich present (Figure 1), there was an additional causal arrow indicating the reputable speculation that explicit reasoning can correct inconsistencies in the norm database. Accordingly, collaborative reasoning may still prove to be a reliable means for making individuals better able to determine what they should value and how to express those values in their behavior. In other words, it may still be a reliable means for revising the set of values. This possibility is important because in our review of sociality so far, an individual's values seem largely beyond their control. I turn to empirical moral psychology in order to evaluate evidence for this possibility. I argue that research findings allow that individuals who internalize moral dilemmas have access to revising any inconsistencies in their set of values, but that factors like conformation bias and the strength of our moral convictions make it difficult for individuals to actually do so. Accordingly, while possible, it is implausible that sociality can reliably revise an individual's set of values in the vast majority of cases. In part with passive inheritance of sets of norms and values, this conclusion shows that those sets are largely out of the

control of individuals, and thereby Doris' account fails to ground agency on an internal feature suitable for self-direction.

Before beginning my own analysis of collaborative reasoning's aptness (or lack thereof) for facilitating moral improvement, consider again the examples Doris uses to make his claim: academic research and deliberative polling. Doris explicitly recognizes that for both examples the right kind of group matters. He describes the academics as "a diverse population of motivated reasoners" (Doris 119). On deliberative polling, he reiterates "participants are, with appropriate structure, able to reason cooperatively (once again, not just any group will do)" (121). Certainly, people discuss and debate especially in regards to moral judgments, but how well do these examples generalize? Dinner table debates are neither moderated, nor accompanied by expert panelists like deliberative polling events. Further, at the neighborhood bar, the diversity will neither be high, nor will the motivated reasoning be focused on objective truth like academics is.

These examples present individuals in situations where reasoning is accepting of diverse opinion and directed toward objective truth. However, a lay reasoner in the wild is more likely motivated to preserve the consistency of her worldview— of her set of norms and values— than actively seek truth like these situations depict. Such a claim about consistency motivation is substantiated both by the type of maintenance mechanism considered possible on the Sripada-Stich account and by the best route proposed within the moral psychology literature for how sociality may effectively operate. I can now turn to research from Richard Campbell and Victor Kumar (2012), Zachary Horne, Derek Powell and John Hummell (2014), Raymond S. Nickerson (1998), and Linda J. Skitka (2010) to demonstrate this (albeit limited) route.

In "Moral Reasoning on the Ground," Richard Campbell and Victor Kumar develop an exemplar for how Doris' sociality can be maintained and fit as the type of maintenance mechanism Sripada and Stich consider possible. On their account, explicit reasoning and implicit affect work together "to expose latent inconsistencies embodied in conflicting moral judgments about cases that are, by their own light, similar in morally relevant respects" (274). This practice is called moral consistency reasoning. Let's take their example to understand the practice. Consider a Norwegian mother faced with the decision of sheltering a World War II Norwegian resistance fighter against the Nazis. Sheltering the boy leaves her open to the risk of their retaliation which could mean the death of her entire village. The mother has an initial feeling of responsibility to protect her village, but she is forced by her son to consider how she would like him to be treated if he were the young man in trouble. This creates a dilemma. Based on the former, she thinks she should turn in the fighter, while the later forces her to think she should protect him. To reconcile the views, she decides to abandon her intuitive sense of responsibility to the village and instead treats the boy how she would hope her son would be treated (273).

The account is promising for Doris' claim. Before evaluating how it should be restricted, a clarification should be made. There is an obvious difference between inconsistencies in moral judgment rather than moral values. Moral values need to be thought of as part of an internal set from which an individual's moral judgments are derived. This clarification is important, but we need not abandon the practice offered by Campbell and Kumar. In "Single Counterexample Leads to Moral Belief Revision," psychologists, Zachary Horne, Derek Powell and John Hummel, set up a similar account

whereby moral dilemmas prompt individuals to actively hold inconsistent moral beliefs and psychologically force them to resolve the issue. Accordingly, I can draw from the resources of either account while addressing the possibility of sociality as being reliably revisionary. But in doing so it is important to keep in mind that a stored set of highly consistent values is the relevant material when examining Doris' account.

The first restriction that should be noted is that the inconsistency must be latent within the set. Confirmation bias is a phenomenon by which individuals experience evidence in a way that confirms their beliefs.[7] While confirmation bias is a well-known phenomenon, it should be stressed that it is "unwittingly selective" (Nickerson 175). Whether an individual is motivated or neutral regarding a desire to maintain the status of their belief, the bias will operate in the same manner rejecting evidence or distorting it to conform to the belief. Accordingly, when a "dilemma" is posed to an individual who fails to internalize the reason for tension, the dilemma will be dismissed out of hand. We can expect the bias to act as a defeater for cases of inconsistency that are externally caused, i.e. responding to a situation differently than another person or having a dilemma forced by a belief you do not already hold (Campbell and Kumar 301; Horne 1951).

To relate this back to the set of highly consistent values, the set will be largely resilient to collaborative reasoning or appeals to implicit affect that fail to properly acknowledge internalized values. We can tweak the example of the Norwegian mother to show the import of confirmation bias. Imagine instead that the soldier who appeared at the house was not an injured resistance fighter but an enlisted deserter, and imagine it

[7] It may be helpful to distinguish that this is a separate phenomenon than conformity bias mentioned in the discussion of norm acquisition.

was him who asked "What if I were your son?" to challenge her intuitions. If the woman held common values against desertion, it is unlikely that she would have been forced into a dilemma. She likely wouldn't have regarded it possible that her son would act with such a lack of patriotism. The dilemma between protecting her community and protecting "her son" wouldn't be recognized.

I can return to the initial question posed in my introduction *Why does sociality facilitate moral improvement in Branch Rickey but not Bobby Bragan?* At this point I can diagnose how sociality helped Rickey determine both what he should value and how to express those values in his behavior. In the same way we are to expect that the Norwegian mother relates to the injured resistance fighter as similar to her son, Rickey reevaluates his relation to African Americans. Because of their service in World War II, Rickey recognizes that African Americans are a group he shares values with and revises the network of values associated with his previously held racial prejudices. To sum it up simply, he couldn't find it appropriate to share a battlefield with men that he wouldn't share a ballpark with.

Previously we analyzed how sociality facilitates the exacerbation of Bragan's racial prejudices, but more can be said of why it fails to facilitate improvement. We do know that Bobby Bragan is privy to the same information as Rickey; he knows that African Americans served in World War II. One possibility is that Bragan fails to recognize a dilemma. Like the Norwegian mother with the enlisted deserter, he may not internalize that information in a way that causes a dilemma. The dilemma between being willing to share a battlefield and not a ballpark wouldn't be recognized. Based on confirmation bias, we can assume if this is the case it is because being raised in the Jim

Crow South, Bragan's values are highly insulated from relating to African Americans. This is not the only way to interpret the case of Bragan. There are more restrictions to the revisionary capacity of sociality.

The second restriction that should be noted is that conviction makes some values more resistant to revision than others. On Linda J. Skitka's account, for an individual who seriously holds a moral conviction "to support alternatives to what is 'right', 'moral' and 'good' is to be absolutely 'wrong', 'immoral', if not 'evil'." (Skitka 267). This feature is somewhat obvious. If individuals didn't take their values seriously, transgressions wouldn't engender punitive attitudes, particularly toward non-group members (268). Conviction makes the given value resistant to change. A given set of values, however, is not a collection of unrelated members. Acquisition creates a network of interrelated values. Conviction may be strongest in a particular value, but on Skitka's account it will also motivate resistance to changing those related values as well. Further, it also supports defying the dictates of authority and/or conformity insofar as either are influenced by the opposing conviction (274).

The strength of our moral convictions can be positive, as I imagine people like Jackie Robinson needed it given the trials he faced in striving toward equality. In regards to revision, however, it (1) exacerbates the effects of confirmation bias and (2) acts as a defeater in moral consistency reasoning. Examples are abundant for the former, while the latter is more interesting. When a dilemma is suitably internal, epistemically an individual should be motivated to resolve the dilemma by revision of their values. However, if the moral convictions within the dilemma have the type of psychology indicated by Skitka, then revision isn't the likely option. Rather, the inconsistency can be

explained away by rejecting the way the values have been mapped into a dilemma (Horne 1972). Accordingly, even if a suitably internal inconsistency is recognized, it need not cause revision. Internal dilemmas with values backed by strong conviction will be rejected rather than revised.

A third restriction also concerns the reach of revisionary dilemmas, i.e. a domino effect. In Horne, et. al, the study's methodology invites some skepticism about the applicability of their conclusions. Participants were asked theoretical questions about organ transplants that posed dilemmas. The study aimed to understand their response to the dilemma both initially and after a 6-hour delay. The authors took consistency between the responses across the delay to indicate an internalized revision to the beliefs challenged in the dilemma. Answering a low stakes dilemma in a lab is different than interacting with dilemmas in one's environment. The results point to the possibility of revision, but as Campbell and Kumar recognize, even if an internal inconsistency appears, revision is tough (Campbell and Kumar 303). Any proposed resolution will lead to further inconsistencies since we are operating with an understanding that the set of values is an interrelated network. Revision becomes an onerous cognitive task when it causes a domino effect of reevaluating the related values. A lab setting may not inspire this level of revision consideration. But Campbell and Kumar take seriously that the onerous task of revising embedded values may lead to another case where dilemmas are rejected rather than resolved.[8]

[8] It is not discussed, but the rejections considered in the previous two restrictions seem to have the same type unwitting selection as confirmation bias is purported to have. Individuals aren't intentionally obstinate, but unwittingly so.

Considering the previous two restrictions, there is a second possibility for why sociality fails to facilitate improvement in the case of Bobby Bragan. Even if Bragan began to internalize a dilemma with his values relating to African Americans, this need not necessitate that he resolve the dilemma. Bragan grew up in 1920s Mobile where values and norms supporting white supremacy and racial segregation were ubiquitous. He internalized this set of values in a highly consistent network. The fact that African Americans served in World War II could be dismissed if (1) the conviction of Bragan's racial prejudice were strong enough to cause the rejection of the dilemma or (2) accepting the fact would cause a much too onerous revision of his network of values, leading again to the rejection of the dilemma. These possibilities, taken together with the earlier possibility that confirmation bias may cause Bragan to outright fail to internalize a dilemma, represent ways the set of values are resilient to change.

The restrictions leave little hope in what was already a small possibility for sociality. Two accounts grounded in empirical psychology and cognitive science proposed the similar scenario by which a dilemma, or inconsistency, would have to be recognized within an individual's internalized set of values for the possibility of revision. This narrowed the type of possible scenarios significantly but still seemed promising because the types of social processes that Doris is concerned with, explicit reasoning and implicit moral emotions, were recognized as possible revisionary routes. However, the possibility became increasingly restricted, such that confirmation bias, strength of convictions and difficulty of revision were presented as defeaters. I conclude that this evidence is enough to motivate that while sociality appears to be a route for revising our values, it does not do so reliably in the vast majority of cases.

Conclusion

The currentist feature of Doris' valuational account of agency and responsibility relies on sociality being a reliable means by which individuals can better recognize what they should value and how to express those values in their behavior. Without this collaborativist feature, the currentism leaves the account vulnerable to the skeptical charge that our values are historically-grounded and unshakable in a way that undermines agency. My arguments have targeted this collaborativist aspect of Doris' account. I engaged literature in empirical moral psychology to demonstrate that sociality in the wild does not reliably facilitate moral improvement in the vast majority of cases. With reference to the norms literature, I argued that individuals are culturally inculcated with a highly consistent set of values by a young age. I then turned to psychology and cognitive science to determine whether sociality can influence values in ways that reliably facilitate moral improvement. I argued that, due to confirmation bias, the strength of our moral convictions, and the difficulties these factors raise for individuals recognizing and resolving moral dilemmas, historically acquired sets of values are highly resistant to change.

What does this mean for Doris' account of agency and responsibility? First, a currentist, collaborativist, valuational account cannot properly source agency in the vast majority of cases. Agency requires self-directed behavior, and values are the source of self-direction on Doris' account. Because I have demonstrated that an individual's set of norms and values is historically-grounded and often impenetrable to revision, values in

the vast majority of cases fail to act as the type of internal feature Doris needs them to be in order to source agency.

Second, based on my argument, skepticism about agency and individual responsibility loom large. However, I take seriously Doris' attempt to shift the focus of theories about human's characteristic nature from an emphasis on capacities for reflection to an emphasis on human sociality. Accordingly, I do not feel hopeless in the face of skepticism about agency and individual responsibility. This discussion has left open an opportunity to recognize the impact of sociality in a fuller sense. Future research may need to be less conservative, less focused on building accounts that maintain current practices of agency and responsibility attribution. Instead, it might focus on the influence of sociality on individuals' values and behavior as well as engineering social institutions and practices that better put individuals in situations conducive for moral success.

Work's Cited

42: The True Story of an American Legend. Directed by Brian Helgeland, Warner Bros, 2013.

Bennett, Jonathan. "The Conscience of Huckleberry Finn." Philosophy, vol. 49, no. 188, 1974, pp. 123–134., doi:10.1017/s0031819100048014.

Campbell, Richmond, and Victor Kumar. "Moral Reasoning on the Ground." Ethics, vol. 122,no. 2, Jan. 2012, pp. 273–312., doi:10.1086/663980.

Doris, John M. Talking to Our Selves: Reflection, Ignorance, and Agency. Oxford University Press, 2015.

Haidt, Jonathan, et al. "Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?" Journal of Personality and Social Psychology, vol. 65, no. 4, 1993, | pp. 613–628., doi:10.1037//0022-3514.65.4.613.

Horne, Zachary, et al. "A Single Counterexample Leads to Moral Belief Revision." Cognitive Science, vol. 39, no. 8, 2015, pp. 1950–1964., doi:10.1111/cogs.12223.

Joy-Gaba, J.A. and B. A. Nosek. "The Surprising Limited Malleability of Implicit Racial Evaluations." *Social Psychology*, vol 41, 2010. pp, 137 146, doi:10.1027/1864-9335/a000020.

Nichols, Shaun. "Norms with Feeling." Sentimental Rules, 2004, pp. 3–29., doi:10.1093/0195169344.003.0001.

Nickerson, Raymond S. “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises.” Review of General Psychology, vol. 2, no. 2, 1998, pp. 175–220., doi:10.1037//1089-2680.2.2.175.

Skitka, L. J., “The Psychology of Moral Conviction.” Social and Personality Psychology Compass, no. 4, Mar. 2010, pp. 267-281., doi:10.1111/j.1751-9004.2010.00254.x.

Sripada, Chandra Sekhar, and Stephen Stich. “A Framework for the Psychology of Norms.” The Innate Mind: Volume 2: Culture and Cognition, 2007, pp. 280–301.,doi:10.1093/acprof:oso/9780195310139.003.0017.

“The Table Read.” *Curb Your Enthusiasm*. Home Box Office, Inc, HBO, 15, Nov. 2009.