Thrasymachus's Blush:
Between the Science of Reason & the Politics of Emotion

William Minozzi*
Michael A. Neblo

Department of Political Science
Ohio State University

March 2nd, 2015

William Minozzi, Associate Professor, william.minozzi@gmail.com
Michael A. Neblo, Associate Professor, neblo.1@osu.edu.

"Thrasymachus conceded all these points, but not in the easygoing way I have just described. He had to be dragged every step of the way, sweating profusely, as you might expect in summer. This was the occasion when I saw something I had never seen before—Thrasymachus blushing." Plato, *Republic* 350d

Plato's Thrasymachus is the first great "realist" in the history of western moral and political thought. He famously defines justice as the advantage of the stronger. Less famously, though just as importantly, he goes on to argue that as an empirical matter, the rhetoric of reason and justice actually tends to further the interests of the powerful. Versions of Thrasymachus's claims have evolved and echoed down through the history of political thought, all the way to contemporary discussions about the efficacy of practical reason and public deliberation. Empirical research on so called "motivated reasoning," "social intuitionism," and "affective primacy"[1] appears to support such pessimism by demonstrating the ways in which people's desires, emotions, and ideological commitments drive the way that they present, assimilate, and process arguments and information. Recent research has radicalized these results, arguing that such phenomena are so predominant as to leave notions of public reason – and, indeed, most of democratic theory – unmoored in practice. For example, Lodge and Taber (2013) summarize their enormously influential, twenty year research program thus:

> "Human beings are consummate rationalizers, but rarely are we rational…[P]olitical behavior is the result of innumerable unnoticed forces, with conscious deliberation little more than a rationalization of the outputs of automatic feelings and inclinations."

[1] Theories of motivated reasoning argue that most human reasoning is "directional" – i.e., designed to shore up pre-determined conclusions – rather than oriented to "accuracy" (though, as we shall see, it is often unclear what accuracy might mean, making it rather too easy to claim that directional goals trump accuracy). Affective primacy posits that emotional processing proceeds and dominates rational processing in the great majority of situations. We elaborate on Haidt's social intuitionism at length below.

Similarly, the prominent psychologist, Jonathan Haidt, argues that reason evolved as a tool of rhetoric rather than the other way around, and more importantly that its original character is destiny when it comes to applications in modern ethics and politics. Indeed, he goes so far to call his theory "Glauconian," after Thrasymachus's ally in the dialectical jousting with Socrates: "In fact, I'll praise Glaucon for the rest of the book as the guy who got it right" (Haidt, 2013: 86).

In the *Republic* though, Thrasymachus, the cynical sophist, blushes when Socrates unmasks his arguments as cynical and sophistic. This detail is curious and striking. As Allan Bloom has noted, "The apparently shameless Thrasymachus, willing to say anything, is revealed in all his vanity, for he blushes" (Bloom, 1968: 336). Such reactions make little sense, though, on the reason-as-rationalization account, in either its ancient sophistic or modern scientific varieties. Indeed contemporary empirical research also supports the existence of cross-cutting forces that hem in and alter our tendencies to behave as cynical sophists – a kind of photonegative of motivated reasoning that we might term 'principled rhetoric.'

In addition to the evolutionary traits that Haidt discusses (which we agree are well established), our common ancestry has made it so that very few of us – mostly sociopaths – are immune to shame and able to behave as cynical sophists through and through. We are disposed to track and respond to reasons in ways that are not purely strategic (Tomasello, 2009). It turns out that even Thrasymachus is what we might call a "theoretical" (not a congenital) sociopath: whatever his professed views about justice, power, and the sophist's vocation, he blushes and feels the force of accountability to good reasoning.

Haidt's research, as well as Lodge and Taber's, purports to show that so-called "rationalist" moral and political theories rest on untenable empirical premises. Their work has been published in the best journals, and their ideas have been discussed and deployed in a wide

range of fields including psychology, philosophy, political science, communications, law, anthropology, sociology, among others. In addition, Haidt has published a best selling trade book and taken to TED talks and the pages of the *New York Times* to advance the potentially revolutionary implications of his research for our moral and political self-understanding. So both in theory and practice, a lot rides on whether the anti-rationalist implications of such work go through.

Below we argue that such implications do not go through. The anti-rationalist research program stands as a remarkable case study of how the translation between normative theory and social science can go fundamentally awry. In the end, the anti-rationalists' empirical results are either incoherent with the purposes to which they wish to put them, or they are better interpreted as supporting and elucidating broadly rationalist theories, rather than undermining them.[2] The point, however, is not merely to correct interpretive mistakes for their own sake (though there is certainly merit in doing so). Rather, we show that the *way* the translation process goes wrong in these cases exemplifies a recurring theme in such attempts, one rooted in the disciplinary divide over facts and values. So, in this light, we sketch an *inferentialist* model of judgment, arguing that it can help to overcome the difficulties in translating between the normative and social scientific domains.

I. Intuitionism Contra Rationalism

Haidt's social intuitionist model (SIM) posits a pivotal role for affectively managed intuition in making normative judgments. When one encounters a situation that elicits a moral judgment, the immediate reaction is a gut feeling that manifests in a quick evaluation. This

[2] Here we do not set out to defend a specific meta-ethical/political theory, but rather to render a broad class of them – any that rely importantly on our rational deliberative capacities – plausible again in the face of the empirical critique.

immediate, intuitive reaction is an analogue and close cousin of, for example, human disgust reactions (Haidt, 1993). For Haidt, the slower, more deliberative process of reasoning plays only an ancillary role as a *consequence* of intuition and judgment rather than its source. Reasoning in the SIM is typically just ex post rationalization in justifying one's own positions, and an attempt to win other people to one's position by any means, not on the basis of "good" or "better" reasons, whatever, if anything, those might be. In the model, the primary causal effects of reasoning are social; reasons may shape the intuitions of *others*. The capacity for reason to tutor one's own intuitions and judgments is accorded a much lesser role, although some provisions are made for variation across individuals and situations, and for this capacity to be cultivated and marginally strengthened with practice. Haidt explicitly calls reasons the "junior partner" to affect-as-intuition (e.g., Haidt 2010).

The SIM is based on a considerable body of experimental research, and it dovetails neatly with some of the other empirical regularities and theories that comprise modern psychology, such as the dual-process model of cognition (Kahneman, 2011) and motivated reasoning (Kunda, 1990). The evidence that forms the core of Haidt's argument for the SIM is exemplified by the reactions of experimental subjects to stylized stories that invite moral judgments. Perhaps the most famous story involves a brother and sister who, while traveling together, decide to have sex. The two use two forms of birth control, never tell anyone their secret, never have sex again, and find the experience to have deepened their relationship; the point of these stipulated details is to block the inference that their action caused harm. When this sort of story is presented to experimental subjects, the experimenters observe what they call *moral dumbfounding*, the "stubborn and puzzled maintenance of a moral judgment without supporting reasons" (Haidt, Bjorklund, and Murphy, 2000, 6). For example, a subject might say that it is not OK for the

brother and sister to have sex, not for any identifiable reasons (in the face of various stipulations and follow up questions by the investigators), but because "it's just wrong."

In presentations of the SIM, Haidt frequently casts the model as a response to and critique of the "rationalist" model of moral judgment of Lawrence Kohlberg. Beginning in the 1960s, Kohlberg was at the vanguard of the cognitive revolution in psychology. His chief contribution is a cognitive-developmental, stage model of moral reasoning (Kohlberg, 1969). The stages refer to the increasingly sophisticated lines of reasoning that children use as they mature. Kohlberg's model posits that those reasons manifest directly in evaluations. Emotions and social interaction are also accorded (ancillary) roles. The principal difference between Kohlberg's stage model and Haidt's SIM model is the direction of the causal arrow between judgment and reasoning: Kohlberg takes reasons to affect judgment, and Haidt takes judgment to affect (post hoc) reasoning. In Haidt's words, Kohlberg posits reason as the "senior partner" to affect-as-intuition (e.g., Haidt 2010).[3]

One pillar of Haidt's critique is that Kohlberg's model cannot account for moral dumbfounding.[4] If Kohlberg were right, the moral judgments that people offer when they hear this story would have to have been caused by reasons. But Haidt reports that very frequently, subjects cannot identify the reasons why they judge some actions to be "just wrong."

Interestingly, Haidt and his colleagues also offer an account of how Kohlberg could have been so wrong. Kohlberg based his theory on a series of experiments in which children were

[3] For a discussion of the role of emotion in deliberative theory generally, see Neblo (2003). For a more concrete discussion of a systemic conception of deliberation, see Lazer et. al. (2011) and Neblo (2005). For applications of these ideas to the case of race politics, see Neblo (2009a,b). On the empirical relationship between reasons, emotions, and speaker attributes in persuasion, see Neblo et. al. (2012).

[4] Haidt actually lists "four reasons to doubt the causal importance of reason," the other three being what he calls "the dual process problem," "the motivated reasoning problem," and "the action problem." All four purported problems are closely related and our arguments in reply will generally apply to the entire syndrome that Haidt offers as reasons against the viability of rationalism broadly understood.

presented with moral dilemmas and asked a series of questions. Perhaps the most famous dilemma involves Heinz, who, unable to afford a potentially life-saving drug, must decide whether to break into a pharmacy to steal it for his dying wife. Follow up queries featured hypothetical questions of right and wrong, e.g., "What if Heinz didn't love his wife? Would it still be OK for him to steal the drug?" According to Haidt, Kohlberg perceives the behaviors he observed his subjects engage in as a process of cognition that is "conscious and used ordinary moral language" (Kohlberg, Levine, and Hewer, 1983: 69, as cited in Haidt, 2001). But elsewhere, Haidt and his coauthors allege that this perception is mistakenly predicated on an odd sort of data, generated by a strange process:

> "Kohlberg may have concluded that moral judgment was based on moral reasoning because the dilemmas he used, such as Heinz, had very salient fodder for post hoc "reasoning-why." In his dilemmas there were always questions of rights and harm (cf. Kohlberg, 1969). Had he used a broader sample of moral judgment tasks he might have come up with a different theory, one that gave greater prominence to moral emotions and the "seeing-that" of moral intuitions. (The tendency for psychologists to confuse a psychological phenomenon with the way they have chosen to study the phenomenon was called "the psychologist's fallacy" by William James, 1890/1950.)" (Haidt, Bjorklund, and Murphy, 2000, 11).

The underlying assumptions here are that Haidt's stories have more verisimilitude than Kohlberg's dilemmas, and that the process of moral dumbfounding is less strange than questioning via hypotheticals.

Building on the framework of the SIM, Haidt has articulated a moral foundations theory (MFT) of intuitive ethics (Haidt and Joseph 2004, Haidt and Graham 2007, Haidt 2012). According to the MFT, people base their moral judgments on intuitive reactions that cluster into several foundations, which include care (roughly utilitarian beneficence), fairness (roughly notions of right), in-group loyalty, hierarchical authority, and notions of "purity" (failures of which elicit disgust reactions). There is variation in the degree to which different individuals

identify these foundations as important, and that variation correlates highly with well-understood categories from politics (i.e., liberal/conservative) and religion (i.e., believing/atheist).

Based on the SIM and the MFT, Haidt appears to encourage us to collapse descriptive ethics and normative ethics – i.e., how, as an empirical matter, we in fact *do* tend to make judgments with how, as a normative matter, we *should* make those judgments. If (1) moral judgments are primarily caused by intuitions, (2) intuitions are analogous to disgust reactions, and (3) different people are affected by different subsets of those moral intuitions, then differences in moral judgments cannot be resolved by saying that one side is right and the other is wrong. For example, the act of denying the moral judgment "homosexuality is immoral" is similar to denying that having a disgust reaction to eating insects is reasonable. Thus, political foes do not properly have access to claims of moral superiority. Instead, they merely have different tastes (Haidt actually uses analogies to taste buds and aesthetics), different intuitive reactions to eliciting situations. So Haidt embraces a fairly strong version of moral relativism (despite protestations to the contrary). Reason plays a negligible role in forming our moral judgments, and only gets deployed to figure out effective ways to bring others around to our pre-existing views. Reason is merely a rhetorical tool to convince others by any means available, and justice becomes the right of the stronger in wielding those and other more avowedly manipulative tools.

And yet, at other points, Haidt deploys his empirical findings and theoretical apparatus for highly prescriptive purposes, arguing that his findings encourage a kind of *moral leveling* in which we should respect our political opponents and attend to differing morals, whatever their contents. At points his account of moral foundations sounds like an updated and socialized version of early twentieth century ethical intuitionism in which humans have a cognitive faculty

that directly perceives non-natural value properties. Regardless of which horn of the apparent dilemma Haidt is inclined to grasp, we will argue that the dilemma is false, resting on a mistaken interpretation of his model.

II. Thinking Fast, Thinking Slow, and Thinking Again

In the present context, we do not intend to critique the SIM or MFT as theories in their own right, so much as to critique the interpretation and implications that Haidt claims for them. As mentioned above, Haidt explicitly pits himself against Kohlberg (Haidt, 2001), with the key controversy being whether reason or intuition ought to be considered the "junior" partner (e.g., Haidt, 2010). We argue that this is a false dichotomy and a positively confusing way of organizing concepts that follows from mistaking a vaguely defined notion of frequency with causal weight and critical standards.

First, as we alluded to above, Haidt's argument assumes that the SIM is superior to Kohlberg's stage theory in large part because the evidence of moral dumbfounding is elicited by more realistic situations than the evidence of the stage theory. This claim is, to put it mildly, debatable. It is hard to see why, for example, a story about a brother and sister who have sex without any chance of the events coming to light or causing harmful consequences is more realistic than a story about a man who steals medicine to save his wife.[5]

Unlike the Heinz dilemma, the incest scenario is maximally artificial in that it requires a God's eye point of view to avoid contradicting the stipulations in the scenario: if no one ever found out about the events, how did I come to be in a position to judge them? Similarly no real human judge could ever be sure that no power or exploitation was involved, that no social or

[5] Haidt's other supposedly more representative examples of moral judgment situations include eating human cadaver flesh or animal carcasses that have been used for masturbation.

psychological harm could ensue, etc. Most of us can easily imagine ourselves as a friend or acquaintance trying to know what to do with such information, as a juror judging a specific case that is being prosecuted, or a legislator trying to decide whether to legalize incest on libertarian grounds. But all of these roles mean that we would be judging without the stipulations that block the standard sort of reasons that people might invoke. What Haidt calls dumbfounding seems like understandable resistance to accepting what respondents see as implausible stipulations and a completely abstract notion of the position from which they are rendering their (lab) judgment.[6]

Haidt argues that the artificiality of the Heinz dilemma is a problem for Kohlberg but the artificiality of the incest story becomes an even bigger problem for Haidt. At least, it is not clear that a model of moral judgment should give more weight to reactions to one type story rather than the other. An ideal model should be able to account for both. Ironically, Haidt accuses Kohlberg of exactly the mistake he makes.

Haidt's second (and more subtle) mistake is to equate "frequency" with causal and conceptual importance. Importantly, Haidt's lab experiments do not test for frequency in any externally valid way, or even employ a well-defined concept of frequency. But even if we were to grant that they did, the significance of such a finding for his critique of "rationalist" theories of morals and politics is not at all clear.

Many broadly rationalist theories make of point of arguing that explicit reasoning is a fairly specialized and episodic process. Habermas (1996), for example, argues that for most everyday social interactions we rely on relatively settled background assumptions and affectively

[6] We should also note that the evidence in support of moral dumbfounding is surprisingly weak for how prominently the phenomenon has become in this literature. The main evidence relies on an unpublished paper with a small college sample and modest statistical and substantive effects. We also conjecture that the results – if they are replicable at all – are sensitive to the time index of the judgment. That is, the fait accompli frame of the judgment along with the interviewer's aggressive follow ups screen off the fact that the action was almost surely reckless, even if it ended up being harmless in the technical sense. (Sketch the idea for the time index prime experiment to test the conjecture.)

managed behavioral dispositions to coordinate our actions and furnish appropriate social judgments. For Habermas, we rely on the *lifeworld* – much of which can be readily glossed as "social intuition" – to manage the great bulk of mundane interactions. Crucially, though, such assumptions and behavioral norms have an implicit presumption of legitimacy (i.e., a genealogical connection to reason, even if remote). It is only when such interactions break down or the presumptions of legitimacy are challenged that we have to thematize the implicit social rules and values that undergird our intuitions, subjecting them to explicit reflection and scrutiny. So for Habermas, at least, Haidt's findings about "frequency" is actually something to be expected (indeed, that is functionally necessary) under his archly "rationalist" model.

But perhaps more importantly, it is not clear that Haidt has even conceptualized reason in a helpful way, nor identified operational tests that could be interpreted as establishing his claims. Although the SIM mixes together both reason and intuition, Haidt often says that reason must be the junior partner (e.g., Haidt, 2010). His argument for the primacy of intuition is based on how often, *in his experiments,* we observe intuition cause actions, relative to how rarely we observe reason cause actions. In addition to the rather crucial sample frame problem, equating frequency and causal importance is hardly obvious or sensible. Intuition and reason can work on different time scales, and it is not clear why a single causal instance in which reason shapes intuitive dispositions ought to be accorded the same degree of importance as each instance in which that intuitive disposition causes a moral judgment.

Consider an analogy to a ship's captain and helmsman. On a ship, the captain orders the helmsman to lay in a course, say once every few hours (time scale 1). The helmsman will not simply turn the wheel to a locked position. Instead, she will keep vigil over her heading every now and then, monitoring changes that have occurred based on her previous actions and her

environment, and issuing course corrections. Suppose that she does so every couple of minutes (time scale 2). Her only action of interest is to direct the rudder. Water flows around and pushes against the rudder. Small changes in the vortices of the water shift the rudder around. Very small changes in the water are corrected by the very small actions of the rudder, at a very small time scale, say every few seconds (time scale 3). In principle, we could continue this story for many smaller time scales, down to the level of quantum mechanics.

However, if we are most concerned about why the ship ends up where it does, we face an identification problem. We can tell causal stories about the captain, or about the helmsman, or even at some more fine-grained level. But, at least in navigation, when we tell such causal stories, we seldom descend past time scale 2. Although actions at smaller time scales are perfectly causal, these actions are simply not very informative about why the ship moves as it does. Normative questions, such as, "Where should the ship go?" or "What is the best way to get there?" have even less interesting or meaningful answers at small time scales. If our principle concern is normative, then focusing on small time scales is necessary only insofar as action therein constrains our abilities to get where we want to go. But by that reasoning, it is then not clear whether it is even worth talking about the helmsman as opposed to the captain.

When Haidt argues that reason should be viewed as the junior partner to intuition because the latter causes moral judgments much more often than the former, he is ignoring the idea that a little bit of tutelage can go a long way. At a short time scale, intuition often seems to be the senior partner, just as the helmsman seems to be more causally efficacious than the captain. But on a longer time scale, intuition recedes in importance, and the relatively rarer, but more influential role of reason seems to be where more of the interesting action is.

If our intuitive judgments rooted in the lifeworld were once the subject of explicit debate and contestation (as many plainly were), there is a sense in which they can inherit a rational genealogy from the indirect, long-term effects of that debate. Similarly, at the level of the individual, many of our habits and character traits were once deliberately cultivated. So the fact that actions flow more or less automatically from them now does not impugn the sense in which we can consider them the product of reasoning and choice. None of Haidt's evidence speaks to such issues at all. It is not even clear, then, how we should parse the direct and indirect effects of reason versus intuition, especially over time. Each is junior and senior, depending on the frame of reference, and as noted, normative theory can reasonably claim the larger frame as most relevant to its concerns.

III. Motivated Reasoning & Democratic Competence

Lodge and Taber's (2013) *The Rationalizing Voter* advances arguments and claims similar to Haidt's, but with special reference to political judgment. They begin: "Grounded in an Enlightenment view of Rational Man, political science has been dominated by models of conscious control and deliberative democracy. (p. 1)" On the contrary, though, their findings show that "[Political] deliberation is a bobbing cork on the currents of unconscious information processing … we have [merely] the illusion of standing at the helm." They conclude that we should be "skeptical of the ability of citizens to reliably access or veridically report their beliefs and attitudes. (p. 22)" If so we might join a sympathetic reviewer in saying that "the book might be better titled 'The Illusion of Choice in Democratic Politics.'" Though Lodge and Taber are more circumspect than Haidt in advancing explicitly normative arguments about democratic

theory, the general thrust of their findings is unmistakable: any account of democracy that relies on public reasoning (or even the rational self-interest of citizens) is built on foundations of sand.

The first thing to note in light of such "affective primacy" findings is that for most "rationalist" democratic theorists, the implicit antonym of reason is not emotion, but rather illegitimate power. Outcomes determined by arbitrary or malignant authority are the bads to be avoided. Affect and emotion, in themselves, do not pose any threat to "reasonable" outcomes in this sense. So relying on the everyday sense in which reason and emotion are opposite is highly misleading in this context. As in our response to Haidt, emotions can be both responsive to reason and to shape our reasoning in ways that are warrantable. So Lodge and Taber's findings that reason and emotion massively interpenetrate does not cause problems for rationalist theories at all, nor does the apparent sense that affect is somehow "primary," (for the same reasons we adduced in our reply to Haidt).

That said, some of Lodge and Taber's more specific findings would seem to cause trouble for a too easy reconciliation between reason and affect. For example, the most apparently damning phenomenon that they elucidate to illustrate their point is so called "motivated reasoning." Citizens are often "motivated more by their desire to maintain prior beliefs and feelings than by their desire to make 'accurate' or otherwise optimal decisions. (p. 150)" The evidence for directed or motivated reasoning rests on four interrelated phenomena: a *prior attitude effect* in which people often evaluate supportive arguments as stronger than opposing arguments; *biased processing* in which people often spend more time and effort thinking about and challenging arguments that go against their priors than those that support those priors; *attitude polarization*, in which exposure to a "balanced" set of consideration often leave people more strongly in favor of their initial position; and finally the *sophistication effect* in which the

previous phenomena are often more prominent in those who are more knowledgeable and sophisticated about politics or the issue at hand.

As with Haidt, in the present context, we are less interested in contesting Lodge and Taber's findings than the interpretation that those findings are given. The first thing to note in this regard is that it is not clear that any of these phenomena necessarily bespeak irrational behavior (or require positing a special motive to maintain prior beliefs whatever they may be, as opposed to "accuracy"). For example, if I have already demonstrated that I am generally sympathetic to conservative arguments by evincing a conservative attitude, it is hardly surprising (or even problematic) that *ceteris paribus* I should find further arguments that tilt conservative more convincing.

Similarly, at least some "biased" processing is entirely reasonable. First, one might spend more time processing counter-attitudinal arguments merely because they are less familiar and require more effort to address (Ross, 2012). Moreover, coherence in our beliefs and commitments is an eminently reasonable goal. So if I encounter some highly incongruent new information it can be reasonable for me to subject it to extra scrutiny. For example, when psychologists recently published a new set of findings purporting to find evidence of ESP, the research community reacted with more than standard skepticism and scrutiny relative to findings that did not pose fundamental challenges to their discipline's belief system (Bem, 2011). A fortiori, reflective equilibrium (as opposed to deductive theory testing) in the face of potential value pluralism creates even more scope for potentially reasonable "biases" in processing.

Moreover, if one subjects challenging evidence to higher scrutiny and then finds the new evidence lacking, there is a sense in which one's prior beliefs have survived a test, in which case it is not always unreasonable to update toward those priors. And we should expect sophisticates,

with stronger priors and more capacity for processing the coherentist implications of challenging evidence, would be more pronounced in manifesting these effects.

The second main thing to note is that these effects are just that – effects. That is to say, Lodge and Taber's evidence does *not* show that people give *no* weight to counter arguments, or *completely dismiss* disconfirming evidence. We do not deny that some motivated reasoning is likely to be problematic from a normative perspective. But without a substantive theory of rational judgment it is difficult to get a sense of how problematic such effects are, or when they merely reflect a version of what Rawls called "the burdens of judgment" (and the way that such burdens interact with our need to process information via different folk theories of politics).

Finally, in cases when we do think that directional biases are problematic, there is reason to believe that the problem can be substantially remediated. For example, simply prompting subjects to think about how they would evaluate a given methodology had it produced the opposite conclusion almost completely eliminated the differential (Lord et. al. 1984). A whole subfield in cognitive psychology has grown up around such techniques for individual level debiasing (Lewandowsky et. al. 2012). Moreover, there is reason to believe that the contestation and pervasive social accountability surrounding politics will often help constrain the degree and scope of the phenomena that Lodge and Taber identified in their lab experiments. People exposed to more than one set of frames and streams of information are less susceptible to framing effects, as are those who process such information in groups (Druckman 2004), as well as those expect accountability to others with differing views (Huckfeldt et. al. 2004). So the evidence for affectively "tainted" reasoning hardly constitutes grounds for despair about the role of reason in democratic politics and the apparent elitism or decisionism that would seem to follow on such despair.

IV. From Descriptive to Normative Ethics & Politics

We have argued that Haidt's and Lodge and Taber's claims about reason being decisively subordinate to affect-as-intuition are 1) not well conceptualized, 2) not well supported empirically, and 3) that even if it were well conceptualized and well supported, that prominent rationalist theories can and do accommodate versions of the idea quite consistently – i.e., it would not count decisively against rationalism in the way they claim. This, however, is a relatively benign mistake compared to the problematic way in which Haidt (and to a lesser extent, Lodge and Taber and other affective primacy scholars) move between descriptive and normative claims about ethics and politics. When Haidt distills prescriptions from his (mistaken) interpretation of moral dumbfounding, the result is to recommend *moral and political leveling*: different moral intuitions and hence political judgments should be accorded equal weight and respect, with little regard to the different lines of reasoning that they may summarize, or how they may have come to exist in the first place.

Returning to the analogy of the ship, a major reason that we do not tend to focus on small time scales when we are talking about navigation is that by doing so, we risk encouraging the inference that actions at the higher time scales are not really actions at all. In our story, the captain's decisions are heavily mediated; they do not directly cause the ship to go anywhere. To the extent that we operate on the principle that more direct causal relationships at smaller time scales are more important than indirect causal relationships at larger time scales, we might decide to forego paying attention to the captain. If the captain is not doing anything, then we could not say whether it would be better to go to one destination or another, or to take one route

or another. From the perspective of the helmsman, there are no good grounds on which to prefer one destination from another.

Rejecting a meaningful role for the captain is the analogue of the moral leveling that Haidt engages in when he moves too quickly between descriptive and normative accounts of moral and political phenomena. According to Haidt, we ought to reject the notion that some of moral foundations are more important than others because, from the quick, intuitive perspective, all of them look the same. But this leaves out the possibility that less frequent causal actions at larger time scales, like those of reasoning, are meaningful, despite their relative "rarity" such as it is. If we had an alternative theory as to why some moral foundations end up being more important than others at larger time scales, we would also have a good basis for rejecting the general principle of moral leveling that Haidt encourages.

To see an example of the sort of mistake that Haidt encourages us to make, consider what often happens when one teaches the famous Monty Hall problem to a student. In this problem, there are three doors. Behind one of the doors, there is a new car; behind the other two, goats. The decision maker first selects one of the doors. Monty then opens one of the other two doors, revealing a goat. Finally, the decision maker chooses whether or not to switch doors. Counterintuitively, the rules of probability suggest that the decision maker should always switch. As many educators know, when one teaches this problem, students often continue to dispute the switching principle, even after following along and agreeing with each step of the relatively complex reasoning process. This outcome bears a striking similarity to moral dumbfounding. Yet, the educator believes that there remains an important sense in which the switching principle is correct, even if, in a statistical sense, many students might disagree. That is, the educator rejects the analogue of moral leveling in this case. But why?

The key assumption that leads the educator to believe that she is correct, despite opposition from her students, is that with enough time and communication the students will eventually agree that they were mistaken.[7] And, crucially, they will do so because they are committed to various other premises and beliefs that constrain what one can coherently maintain, even if few people recognize such constraint immediately. Indeed, there is the possibility that *everyone* could be wrong (initially) about something eventually revealed to be incoherent with stronger commonly held beliefs and commitments.

This sort of assumption is missing from Haidt's model, and so, therefore, is the ability to be wrong in a meaningful sense. In Haidt's model, an argument can be more or less persuasive in the descriptive sense of garnering support or changing more individuals' minds. The process involves trading one intuition for another. But there is no account for why one intuition should be stronger than another, beyond the mere fact that it is, empirically. Similarly our attempts to persuade are either effective or ineffective, but there is no way to distinguish between persuasion for good reasons versus bad reasons, or for that matter, between persuasion per se and manipulation. Intuitions and emotions in his model are fundamental, and, although they may be influenced by other people's rationalizations, such influence is a bare, contingent fact.

Thus, when Haidt argues that we should abandon rationalist theories of normative ethics and politics, his potential grounds for doing so are all rather unattractive. First, he could admit that he fully conflates descriptive and normative ethics and politics, in which case he is merely exhorting us to switch without any grounds for thinking that we should (beyond his intuitions).

Second, he could embrace a revamped theory of ethical intuitionism, which widely fell out of favor because of its anti-naturalism and its inability to account for persistent moral

[7] Or is it a pragmatic money pump? Do these end up being the same in practice? Does this analogy work?

disagreement. This would be the most interesting option for Haidt, since his lab experiments grounded in evolutionary theory appear to provide a way to reconnect such theories to naturalism, and his theory of the five moral foundations can provide an account of persistent (if constrained) disagreement. But Haidt shows few signs of wanting to embrace moral realism of this sort, and doing so would be more difficult than it appears, since the evolutionary elements of his theory still do not provide a way to jump from natural (i.e., descriptive) processes of how we make value judgments to actual values (i.e., normative) phenomena.

Though not made explicit, the line that Haidt most often seems to take is a second-order appeal to either liberal respect and toleration (i.e., rights) or social utility (i.e., harm and beneficence). At various points Haidt suggests that recognizing varying moral foundations and our intuitionist behavior patterns may be instrumentally useful in promoting beneficial outcomes. This may be true, but it seems to concede that our substantive normative theory could still be a rationalist one like utilitarianism (which would also contradict his insistence on a plurality of moral foundations).

At other points, he argues that, for example, liberals in politics should not be so dismissive of conservatives because the latter are actually more attuned to all five moral foundations, rather than just care and fairness (e.g., Haidt, 2012, inter alia). Here he seems to suggest that such descriptive diversity in political values deserves our respect, which is quite plausible, but highly ironic. Such a rationale is close to the one – dismissed and even mocked by Haidt – that Rawls uses to defend his theory of justice. For Rawls, respect and fairness undergird a theory that protects our rights against overly expansive claims of utility (i.e., exclusive reliance on his harm foundation). But one of the main purposes of such rights is to allow us to develop forms of *private* life that often place great value on submission to hierarchical authority (e.g., to

the Catholic magisterium), or in-group loyalty (e.g., to co-ethnics), or to avoid those we regard as violating purity and sanctity (e.g., homosexual couples).[8]

Thus Haidt's own ambiguity about the status of his argument ends up being a consequential mistake for both science (at least in terms of interpretation, and the conceptual set up for future inquiry) and for the way that it purports to constrain plausible moral and political theories, and especially the interaction between the two: how the normative categories (e.g., moral foundations) get operationalized and interpreted, and their significance for normative inquiry.

V. Making Our Intuitions Explicit

We still find ourselves in search of a theory of judgment that can accomplish four tasks. First, this theory ought to explain the empirical evidence offered by Haidt as well as that offered by Kohlberg. As such, it ought to subsume both Haidt's SIM and Kohlberg's stage theory. Second, the enriched theory should feature a key role for time scale. Practically speaking, we want a theory that looks like Haidt's theory at small time scales and looks like Kohlberg's at larger time scales. Third, the theory should be able to help us adjudicate between benignly "motivated" reasoning, and instances that we rightly judge as a defective form of reasoning. But most importantly, this theory ought to provide a road map that can be used to avoid moral leveling, or at least to know it can be avoided. That is, the theory ought to provide a more convincing bridge between descriptive and normative ethics and politics, and thus explain how we can avoid moral leveling in a *principled* fashion.

[8] Perhaps Haidt would like those three moral foundations to operate co-equally (with utility and rights) in a directly public and political fashion. If so, he would need to address the historical grounds against doing so (e.g., Rawls's invocation of the lessons learned via the wars of religion) in some substantial way. [Might we expand the Rawls discussion in to a section of its own?]

Robert Brandom (1994) constructs an inferentialist model that can be extended to incorporate both the social intuitionist model (SIM) of Haidt and the rationalist stage model of Kohlberg. Brandom identifies the process of giving and asking for reasons as the empirical basis for developing principled justifications for normative claims from otherwise mere rationalizations of judgments. The title of Brandom's major book, Making It Explicit, nicely illustrates the connection between reason and affectively mediated social intuitions.[9] Inferentialist articulation is the process by which we take our intuitions and try to explain and justify them explicitly by reference to mutually interpretable standards. Not coincidentally, then, social reasoning is a linchpin of both Haidt's model and Kohlberg's model. Combining the work of Brandom, Kohlberg, and Haidt, we deploy Brandom to sketch an inferentialist model of judgment that features an account of the social emergence of culturally meaningful normative principles via interactive reasoning (time scale 1), and an account of split-second judgments via affectively mediated intuitions (time scale 2). Transitions between time scales are accomplished by viewing intuitions and reasons each in terms of the other. Intuitions can be interpreted as encoded reasons as in the on-line model of memory and judgment (Hastie and Park 1986; Lodge, McGraw, and Stroh 1989). Reasons can be interpreted as coherent and self-stable bodies of intuitions accrued in an evolutionary process of interactive reasoning (cf., Bowles and Gintis 2011). Neither reason nor intuition can operate well without the other.

Brandom casts giving and asking for reasons as elements of a game, in which players track each other's behavior through a process of *deontic scorekeeping*. Essentially, a player's score is just a summary statistic to keep track of when they engage in incoherent, hypocritical

[9] Does this mean that we need to spend even more time directly addressing dumbfounding and rationalization – i.e., because the point that Haidt and Lodge and Taber are implicitly making is that we are bad at making the real grounds for our intuitions explicit – which would seem to mean that we are bad at inferential articulation?

behavior. In the game, making a claim has the dual consequences of (1) entitling one to assert any statement that is an inferential consequence of that claim, and (2) prohibiting one from asserting any statement that is incompatible with that claim. Just what counts as an inferential consequence or an incompatible statement is defined by bootstrapping the rules of the game. For example, a player could hypothetically claim that a rule of elementary logic is invalid. But in so doing, that player would back herself into a corner, wherein she would inevitably be forced to rely on the socially articulated consequences of that rule to justify other claims she might want to make. Ultimately, she would have to drop her original claim and agree to live by the logical rule.

Brandom's key move is the bootstrap, in which the simple rules of deontic scorekeeping explode into a universe of inferential consequences. Although Brandom formalized the bootstrap, it is implicitly prefigured in Kohlberg's stage model. In Kohlberg's model, the social act of reason giving is the primary means of development from one stage of moral reasoning to another (Kohlberg 1969). His model isolates the deontic element of Brandom's game of giving and asking for reasons, in which, once another player has noted an inconsistency, the player responsible for the inconsistency must reason her way through it to a more coherent inference, where coherence is itself socially defined. Taken at the time scale of cognitive development, this process explains how Kohlberg can "commit the naturalistic fallacy and get away with it" (Kohlberg 1971). Kohlberg's model shows how meaningful moral principles can emerge from empirical social practice, just as Brandom's model shows how logical principles can emerge from deontic scorekeeping.

Brandom's bootstrap is also a key component in Habermas' conception of the *lifeworld* and the processes of its evolution. The lifeworld is the shared, common understanding of what is valid or good; it is the sum total of what is taken for granted in a conversation. When a piece of

the lifeworld is thematized, for example by becoming the subject of an argument, it ceases to be taken for granted, and thus ceases to be a part of the lifeworld. When people argue in good faith, they attempt to warrant their judgments and actions by relying on the remaining totality of the lifeworld. Over very large time scales, new pieces of the lifeworld come into being through many acts of communication. The evolution of the lifeworld is essentially another manifestation of Brandom's bootstrap.

But the bootstrap is not cheap. A common criticism of Brandom's model is to question its empirical value: how could a human being possibly keep the explosive multiplicity of inferences that his model identifies? Habermas (1996), and even Kohlberg to some extent (1969), explicitly admit that the great bulk of our everyday normative judgments do and even must occur via social intuition. That is the whole point of the theory of the lifeworld, and our limited ability to problematize ever larger swaths of it. Deliberation, discourse, and explicit moral reasoning are quite specifically exceptional.

We take Haidt's model to be a psychological theory of the lifeworld. Humans do not track each inference in the game of giving and asking for reasons, just as they do not constantly trace all lines of arguments down to foundations. Instead, humans track summaries of these inferences, encoded as automatic intuitions and emotions. A similar process is at work in the on-line model of memory and judgment (Hastie and Park 1986; Lodge, McGraw, and Stroh 1989). Haidt's model shows how judgments can be issued quickly and cheaply. The SIM even identifies how the larger processes of interactive reasoning and lifeworld formation can enter back into our intuitive processes, as the model includes a role for others' people's reasoning to affect one's own intuitions.

Because the inferentialist model subsumes Haidt's model at small time scales and Kohlberg's at larger time scales, it can also explain their empirical findings. But then a reasonable response to this model would be that it is too complicated, that Haidt's simpler model (without his mistaken interpretations) is in some sense enough. In rejoinder, we can point to two sorts of things that the inferentialist model does that Haidt's model does not. First, the inferentialist model helps to make sense of an emerging body of experimental evidence that cannot easily be accounted for by the SIM. One set of experiments tests how actions and judgments change when people are forced to pause before acting or judging. For example, Rand, Greene, and Nowak (2013) find that when people are forced to wait as little as ten seconds before deciding how to act in a collective action game, they tend to donate less to the collective good, an action which is consonant with individual utility maximization. Paxton, Ungar, and Greene (2012) observe that when people are forced to wait for several minutes before rendering judgments in the incest dilemma, they tend to be more permissive. Haidt's model does not offer a good reason why different intuitions ought to crop up differently at different time scales, but the inferentialist model does. And Sklar et. al. (2012) show that people can read and do arithmetic nonconsciously, which suggests that Haidt's partitioning of reason and intuition does not have the force against the rationalist notion of persuasion that he would like to claim for it.

The second and more important advantage of the inferentialist model over Haidt's model is that it avoids moral leveling. Ironically, the element that is missing from the SIM is the ability to be wrong in a normatively meaningful sense. In Brandom, Kohlberg, and Habermas, "being wrong" is the practical equivalent of eventually coming to agree with an interlocutor that one has made a mistake and to retract it, typically after much communication. Haidt (2003) admits that persuasion happens, but glosses it as activating new moral intuitions for the most part. Doing so

begs the question of how we construe "persuasion" (and whether, in a funny turn, the way he uses it is a case of "persuasive definition" since it assumes that "rational" is not well defined or causally efficacious). The inferentialist model not only provides a way to be wrong, it turns being wrong into the fundamental building block of what it can mean to be right. In so doing, we can stake a claim that some moral intuitions are better than others, and thereby escape the torpor – or worse, the realist equation of power and justice – that attends moral leveling.[10]

VI. Conclusion

For two millennia, philosophers were typically also the best social scientists of their day. Their descriptive psychology was designed from the beginning to serve as a logical base for their moral psychology, which served, in turn, to underwrite their ethical theory, and on to their political theory. And moving in the other direction, the descriptive psychology had to be compatible with, and preferably entailed by their epistemology and, in turn, their metaphysics and ontology. In short, these were often systematic thinkers whose work spanned the practical, the scientific, and the philosophical (Neblo, 2007).

The rise of modern social science has created a necessary, and in many ways salutary, division of labor between philosophers and social scientists, driven primarily by the need for specialization in the face of technical advances. This is not to say that the two have proceeded in pristine isolation from each other. Many social scientists aspire to be practically relevant and regard their research as having important implications for normative theory and practice – "giving hands and feet to morality" in the words of one (Lasswell, 1941). Similarly, many philosophers question the sharpness of the divide, or at least believe that their conceptual

[10] Find some non-clumsy way to pair this point with the scientific point about effects in the absence of a substantive theory of rationality.

apparatus should help nudge the social scientific research agenda, just like the normative category of "disease" guides medical research without compromising its scientific status.

And yet, because the division of labor has only intensified, the ability to manage good integration of normative philosophy and social science has become fraught with dead ends and positively harmful missteps in translation and transposition. The current incarnation of the ancient debate between rationalist and sophistic theories of morals and politics illustrates this phenomenon and the stakes that attend it. Haidt's and Lodge and Taber's attacks on rationalism, unlike Thrasymachus's, come with the force of modern social behind them. But unlike past debates, such critiques lack an integrated philosophical framework to help place the scientific findings in a more systematic interpretive context. We hope to have shown that even if we accept the social science at face value, anti-rationalist conclusions do not follow from them. Moreover, by sketching a substantive theory of rationality that can link the descriptive and the normative, we can create a more stable bridge to link scientific and normative theorizing, to the benefit of both. The normative force of giving and asking for reasons explains why even the cynical sophist blushes. And that, in turn, should give us hope that, despite the scope of the cynical use of power, justice in practice need not reduce merely to the advantage of the stronger.

References

Audi, Robert. 2004. *The Good in the Right: A Theory of Intuition and Intrinsic Value*, Princeton University Press.

Baril, Galen L., and Jennifer Cole Wright. 2012. "Different Types of Moral Cognition: Moral Stages versus Moral Foundations." *Personality and Individual Differences* 53 (4): 468-473.

Bowles, Samuel and Herbert Gintis. 2011. *A Cooperative Species.* Princeton, NJ: Princeton University Press.

Habermas, Jurgen. 1996. *Between Facts and Norms*. Cambridge, MA: MIT Press.

Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108(4): 814-834.

Haidt, Jonathan. 2010. "Moral Psychology Must Not Be Based on Faith and Hope: Commentary on Narvaez (2010)." *Perspectives on Psychological Science* 5(2): 182-184.

Haidt, Jonathan. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage.

Haidt, Jonathan, and Jesse Graham. 2007. "When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize." *Social Justice Research* 20(1): 98-116.

Haidt, Jonathan, and Craig Joseph. 2004. "Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues." *Daedalus* pp. 55-66.

Haidt, Jonathan, and Craig Joseph. 2011. "How Moral Foundations Theory Succeeded in Building on Sand: A Response to Suhler and Churchland." *Journal of Cognitive Neuroscience* 23(9): 2117-2122.

Haste, Helen. In press. "Deconstructing the elephant and the flag in the lavatory: promises and problems of moral foundations research." *Journal of Moral Education.*

Hastie, Reid, and Bernadette Park. 1986. "The relationship between memory and judgment depends on whether the task is memory-based or on-line." *Psychological Review* 93: 258-268.

Huemer, Michael. 2005. *Ethical Intuitionism*. Palgrave Macmillan.

Joseph, Craig M., Jesse Graham, and Jonathan Haidt. 2009. "The End of Equipotentiality: A Moral Foundations Approach to Ideology-Attitude Links and Cognitive Complexity." *Psychological Inquiry* 20: 172-176.

Kohlberg. Lawrence. 1969. "Stage and sequence: The cognitive-developmental approach to socialization." In D. A. Goslin (Ed.), *Handbook of Socialisation Theory and Research* (pp. 347-480). Chicago: Rand McNally.

Kohlberg, Lawrence. 1971. "From is to ought: How to commit the naturalistic fallacy and get away with it in the study of moral development." In T. Mischel (ed.) *Cognitive Development and Epistemology* (pp. 151-235). New York: Academic Press.

Lazer, D., Neblo, M., & Esterling, K. (2011). The Internet and the Madisonian Cycle: Possibilities and Prospects for Consultative Representation. *Connecting Democracy: Online Consultation and the Flow of Political Communication*, 265-285.

Liu, Brittany S., and Peter H. Ditto. 2013."What Dilemma? Moral Evaluation Shapes Factual Belief." *Social Psychological and Personality Science* 4(3): 316-323

Lodge, Milton, Kathleen M. McGraw, and Patrick Stroh. 1989. "An impression- driven model of candidate evaluation." *American Political Science Review* 83:399-420.

Mericer, Hugo, and Dan Sperber. 2011. "Why do humans reason? Arguments for an argumentative theory." *Behavioral and Brain Sciences* 34(2): 57-74.

Moore, G.E. 1903. *Principia Ethica*. Cambridge University Press.

Narvaez, Darcia. 2010a. "Moral Complexity: The Fatal Attraction of Truthiness and the Importance of Mature Moral Functioning." *Perspectives on Psychological Science* 5(2): 163-181.

Narvaez, Darcia. 2010b. "The Embodied Dynamism of Moral Becoming: Reply to Haidt (2010)." *Perspectives on Psychological Science* 5(2): 185-186.

Neblo, M. A. (2003). Impassioned democracy: the role of emotion in deliberative theory. In *Democracy Collaborative Affiliates Conference*.

Neblo, M. (2005). Thinking through democracy: Between the theory and practice of deliberative politics. *Acta politica*, *40*(2), 169-181.

Neblo, M. A. (2007). Family disputes: diversity in defining and measuring deliberation. *Swiss Political Science Review*, *13*(4), 527-557.

Neblo, M. A. (2007). Philosophical psychology with political intent. *The affect effect: Dynamics of emotion in political thinking and behavior*, 25-47.

Neblo, M. A. (2009). Three-fifths a racist: A typology for analyzing public opinion about race. *Political Behavior*, *31*(1), 31-51.

Neblo, M. A. (2009). Meaning and Measurement Reorienting the Race Politics Debate. *Political Research Quarterly*, *62*(3), 474-484.

Neblo, M. A., Esterling, K. M., Lazer, D., & Minozzi, W. (2012). Logos, Ethos, & Pathos: Mechanisms of Persuasion in a Deliberative Field Experiment. In *APSA 2012 Annual Meeting Paper*.

Pava, Moses L. 2009. "The Exaggerated Moral Claims of Evolutionary Psychologists." *Journal of Business Ethics* 85: 391-401.

Paxton, J.M., L. Ungar, J.D. Greene. 2012. "Reflection and reasoning in moral judgment." *Cognitive Science* 36(1): 163-177.

Plato. *Republic*. Ed., G.R.F. Ferrari. Trans., Tom Griffith. Cambridge University Press.

Prichard, H.A. 1912. "Does Moral Philosophy Rest on a Mistake?" *Mind* N.S., Vol. 21.

Rand D.G., J.D. Greene, and M.A. Nowak. 2012. "Spontaneous giving and calculated greed." *Nature* 498: 427-430.

Ross, W.D. 1930. *The Right and the Good*. Oxford: Clarendon Press.

Sklar, Asael Y., Nir Levy, Ariel Goldstein, Roi Mandel, Anat Maril, and Ron R. Hassin. 2012. "Reading and doing arithmetic nonconsciously." *Proceedings of the National Academy of Sciences* 109(48): 19614-19619.

Suhler, Christopher L., and Patricia Churchland. 2011. "Can Innate, Modular "Foundations" Explain Morality? Challenges for Haidt's Moral Foundations Theory." *Journal of Cognitive Neuroscience* 23(9): 2103-2116.