Moral realism and the heuristics debate

Mark Kelman
Stanford Law School
September 2012

Abstract

There has been substantial debate in the legal academy centered on the questions of whether universal moral intuitions *exist* and, if so, whether these intuitions have a privileged normative status, a debate both reflecting and partly reinterpreting classical jurisprudential debates about the existence of "natural law" and "natural rights." There is a strong but underappreciated homology between the debates about the nature and quality of intuitive *moral* reasoning, and debates, associated with the Heuristics and Biases (H&B) school on the one hand and the "Fast and Frugal" (F&F) school on the other, about the nature and quality of our capacity to make *self-interested* decisions (decisions requiring both *factual* and a-moral *evaluative* judgment and decision making ability.) There are those in the legal academy, most prominently Cass Sunstein, who accept that people indeed often have strong moral intuitions but believe these predispositions deserve little or no normative deference because the intuitions are frequently merely reflect the use of inapt rules of thumb. Other, most prominently, John Mikhail believe people readily make non-reflective moral judgments that we cannot readily explain or justify logically that are grounded in our capacity to process a quite small number of critical features of a decision situation in precisely the way that F&F theorists believe we make most judgments. I explore the degree to which some of the virtues, and, more importantly, most of the problems, in both Sunstein's and Mikhail's work are the features and shortcomings that have bedeviled the work of each of the schools on heuristic reasoning.

I. Introduction

There has been substantial debate in the legal academy centered on the questions of whether (more-or-less) universal moral intuitions *exist* and, if so, whether these intuitions have a privileged normative status because they are reasonably close to universally held manifestations of basic, fixed features of human cognition. Not only does the debate seem both to reflect and partly reinterpret classical jurisprudential debates about the existence of "natural law" and/or "natural rights" – one could readily argue that some particular positive laws reflect "natural law" to the degree that it is in our biological nature as human beings to embrace only certain sets of propositions about the permissibility and impermissibility of particular actions -- but it plainly has some immediate practical, or at least rhetorical, relevance as well as theoretical interest. Binding international legal norms would seem far less like the imposition of the will of powerful sovereigns if they simply expressed beliefs that were both universal and either absolutely unalterable or altered only through labored and artificial efforts to overcome powerful intuitions.[1] Moreover, many questions that might otherwise seem morally (and legally) vexing – the permissibility of various forms of active or passive euthanasia, questions about the permissibility of military actions that jeopardized but did not target civilians, questions about whether torture (if efficacious) is permitted – *might* well be

[1] The supposition that human rights law is more readily justified if its content reflects deeply biologically shared ways of approaching the subjects of such law is made explicit in the work of one of the two main subjects of this essay, John Mikhail, the most prominent of the biologically-focused neo-natural law thinkers in the legal academy today. See especially John Mikhail, "Moral Grammar and Human Rights: Some Reflections on Cognitive Science and Enlightenment Rationalism," in *Understanding Social Action, Promoting Human Rights* (Ryan Goodman, Derek Jinks, and Andrew Woods eds. 2012) and John Mikhail, "Law, Science, and Morality: A Review of Richard Posner's 'The Problematics of Moral and Legal Theory,'" 54. *Stan. L. Rev.* 1057, 1098-99 (2002). One can see further signs of the interest that biologically focused neo-natural lawyers show in the universality of at least some laws in John Mikhail, "Is the Prohibition of Homicide Universal? Evidence from Comparative Criminal Law," 75 *Brooklyn L. Rev.* 497 (2009).

considerably less vexing than they first appear if certain responses to these issues are both strongly counter-intuitive and unstably held because they cannot be "computed" by those (physical) "portions" of the brain (or distributed processes) that create both moral *sentiments* (emotions attached to both norm-compliant and wrongful behavior) and *beliefs* that can aptly be characterized as moral.[2]

One of the most interesting features of this important debate in my view is one that has largely escaped sustained attention: There is a strong homology between the debates about the nature and quality of intuitive *moral* reasoning, and debates about the nature and quality of our *self-interested* reasoning, associated on the one hand with the "Heuristics and Biases" (H&B) school and the "Fast and Frugal" (F&F) school. There are those in the legal academy, most prominently Cass Sunstein, who accept that people indeed often have strong moral intuitions but believe these predispositions deserve little or no normative deference. Sunstein has argued that we ought to be wary of relying on shared moral intuitions, in making public policy or in evaluating our ethical commitments, because many of our intuitions are merely "rules of thumb" that apply poorly in some of the situations in which we apply them. His argument to this effect quite explicitly draws on the "heuristics and biases" tradition.[3] Sunstein self-consciously analogizes "moral heuristics" to the heuristics that H&B theorists argue that people use in making *factual* judgments, especially judgments about the probability that certain events will occur: These judgments, he notes, are triggered by the presence of particular features

[2] It is a difficult to characterize what the domain of "moral" emotions and beliefs might be and how we might contrast an emotion or belief generated in accord with a Universal Moral Grammar (UMG) from other sorts of beliefs or feelings that neither "fit" the purported UMG nor "violate" it; I return to, but by no means purport to resolve, this difficult question.

[3] The basic texts by Sunstein addressing this issue that I will be examining are Cass Sunstein, "Moral heuristics," 28 *Behavioral and Brain Sciences* 531 (2005); Cass Sunstein, "Moral Heuristics and Moral Framing," 88 *Minn. L. Rev.* 1556 (2004); Cass Sunstein, "Hazardous Heuristics," 70 *U. Chi. L. Rev.* 751 (2003).

of the decision environment or by the way we automatically construe the environmental features that generally cue us to make correct judgments but misfire in some substantial number of non-trivial settings. Such judgments tend to be made with little reflection or consideration, especially in situations in which it is difficult to cognize problems deeply. He is also fairly explicit that the sorts of framing/elicitation effects that H&B theorists have argued are in play when people *evaluate* routine end-states are also operative when we try to elicit considered moral judgments, which are, in this respect, merely a particular variety of evaluative judgment. Finally, the basic method he uses to justify the claim that the use of heuristics may be *problematic* is reminiscent of (if not explicitly derivative of) the technique that H&B scholars have used in criticizing the propriety of other frame-sensitive evaluative judgments.[4]

At the same time, John Mikhail's reply to Sunstein and other anti-intuitionists[5] (and others who not only question the normative validity of moral intuitions but argue that no such strong intuitions exist that are not culturally contingent and learned[6]) draws,

---

[4] H&B researchers generally, like all those committed to the existence of a strong fact/value distinction, find it challenging to construct arguments that any evaluative judgment can be in error in the same, transparent way that certain forms of factual judgments can be deemed mistaken. For a fuller discussion of both the distinction between criticizing factual and normative judgments and the techniques the H&B scholars use to criticize evaluative judgments, see Mark Kelman, *The Heuristics Debate* 25-32 (2011).

[5] The texts by Mikhail that I see as most central are: John Mikhail, *Elements of Moral Cognition* (2011); John Mikhail, "Moral heuristics or moral competence? Reflections on Sunstein," 28 *Behavioral and Brain Sciences* 557 (2005); John Mikhail, "Universal moral grammar: theory, evidence and the future," 11 *Trends in Cognitive Science* 143 (2007); Matthias Mahlmann and John Mikhail, "Cognitive Science, Ethics, and Law," Marc Hauser, Fiery Cushman, Liane Young, R. Kang-Xing Jin, and John Mikhail, "A Dissociation Between Moral Judgments and Justifications," 22 *Mind & Language* 1 (2007.) Mikhail's arguments are frequently parallel to arguments made by the psychologist Marc Hauser (though I will try to indicate ways in which they seem interestingly distinct as well.) For a good overview of Hauser's views on "moral realism" see Marc D. Hauser, *Moral Minds* (2006). I will also draw a bit on Frans de Waal, *Primates and Philosophers: How Morality Evolved* to help clarify what I take Hauser's argument to be, in part because I think it is more difficult to understand Mikhail's argument without comprehending Hauser's, and difficult in turn to understand Hauser's without understanding de Waal's.

[6] See especially Mikhail's extensive attack on Richard Posner's general moral relativism and Posner's more particular agnosticism about whether there are morally valid answers to questions about physician-assisted suicide and other issues in which an answer could at least arguably be provided by invoking "double effects" doctrine. Mikhail, "Law, Science, and Morality" supra note --, at 1098-1110 and 1118-26.

if somewhat less explicitly, both on aspects of what I will label massive modularity (MM) theory and aspects of the "fast and frugal school," each of which represents a substantial departure from both conventional rational choice theory and the H&B school that, in ways I will describe, can be quite profitably understood as merely "qualifying" conventional rational-choice theory. Once more, Mikhail is quite explicit that he believes that we are readily able to make non-reflective moral judgments that we cannot readily explain or justify logically that are in fact grounded in our capacity to process a quite small number of critical features of a decision situation in precisely the way that modularists and F&F theorists believe we make many other judgments. He implies, albeit less explicitly, that these specific-cue responsive judgments attend to the most critical features of our decision-making environment. Mikhail argues, more generally, that (near) universal moral judgments reflect the (inexorable) workings of a highly constrained, modularized morality-acquisition system (parallel to the modularized language acquisition system first posited most strongly by Chomsky in proposing the existence of a Universal Grammar and taken up by Fodor[7] and Pinker[8], among others). Mikhail is far less explicit, but arguably implies, that the existence of a "module" to represent a (smaller-than-abstractly-imaginable) set of morality-relevant situations in a particular way is likely adaptive (and that the specific bottom-line moral judgments such fact-representing "modules" render either more plausible or downright inevitable are judgments that serve adaptive ends well.) One might further say that Mikhail implies (rather indirectly) that we should not be too worried that our moral reactions are unalterable, or at least extremely tenacious, given that they are in some weak sense

[7] See especially Jerry A. Fodor, *The Modularity of Mind* (1981).
[8] See especially Steven Pinker, *The Language Instinct* (199-).

adaptive and entrenched in something that could be described as "human nature."[9] But I think it would be far fairer to say with a great deal more assurance that Mikhail's primary mission is descriptive, rather than normative: It is his task to investigate the rules that he believes constrain (or perhaps even determine) both our moral development and our moral reactions rather than to extol the rules that he discovers or the more particular moral views that we observe given the rule-constraints. Still, I will return to discuss briefly several distinct interpretations of the possible normative implications of Mikhail's view that a universal moral grammar both exists and at least constrains the sorts of moralities we intuitively employ.

I have written a good deal both about the nature of the debate between the H&B and F&F schools, and what this debate entails for thinking about a variety of legal issues.[10] In Part II of this paper, I will simply summarize my past efforts to synthesize the heuristics

[9] Hauser seems to me considerably more committed than Mikhail to making a normative argument of the following form: If we can locate some set of capacities that sub-serves moral judgment making that is unique to humans, we can locate something close to the core of "natural" human morality. (See, e.g. *Moral Minds* at 358-359, 411-418) On the other hand, de Waal not only is interested, above all, in showing *continuity* between the behavior of humans and other primates, he never even remotely implies that a particular judgment is "more moral" because it is one that only humans can make. I return to the question of whether he implies that judgment *making* generally, in form if not in substantive content, is more moral when people engage in it because people, uniquely, reflect on judgments and think of them as obligatory.) To be honest, I am never sure what to make either of the argument that what is (most?) natural is "good" (particularly when it appears that the functional translational of "natural" is frequently nothing more than "easily learned") – there are all the problems of deriving an "ought" from an "is" that Hauser adverts to but leaves dangling (see, e.g. *Moral Minds* at 3-4) – or, even more important for the moment, that what is most "natural"(or revealing of "human nature") is what is most *distinctive* to humans.

[10] See Mark Kelman, *The Heuristics Debate* (2011). The book contains a far fuller discussion than I will present here of the underlying debate between those associated with the H&B school and those associated with the F& F school, laying out not just the critical claims each group of scholars makes but the critiques each school makes of the other's work (p. 19-116), as well as discussions of the implications of the debate for thinking about criminal punishment (119-151); the regulation of markets to limit discrimination (p. 155-158) or to increase the degree to which consumers make "prudent" judgments, either through soft paternalist "nudges" or through improving information flow (p. 152-155, 159-177); whether the value of end-states is properly thought of as commensurable (p. 178-201); and whether Langdellian orthodoxy implicitly drew on an F&F-like theory of cognition while Holmesian critics of the classical orthodoxy implicitly embraced H&B positions (p. 202-225).

debate. Then, I will come back, in Part III, to the debate between Mikhail and Sunstein over the nature of moral intuitions.

There are many ways of framing what the debate over moral intuitions entails, and so my most important task in Part III will simply be to try to explicate what I think the most interesting interpretation of what I think that "neo-natural lawyers" like Mikhail are arguing and one version of what those (like Sunstein) who believe we are frequently misled by the use of moral heuristics are arguing. Along the way, I will inevitably consider aspects of what those who believe either that moral judgments are conventional (and learned) believe. I will also make some reference to strongly anti-intuitionist normative arguments that moral beliefs are best thought of as the product of fully rational, self-conscious deliberation, sensitive both to the restricted procedural methods of moral reflection and judging the relevance of arguments with no regard to their ease of accessibility or their immediate intuitive appeal.[11] I hope as well to note some sub-set of the many ways in which the arguments are richer, more qualified, or interestingly distinct from the ideal-types I will ultimately consider. I then return to discuss the connection of this debate to the heuristics debate. Part IV is a brief conclusion.

II. A summary of the heuristics debate

A. Similarities and distinctions between the Heuristics and Biases School and the Fast and Frugal School

[11] Though I in fact largely share the anti-intuitionist belief that the normative persuasiveness of a moral argument does not significantly depend on how well it matches intuitions, however such intuitions are defined, my main goal in this paper is *not* to make the general case against intuitionism but to explicate problems in two distinct ways of approaching moral intuitions that correspond to two distinct ways of approaching other sorts of heuristic-based reasoning. I understand that Sunstein can be read as making one particular form of critique of intuitionism – he can be read as saying intuitions are untrustworthy for a particular sort of reason – but I am more interested in exploring ways in which his approach has the virtues and flaws I associate with the H&B school more generally than using his arguments to bolster the case against intuitionism.

If one stayed at a fairly high level of abstraction, one might argue that everyone interested in heuristics at all thinks about heuristics in the same way: People are employing heuristics whenever they make a judgment or reach a decision without making use of some information that could be relevant or some computational abilities that at least some people possess. Again, looked at quite generally, theorists associated with both the Heuristics and Biases (H&B) school as well as those associated with the Fast and Frugal (F&F) school agree that using strategies that are plainly not formal optimization strategies is, sometimes, absolutely *necessary* because we are incapable of employing formally optimal methods. Many of us can "know" enough about the flight of a fly ball in baseball to catch the ball even though there is lots of (potentially and actually) available information about where a batted ball will land that we don't use at all (e.g. information about wind, spin, the force with which the ball was hit) and computations that many of those capable of catching a fly ball either don't know how to perform or could not perform nearly quickly enough to make use of them (e.g. about how far a ball will go if there is a particular angle of ascent). The one-input heuristic (the "gaze heuristic") we use to "solve" the problem appears to work just fine. People first crudely estimate whether the ball will land in front of or behind them, then run in that direction fixing their eye on the ball. They adjust their running speed so that the angle of gaze – the angle between the eye and the ball – remains constant or within a small range.

At this same high level of generality, too, no one questions that it is often "functional" to use heuristics – using them produces answers that meet our ends well, however these ends are defined – and that they may also, more or less frequently, be used in situations in which their use is dysfunctional. Moreover, there is widespread agreement

that in a multi-actor setting in which one actor may not treat another's interests as if they were her own, the fact that we employ heuristics can be *exploited* by those who have the capacity to manipulate an environment so that it has, or appears to have, traits that trigger a particular judgment, inducing behavior that the manipulator desires rather than the behavior that the agent would engage in if he either had (and used) fuller informational cues or if he encountered the (single or simple) cues that he would have encountered absent the manipulation. Thus, everyone who writes about heuristics worries, at least on occasion, about both advertisers and sneaky lawyers.

Moreover, all agree that it is often easier or preferable to change the environment in which people make decisions or to delegate decisions from a badly positioned to a well-positioned decision maker than to try to change how each individual processes fixed cues: There is a strong consensus, then, that the disposition to use heuristics may typically be rather recalcitrant. If, for instance, patients are more likely to figure out how likely it is that they are actually HIV-positive given that they have tested positive when information is presented in one form rather than another, it might be better to present it in the fashion that most people more typically understand rather than to attempt to train them to "think better," remind them to focus, or even give incentives to do a better job.

The vast bulk of the literature in both law and the policy sciences that has made use of the concept of heuristics has been literature drawing on H&B, most associated with the Nobel Laureate, Daniel Kahneman and with his one-time collaborator Amos Tversky. Those in the "H&B" school are prone to emphasize the degree to which the use of heuristics often leads us to fail to maximize expected value in the way that conventional rational choice theorists believe we do because we both miscompute

probabilities and misevaluate end states. But in many ways, the H&B school can best be seen as *qualifying*, rather than rejecting, conventional rational choice theory: While its proponents believe that we may often fail to maximize (appropriately risk-discounted) expected value when we make choices, they do not really deny either that we typically do, or should, seek to do just that.

Proponents of those who think of heuristics as "fast and frugal" techniques to make decisions that achieve an organism's ends in a given environment whether the problem-solving techniques are formally rational or not, most associated with Gerd Gigerenzer, are considerably less interested in "biases" or errors than in *achievements*. They typically emphasize the degree to which the heuristics that we use will far more typically, though not invariably, be not only adequate to the decision-making tasks at hand, but typically even be superior to formally rational decision-making, given the interplay between our capacity sets and the actual features of the problems that we confront in the environments in which we must solve problems.[12] Interestingly, to the

[12] If one wanted to use a single, Take the Best, fast and frugal heuristic to distinguish the schools – perhaps merely in ironic tribute to the F&F scholars who speak frequently of the purported tendency of people to make a judgment or reach a decision attending only to one single best cue-- one could probably say that the "heuristics and biases" people are conventional political liberals and that "fast and frugal" optimistic functionalists are conventionally conservative. *Everyone* notes, as I said, that the use of heuristics can misfire in particular situations, and (nearly) everyone has a (broadly) similar evolutionary story for this: cognitive capacities that served us well in the circumstances in (the hunter-gathering) environment in which the evolved may serve us poorly in modern life.

It is no great surprise, though, that when optimistic functionalists like Cosmides and Tooby search for an example of how functional hunter-gatherer capacities sabotage us in the modern world, they pick on programs (advocacy of rent control) that conventional political conservatives attack for perfectly conventional politically conservative reasons: just another case of good-hearted, mushy liberals missing the unintended consequences of their misguided efforts to help the poor. But instead of *describing* this form of misdirected empathy as sentimental ideology gone bad or as a pernicious power-grab by self-interested state bureaucrats interested in expanding their own power or securing their jobs, they tell us it is the (rare?) case of misfit between our hunter-gatherer intuitions [to help those who are victims of misfortune that they could not avert, so that they will help us when we are similarly victimized] and modernity…

At the same time, it is no great surprise that writers in the heuristics and biases literature often throw the kitchen sink of familiar liberal complaints about Western market and political culture at you when (ostensibly merely) trying to emphasize the point that even if our judgment heuristics were "good enough" to deal with many of those tricky hunter-gatherer conundrums, they aren't quite up to the complex

extent that law and policy academics were even aware of the F&F school, they tended to treat it as supplementing rational choice school attacks on one aspect of the H&B school, generally first offered by people defending consumer sovereignty and pluralist politics: the H&B school's pessimism about the capacity of consumers (or voters) to make welfare-enhancing choices if fully informed. While it is true that F&F scholars' work often gave comfort to those who saw less need for "experts" to intervene to protect people against almost-inexorable imprudence, H&B work is actually structurally more closely aligned with rational choice theory than F&F work is, for it is the F&F scholars who disclaim far more thoroughly the idea that people make choices designed to maximize satisfaction, quantifying and weighing the value of a multitude of option traits given their own idiosyncratic tastes, rather than making choices grounded in single simple cues (e.g. mimetically, habitually, with regard to the option's most salient feature). H&B theorists think we are unskilled rational choosers; F&F people think we are quite skilled at making good choices, but not because we choose as rational choice theorists postulate.

I believe that the most important distinctions among the schools can be understood if we see that they answer the following sorts of questions differently:

- What is each theoretical school fundamentally trying to explain? To what extent does the theorist start with an idealized picture of judgment and decision-making,

---

tasks of modernity: Thus, the following is an entirely typical "defense" of the idea that non-adaptive uses are ubiquitous by a partisan of the heuristics and biases tradition, Keith Stanovich: "Meliorists [his term for the people I am describing as proponents of the heuristics and biases program] see a world seemingly full of shockingly awful events – pyramid sales schemes going "bust" and causing financial distress, Holocaust deniers generating media attention, $10 billion spent annually on medical quackery, respected physical scientists announcing that they believe in creationism, savings and loan institutions seemingly self-destructing and costing the taxpayers billions – and think that there must be something fundamentally wrong in human cognition to be accounting for all this mayhem."

Still, I think this wholly "political valence" contrast is ultimately not especially instructive or true.

typically associated with rational choice theory, and then look to see how frequently there are departures, why they occur, and how one would describe the non-ideal mechanisms? To what extent, instead, does the theorist start with the supposition that our judgment and decision-making processes developed to solve a concrete set of problems in the environments in which we must solve problems, so that our task is first to understand the *fit* between cognitive capacity and environmentally-established problems?

- What criterion does each school use in evaluating whether a judgment or decision-making process is "rational"?
- To what degree do theorists in a particular school believe that judgment and decision-making is (mildly, substantially, or absolutely) "informationally encapsulated"? Are people capable of "overriding" heuristics when they make a judgment, using cues beyond the informationally limited ones that would trigger a particular judgment outcome if they simply employ a particular heuristic?
- Similarly, to what extent does the theorist believe that we can think about problems using "generalized," non-problem-specific cognitive mechanisms, and if the theorist believes that there are (at least some) *general* cognitive mechanisms, how should these mechanisms be described and what is their functional domain?
- To what degree does the theorist see the use of heuristics as arising almost exclusively from limitations on *internal* mental processes – time, attention, computational power – and to what degree does the theorist emphasize instead the limits on the number of significant naturally occurring tasks that could be solved

using ordinary optimization methods, even by an unlimited mind? Would we use heuristics less frequently if we were somehow "smarter"?

- Does the theorist assume that all functional adults are equally likely to use both useful and dysfunctional heuristics? If some people with particular traits (e.g. higher intelligence, conventionally defined; certain personality traits that are generally associated with "open-mindedness") are less prone to use some dysfunctional heuristics, does this imply that we use heuristics because some, but not all of us, are computationally limited or inadequately motivated to solve problems "well"? Would individual differences (if real) suggest that people use heuristics because of their internal limitations, not because they are the ideal decision making mechanism given the features of the external environment? Does the fact that some people "avoid" using heuristics more than others imply distinct things about what rationality is and whether the use of heuristics is rational (under a host of distinct definitions of rationality)?

B. Descriptive notes on the heuristics and biases school

What I think is most critical for lawyers and policy-makers to understand about the heuristics and biases school is that it is framed, fundamentally, as a critique of the realism, but not the desirability, of making decisions in accord with the dictates of classical rational choice theory. At core, what rational choice theorists both counsel and observe is that, as a prelude to a choice between two options, each of us should (and often either does, or tries to) assess the *probability* of each ultimate outcome that might arise if a particular action-option is taken and the *value* of each such outcome. It is rational to

choose that action-option that maximizes the expected value of the possible outcomes, weighting preferences about risk-seeking or risk-avoidance appropriately.[13]

At any rate, if people are to perform the task of selecting an option that maximizes expected utility (setting aside risk preferences), one must assess accurately the probability that each of a series of conceivable outcomes would arise if one chose a particular option. Thus, the first aim of the H&B researchers was to show that people did *not* assess probabilities in a fashion that was likely to reflect the best available information about the probability of future events. People may have *thought* they were assessing how frequently some event X, not Y, would occur on the basis of how often it had occurred in the past, but their judgment of how often it had occurred inaccurately reflected the actual relative frequency of X and instead reflected things like its availability or its representativeness or the fact that one anchored to some prior estimate of frequency (even a rather transparently arbitrary and uninformed one) and adjusted inadequately. At core, people *substitute* one feature of a cue (e.g. its availability or representativeness) for the more rationally relevant one (its probability.) For instance, when using the availability heuristic, individuals estimate the frequency of an event and therefore the likelihood of its past occurrence (or future recurrence) "by the ease with

[13] It is an important point, in thinking about the contributions of the heuristics and biases school generally, but not so much in thinking about the contributions most central to the issues I raise in this article, that H&B scholars believe that the traditional account of risk-preferences is wildly inaccurate, so that thinking about subjects as trying to maximize expected utility given certain attitudes towards risk is quite misleading. But the H&B material on the infirmities of conventional rational choice theory about risk proclivity and aversion – Kahneman and Tversky's "prospect theory" – is largely outside the scope of the debates between H&B and F&F theorists that I am attending to. For a fuller discussion, see *The Heuristics Debate*, p. 10, p. 247-248, note 3. For the classic discussion of prospect theory, see Daniel Kahneman and Amos Tversky, "Prospect Theory: An Analysis of Decision Under Risk," 47 *Econometrica* 265 (1979). For the classic F&F response, arguing that neither conventional rational choice theories of risk proclivity *nor* prospect theory accurately describe how people make choices involving uncertain outcomes, see Gerd Gigerenzer and Ralph Hertwig, "The Priority Heuristic: Making Choices Without Trade-Offs," 113 *Psychol. Rev.* 409 (2006).

which instances or associations come to mind."[14] While making frequency judgments on the basis of availability will typically work well – people typically most readily recall events that they have been exposed to frequently and they most typically been exposed most frequently to events that occur most often – it may misfire. Events may be easily recalled, for instance, when they are emotionally salient, even though they are infrequent. Heuristics users are, in the H&B view, like any people making use of a generally accurate but over and under inclusive rule or proxy.

According to H&B theorists, not only do people often fail to assess probabilities accurately, they often do so in a fashion that is logically incoherent. It is, of course, generally more straightforward to detect incoherence than inaccuracy; assessing inaccuracy requires that the experimenter herself knows the actual probabilistic distribution of the phenomena at issue.[15] For example, people who judge probabilities on the basis of the representativeness of an outcome might believe that it is more likely that 1000 people will perish in an earthquake in California in the next twenty years than that 1000 people will perish in a natural disaster West of the Rockies, though an earthquake in

[14] For a fuller discussion, see *The Heuristics Debate* p. 21-25. The classic works in the H&B tradition are Amos Tversky and Daniel Kahneman, "Availability: A Heuristic for Judging Frequency and Probability," 5 *Cogn. Psychol.* 207 (1973 (on availability, the ease of recalling an event or attribute); Daniel Kahneman and Shane Frederick, "Representativeness Revisited: Attribute Substitution in Intuitive Judgment," in *Heuristics and Biases* 49 (T. Gilovich, D. Griffin and D. Kahneman eds. 2002) (on representativeness, the degree to which the event is prototypical of the kind of events the subject is trying to recall); and Amos Tversky and Daniel Kahneman, Judgment under uncertainty: Heuristics and biases," 185 *Science* 1124 (1974) (on anchoring, the tendency to adjust inadequately from some preliminary estimate of frequency, probability or value, even when the initial estimate to which one anchors is a transparently irrelevant figure).

[15] One may, of course, be mistaken even when one makes perfectly coherent, contingent judgments. It *may* simply be wrong that there are fewer English words beginning with "r" than words whose third letter is "r", even though most of us think the opposite, because we can more readily think of words beginning with "r", but the belief is not *logically* wrong.

California is included in the set of natural catastrophes West of the Rockies so it cannot be more probable than the set in which it is included.[16]

Not only do H&B researchers detail ways in which people fail to assess accurately (or even coherently) the probability that certain outcomes will arise if they choose a particular option, they also attempt to demonstrate that people may make "mistakes" in *evaluating* the end states whose probability of occurring, given any course of action, they have already assessed, however inaccurately. Given conventional commitments to the gap between (objective) fact and (subjective) value, the criteria H&B authors can use in criticizing a value judgment are at once both narrower and almost invariably more controversial than the criteria for critiquing a factual judgment. Value judgments are most obviously troublesome when they violate coherence rationality – they are, for instance, intransitive or violate dominance rules. Not surprisingly, then, H&B researchers frequently attempt to demonstrate that the use of heuristics generates intransitive preference orderings or violations of dominance rules.

Further, and more significantly, the H&B theorists typically argue that they need not have substantive views on what tastes are "objectively preferable" to argue that people are not evaluating end-states properly if the evaluation of such end-states is frame-sensitive. H&B theorists have been especially adept at exploring situations in which some end-state X is evaluated as better than Y if the outcome X is described in one fashion but not another or if X is evaluated as better than Y only if there is some irrelevant third

[16] For H&B discussions of this sort of conjunction fallacy, see, e.g. Amos Tversky and Daniel Kahneman, "Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment," in *Heuristics and Biases*, supra note – at 19.

alternative Z present as part of the option set. Once more, much of the H&B literature focuses on just these sorts of framing effects.[17]

Of course, H&B proponents want to be able to critique evaluative mechanisms even when they don't generate either incoherent preference-orderings or demonstrate irrational frame sensitivity. While unwilling to adopt full-blown perfectionist critiques that declare that some choices are substantively unacceptable, they are prone to argue that the choices made by subjects who are "misusing" heuristics are apt to regret their choices, and that the regret bespeaks a substantive problem. Obviously, whether regret bespeaks "error" is hardly obvious; we may forget or discount the gains that prior choices generated and focus unduly on the unwanted aspects of our current situation that might have been avoided when we evaluate decisions retrospectively.

Since they believe that people will frequently fail to behave "rationally," the question that arises is: "Why? There is a dominant generalized story that goes something like this: Our brains have two "systems." Cognition that occurs in System One (including the rationality-distorting heuristics) is associative, effortless, unreflective, rapid, intuitive, and fairly automatic or tacit rather than conscious. Virtually all (functioning) adults engage in System One cognition pretty much equally well.[18] System Two thinking is, in this view, pretty much the opposite: It is at core rule-based, analytical, conscious and

[17] One of the most familiar H&B heuristics (grounded in "endowment effects" and "loss aversion") tells us that the same mortality outcome may either be deemed preferable to or inferior to some other outcome depending on whether the outcome is described as saving a certain number of lives or resulting in a certain number of deaths.

[18] Many (but by no means all) H&B theorists believe that System One thinking is highly contextual rather than abstract. People engaging in System One thinking are unable to draw inferences about situations they have not directly experienced simply on the basis of the formal features of the situation. The canonical example comes from anthropology. An illiterate Uzbek (with high reliance on System One thought?) is presented with a syllogism: "In the Far North, where there is snow, all bears are white. Novaya Zemlya is in the Far North and there is always snow there. What color are the bears there?" The respondent could not answer, but merely stated that he had only encountered black bears in his own experience and could not speculate on what bears would look like in places he'd never been.

explicit. It requires hard work, and tends, therefore, unlike System One thinking, to be disrupted by distractions, stress, and time pressure. It is less sensitive to the factual content and context of propositions than to the formal analytic properties of these propositions and what the propositions logically entail. Generally, H&B theorists imagine that System Two works to insure more rational judgment by sometimes overriding and sometimes accepting System One intuitions.[19]

At any rate, the capacity to engage in System Two thinking is influenced not merely by situational mediators (like time pressure or distraction) but by innate or learned individual distinctions in the *capacity* to engage (in more situations) in System Two thinking.[20]

C. Sketching the features of the "fast and frugal" school

H&B theorists typically start with the assumption that people do and should seek to make conventionally rational decisions, and fail to do so because they lack the *internal* resources (time, attention and computational power) to do so. F&F theorists are far more prone to emphasize that making formally rational decisions does not inevitably serve the

[19] For a discussion of the System One/System Two model, see *The Heuristics Debate* p. 32-38. The classic work on the dual system hypothesis is Stephen Sloman, "The Empirical Case for Two Systems of Reasoning," 119 *Psychol. Bull.* 3 (1996)

[20] Thus, people who are trained in statistics are (modestly) more likely to override the use of (many) heuristics. Similarly, people who are more "intelligent" (in the sense measured by traditional "g-loaded" tests, like IQ tests or the SATs) use many of the heuristics less frequently. The point, for this group of H&B theorists, is not that the "sort" of intelligence that g-loaded tests measure is the only sort of relevant intelligence (or even the most important), but that it is a genuine measure of *something*. That something appears to be the capacity to manipulate non-contextualized formal symbols in accord with the dictates of conventional rational choice theory. For the classic discussion of individual differences in the use of dysfunctional heuristics, accompanied by the claim that the fact that "smarter" people use dysfunctional heuristics less frequently belies the claim that heuristics are the source of superior performance, see Keith E. Stanovich and Richard F. West, "Individual Differences in Rational thought," 127 *J. Exp. Psychol: General* 161 (1998) and Keith E. Stanovich and Richard F. West, "Individual Differences in Reasoning: Implications for the Rationality Debate," 23 *Behavioral & Brain Sciences* 645 (2000). See also *The Heuristics Debate* p. 34-38.

organism's goals; thus, we ought not to optimize in the fashion H&B theorists suggest we should even if we had limitless computational powers.[21]

Broadly speaking, the F&F researchers believe that one *cannot* employ optimizing efforts when a decision task has some or all of the following traits: the problem may be computationally intractable, pay-offs from the projected outcome of the decision are ambiguous, and the future is uncertain.[22]

What one can see, more generally, then, is the F&F people do not start with the assumption that our goal is or should be to be logical – to follow abstract, context-free norms. We do not and should not seek *logical* rationality, we do and should seek *ecological* rationality. We do and should seek to use our inevitably limited capacities in such a way that we meet our ends, and we do so by having developed cognitive capacities that fit our environment. When an environment provides certain readily processed cues that can lead to decisions that lead to choices that meet our ends, it is of little moment whether or not our views are as veridical as they could be (if we accounted for more cues) or as logically consistent as they might be.

---

[21] Of course F&F people frequently and forcefully emphasize that optimizing is not feasible because of limitations that could best be described as internal.

[22] It is not central to my task of setting out the basic features of the F&F school, but it worth noting that I find all three of these points problematic: Problems are not intrinsically intractable or tractable, rather than tractable or not relative to some intellectual skill set; values may not be incommensurable in relevant senses; and the fact that the future is uncertain seems to be less of an argument against optimization than it is an argument against modularization, which commits us to making the same judgment once a certain cue is present even if the environment in which the cue is perceived has altered. Often, though, it seems that the F&F argument about the uncertain future is not really so much an argument against general efforts at optimization, but an argument against particular forms of statistical reasoning. It is, Gigerenzer repeatedly (and rightly) notes, troublesome to rely on regression equations that fit (or, as he rightly puts it, over-fit) a particular data set. It can indeed be misleading to establish relationships between some dependent outcome variable V and a host of factors that have been present or absent in the past when V occurred if our goal is to predict whether V will occur in the future. This is true because many of the factors that seemed to influence the occurrence of V were accidentally related on a single, non-recurring occasion, or the relationship between some of these factors and the occurrence of V will alter. There may, instead, be a small number of cues that persistently co-occur with V, even in a changing world, but many others that do not: heuristic decision makers may focus on the few best cues that turn out to be persistent.

F&F researchers not only posit that boundedly rational thought arises in a particular fashion, but that boundedly rational thought has typical structural features. At core, the structural features are as follows: The subject first follows a simple search rule. This rule tells her what cues to look for. She then employs a simple stopping rule that tells the subject that she needn't search for more cues, either because she has learned enough to make a decision that reaches an aspiration level or because she has found an informational cue that provides her with adequately accurate information. Finally, she uses a simple decision rule that directs her to take the action that the positive cue value specifies. Think in this regard of one of the simplest of the heuristics: the recognition heuristic and think about it in the context of a subject trying to make a judgment about which of two cities has a larger population. Structurally, what I want to emphasize is that the subject using the recognition heuristic employs a simple search rule (search first for the city whose name one recognizes), a simple stopping rule (stop looking for other cues to city size if one recognizes one city in a pair and not the other), and a simple decision rule (decide that the recognized city is more populous.)

The cognitive process envisioned by F&F researchers is not *strongly* informationally encapsulated in the sense used by massive modularity (MM) theorists.[23]

[23] I discuss "massive modularity" (MM) at a bit more length in *The Heuristics Debate* 59-67. Generally speaking, MM theory was designed to revitalize the traditional idea that the mind possesses specific "faculties" rather than a more general capacity to learn and reason. The modules – which give us the capacity to process domain-specific information and make domain specific decisions – have a number of critical properties: not only are they domain specific not in the trivial sense that all knowledge could be said to relate to a particular subject but in the sense that there are distinct algorithms to process cues about each appropriately delimited category. They are also mandatory (cues trigger reactions automatically), opaque (we don't have a good sense of why we came to the judgments or made the decisions we do), and *strongly* informationally encapsulated. MM theory solves certain recurring quandaries in F&F theory: while F&F theorists have never really figured out how subjects choose whether a heuristic should be used or which heuristic should be used, MM theorists posit that certain mandatory perceptions of the external world trigger the activation of the relevant module, without the need to learn anything about the relevance of the cues, and that they do so because it has proven adaptive, in an evolutionary sense, to act on these automatic triggers. There are many reasons to believe this MM picture is implausible – it is very difficult to figure out

A decision about city size, for instance, is not committed to a module that cannot be penetrated by any information but recognition information. But heuristic-based cognition is "*softly*" informationally encapsulated in the sense that people typically will "stop" once they have found the discriminating single cue rather than incorporate any additional non-recognition information once they have passed their "stopping point."[24]

The interesting point for now is how F&F researchers have reacted to H&B findings that people in fact *do* use compensatory information, in terms of how they model heuristic reasoning. Some argue that the relative city size judgment, for instance, is only *sometimes* made heuristically, and that when it is, it is made without the use of compensatory information. Thus, from this vantage point, the interesting question is how we define the *domain* in which we will use heuristics, not what it means to use heuristics (or a particular heuristic) *if* we are using one. Conceptually, the problem is a quite thorny one: If we need non-modular (or "slow and informationally rich" rather than "fast and frugal") cognitive processes to determine *whether* to assign a cognitive task to a module (or heuristic decision-making process) and, if so, to what "module" (or heuristic) to assign it, then it is not at all clear that we should describe *cognition* on the whole as either modularized or heuristic. Full-blown rational choice theory plainly contemplates the use of rules of thumb (single cues) *when* the decision maker thinks them apt or sufficient. If F&F (and MM) differ from rational choice theory (with or without dysfunctional biases)

---

what perceptually salient cues are supposed to trigger the activation of most of the modules MM theorists discuss (e.g. if we have a "cheater detection module," as most MM theorist postulate that we do, situations that pose the possibility of cheating do not have a proprietary trait, like a color or a taste, that we can instantly perceive) – but it does solve the heuristic selection problem F&F scholars face.

[24] I explore in detail in *The Heuristics Debate* the unambiguous finding – both in my own experiments and the experiments of other researchers – that subjects actually use non-recognition information in a compensatory fashion when assessing things like (and including) relative city size. (That is, they sometimes will believe that a non-recognized city is bigger than a recognized one.) See Mark Kelman and Nicholas Richman Kelman, "Objections to the Fast and Frugal School Program," in *The Heuristics Debate* 90.

it is because these critics of rational choice theory believe that subjects need not generally *choose* what sort of decision-making process (or how many cues) to use.

D. Cross-cutting critiques: what the debaters emphasize

1. F&F critiques of H&B work

At core, the most basic critique that F&F theorists level at H&B research is that subjects *seem* to perform sub-optimally in H&B experiments only because they are given problems in these experimental settings that do not mimic problems that they would confront in natural environments. What ultimately *creates* the gap between performance on "real world problems" and laboratory problems is that the mental capacities that evolved are the capacities to solve recurring problems that increase inclusive fitness, not the more general capacity to be an abstractly better calculator (e.g. of expected values). In this view, H&B researchers fashion lab problems that merely test formal problem-solving capacity and then interpret formal failures on these problems as functional failures. Whatever its ultimate *origins*, though, the gap between good "real world" performance and bad lab performance may be *manifest* in four distinct ways:

- *H&B theorists may present material in a fashion that is formally mathematically equivalent to an alternative presentation that subjects would find more tractable.*

In experiments that the F&F theorists believe are vulnerable to this particular critique, subjects indeed make what even F&F theorists concede are "mistakes." That is to say, in this class of cases, the F&F scholars are not arguing that the subjects' answers are "better than rational." However, the mistakes, they say, come from the artificiality of the way in which the problem is presented. The fact that the subjects make mistakes in the lab setting does not imply that they will typically make mistakes coping with problems "of a

similar sort" in ordinary life. The material the H&B experimenters present might well be more tractable if presented in the manner that it is (ostensibly) confronted in natural settings, generally, or at least in the natural settings that were prevalent when humans developed their cognitive capacities. This criticism was perhaps most prominent in disputes over whether people would exhibit the sort of base rate neglect that H&B theorists had demonstrated if the information had been presented in frequentist rather than probabilistic fashion.[25]

- *A sub-set of material that is formally, mathematically equivalent to other material may be less readily solved because – though formally equivalent – it does not involve the solution of a problem that we have learned to solve (without understanding the formal mathematically or symbolically identical computations involved) because of its practical importance in increasing inclusive fitness*

Once more, the basic idea here is that we solve the problems we solve using dedicated problem-solving algorithms, not by reducing all problems to a form in which they are tractable for a general computing machine. We can thus demonstrate that people are poor problem solvers if we give them problems they have little reason to solve in real life (or at least real life in the Environment of Evolutionary Adaptation or EEA), even though solving the problem seems to involve no more formal math skill than solving

[25] According to F&F theorists (as well as some Massive Modularists), people have a great deal of trouble processing information presented in the following (probabilistic) form that H&B researchers had presented it in: "99.8% of those who are HIV-positive test positive. Only .01% of those who are not HIV-positive test positive. The base rate for the disease among heterosexual men with few risk factors is .01%. How likely is it that a particular low-risk factor heterosexual man is HIV-positive if he tests positive?" On the other hand, most people find it relatively easy to deal with the same information presented in the following (frequentist) way: "Think about 10,000 heterosexual men with few risk factors for acquiring HIV. One is infected, and he will almost certainly test positive. Of the remaining 9999 uninfected men, one will also test positive. Thus, we'd expect two of the ten thousand men will test positive and only one of them has HIV. So what are the chances that the person who tests positive is infected?" For a fuller discussion, see *The Heuristics Debate*73-75; for a discussion of H&B responses (including the response that it is not "frequentist" presentation that helps make the problems tractable but set inclusion), see *The Heuristics Debate* 267-268, note 12.

problems that they solve readily when the problems must be solved to cope with a practical predicament. We do not really solve those practical problems by first reducing them to abstract, generalized mathematical form; instead, we have domain-specific solution techniques to solve them. Not surprisingly, given the prominence of the task in debates over the general persuasiveness of evolutionary psychology, one of the key disputes in this area centers on poor performance on the abstract, but not cheater-detection form, of the "4 card" Wason selection task.[26]

- *Subjects may make what appear to be "mistakes" playing games with formal pay-off rules because the "games" resemble real-world problems in which the pay-offs are subtly distinct from the pay-offs that are defined in the formal game and people solve the "real world" (mild) variant of the problem that they have been presented, rather than the precise problem they have actually been presented*

  Once more, in this class of cases, the F&F researchers concede that the experimental subjects perform poorly on the task they have been given. That is to say, once more, the behavior is not "better than rational" given the precise pay-off structure of

[26] At the high level of abstraction (that H&B theorists associate with System 2 thinking), *all* selection task problems might be seen as the same. (Some H&B theorists are skeptical of the claim that all "selection task"" problems are indeed formally identical, but my main point for now is to clarify the F&F critique, so one should assume that is at least plausible to describe them as invoking the same formal solution procedures.) If given a proposition of the form, "If P, then Q" a person who wants to take the steps necessary to discover whether the proposition is true must investigate both whether the Ps he encounters always entail Qs *and* whether some of the not-Qs he encounters are accompanied by Ps. He need not, though, investigate whether some not-Ps are accompanied by Qs *or* whether some Qs are accompanied by not-Ps because the rule is not violated in those cases. This is true whether the proposition is of the form, "If a card has an even number on one side, it has a vowel on the other" (the abstract 4-card Wason selection task form) or of the form, "If you are drinking beer, you must be over 18" (the "cheater detection" form) People do quite badly figuring out what steps they need to take to find out if the first, more abstract 4-card selection task proposition is true. Most subjects know you have to turn over the card showing an even number to discover if there is a vowel on the other side but very few recognize you have to turn over the card with a face-up consonant to make sure it doesn't have an even number on its flip side. On the other hand, far more people solve the problem in the second "cheater detection" form: They know that they must both check beer drinkers to make sure they're over 18 *and* check 17 year olds to make sure that what's in their glass is root beer, not beer. For a fuller discussion of this controversy, see *The Heuristics Debate* 75-79.

the laboratory game. Fundamentally, they do so, however, because they ignore the instructions they have been given – they have confronted these instructions for the first time in the experimental setting – and instead assume that they are playing a game whose pay-offs are those that obtain in "games" that resemble the laboratory game that they either play often in real life, or played often when people developed relevant cognitive capacities. The debate over "probability matching" is especially instructive in understanding this aspect of the dispute between F&F theorists and H&B researchers.[27]

- *Subjects may appear to make computational "mistakes" because they reinterpret the experimenters' instructions or assume that the experimenter has implied more than he has explicitly stated: making these sorts of conversational implications is a necessary part of being able to communicate (and, of course, being able to communicate is adaptive)*

F&F researchers often argue that H&B researchers have assumed, incorrectly, that subjects are giving non-normative responses to a set of questions they intended to ask,

[27] Assume that experimental subjects are shown an urn with 70 green and 30 yellow balls. They are told that 10 balls will be drawn from the urn, and the ball that is drawn will be put back in the urn after it is drawn. Subjects are asked to guess which color ball will be drawn on each of the ten occasions. They win a prize for each correct answer. Rational subjects should pick green all ten times (unless the subject has non-monetary goals, e.g. a desire to keep himself more interested in the contest): The expected value of choosing green for all ten selections is 7 (you've got a .7 chance each and every time.) Most people, though, choose green seven times and yellow three: that is to say, they engage in what is usually dubbed "probability matching" for the set, making their choices match the most probable outcome of ten draws. They do so even though the expected value of that choice is .7 X 7 + .3 X 3 or 5.8 rather than 7.
One *could* figure out what choices to make using some (undefined) general cognitive mechanisms (that permit the calculation of expected values in all sorts of situations). Alternatively, one might have developed (at least relatively) domain-specific cognitive mechanism to solve the problem of picking an optimal mix of distinctly risky gain-seeking activities from a small option set that dictates that one will engage in probability matching. F&F theorists, echoing evolutionary psychologists prone to believe that people have developed narrow domain-specific "answers" to problems that presented themselves to our ancestors facing evolutionary pressures argue, for instance, that the "cognate" problem to the urn problem in a natural environment is to pick between foraging sites with distinct probabilities of finding food. The optimal strategy in that setting may not be to maximize expected value, though, but to both get more food and to learn more about unexplored environments, at least when one is satisfied that one has gone to enough high-odds sites to insure that one will be a bit flush with food. (I should note that I remain utterly befuddled by the claim that experimental subjects should be expected to "confuse" these two games.) For a fuller discussion, see *the Heuristics Debate* 79-81.

when they are really giving normatively appropriate responses to the questions that a socially adept communicator, interpreting linguistic cues as they would ordinarily be interpreted in real conversation, believes have been posed. It is important to note what are really two separable points: First, subjects may be giving perfectly good answers to the questions they hear (even if there is no compelling reason for them to interpret the questions as they do). Second, as a matter of fact, their interpretations of the questions the experimenters pose are typically more sensible, given general norms concerning how we draw implications from literal language that are necessary for communication to proceed. One can probably understand this particular general controversy well by reflecting on certain F&F critiques of the conjunction fallacy experiments.[28]

- *While the most central criticism that those associated with the F&F school level at H&B researchers is that they see irrationality where it does not ultimately exist, or find it in settings of little or no practical moment, it is important to note*

[28] F&F critics argued that those who (ostensibly) committed the conjunction fallacy in the "Linda problem" did not do anything problematic, even though they believed it more probable that Linda was a feminist bank teller than a bank teller, though the former is a sub-set of the latter. (They did so, from the vantage point of H&B theorists, because Linda was described as having had traits in college far more prototypical of a feminist than an ordinary bank teller and then made judgments of probability based on "representativeness."). Instead, they were actually behaving more intelligently by observing the standard Gricean norms about conversation and reinterpreting the "intended" question. Grice posits that those committed to a cooperative principle of conversation that permits listeners to draw proper inferences from words spoken in a conversational context assume that what we offer our conversational partners must be relevant. According to the F&F critics, rational social creatures recognizing the cooperative nature of Gricean conversation would not think that the experimenter would have offered information about Linda's left-wing politics or counter-cultural style *unless* the experimenter intended to signal that she was indeed a feminist bank teller now (maxims of both relevance and quantity are implicated): thus, the "conjunction fallacy" response is normative, not irrational, in accounting for implicit information that those who avoid the fallacy simply neglect.
Another way of putting the point is that the subjects hear a different question than the experimenters claim to have asked. At core, the claim is that those who make appropriate inferences from the prior "conversation" (in which they have already been told about Linda's past political/cultural identity) is to hear (or read) the explicitly uttered phrase "Linda is a bank teller" as "Linda is a bank teller but not a feminist." (It is also plausible, in this view, that subjects hear the statement "Linda is a bank teller" as an implicit conditional – i.e. "*If* Linda is a bank teller, she is a feminist"). For a fuller discussion, see *The Heuristics Debate* 81-83.

[29] The truth is, the H&B theorists have ignored the F&F theorists far more than the other way around. I am constructing many of the H&B critiques of F&F theory far more fully than they have been constructed in the literature.

that they also perpetually complain that the H&B theorists neither explain why people use the precise heuristic problem-solving mechanisms that they allegedly use, nor do they typically define the mechanisms in adequate detail.

Their *explanation* for this second deficiency in the H&B program is pretty similar to the explanation of the perceived failure of H&B theorists to test performance on "real world" problems. F&F theorists start (like all influenced by variants of evolutionary psychology) with the idea that mental capacities are adaptive. Given that preconception, they believe we are most likely to be able to identify mental capacities/mechanisms not simply by observation, but by reasoning backwards. We should first note the "need" (in inclusive fitness terms) that the organism had to meet and then intuit the capacity it must have developed to meet that need.

Because, for example, H&B theorist do not typically even attempt to specify precisely what adaptive role it might have played to make certain forms of (purportedly bad) judgments – e.g. to neglect base rates, to encode gains and losses asymmetrically, to assess probabilities on the basis of availability – they (purportedly) have more difficulty describing the form base rate neglect may take. On the other hand, the F&F "adaptive toolbox" approach *starts* with the supposition that we can identify a series of tools, with some precision, that would have been useful in increasing reproductive success. These *are* the heuristic mechanisms.

Whatever the cause of the (purported) problems that beset H&B research, it is plain that F&F theorists frequently note critically that the H&B heuristics are poorly defined, very hard to operationalize, and – as a result – give us little to work with if we want to make predictions that can be falsified or verified.

2. H&B critiques of F&F work

At core, the most fundamental critiques articulated by heuristics and biases (H&B) researchers of the work associated with the fast and frugal (F&F) school simply mirror or reverse the F&F critiques. While F&F theorists deride H&B theorists because they (purportedly) fail to account adequately for the ways in which cognition is adaptive to the problems people actually face, the H&B theorists think that the F&F scholars' fixation on the ways in which capacities must be adaptive may often lead the F&F theorists badly astray.

The most contentious claim H&B scholars make[29] is that F&F theorists are simply wrong when they declare that they offer descriptions of the heuristics people use that are both more detailed than those H&B theorists provide and more accurate. Instead, say the H&B critics of the F&F school, the heuristics the F&F people identify are frequently inaccurate *idealizations* of actual capacities or cognitive strategies – ungrounded both in behavioral observations and in neurobiology – that merely restate (imputed) adaptive *goals* as-if they were capacities. To put that point another way, H&B scholars believe to a considerable extent that the F&F theorist (too) typically describes a heuristic or cognitive process without regard to its real nature, but only as the projected solution to the adaptive problem the F&F theorist *imagines* the organism both needed to solve and must have solved in the fashion the theorist projects.[30] It is vital to recognize that this derogatory

[29] The truth is, the H&B theorists have ignored the F&F theorists far more than the other way around. I am constructing many of the H&B critiques of F&F theory far more fully than they have been constructed in the literature.

[30] In *The Heuristics Debate*, I discuss this point at great length in the context of the "recognition heuristic." In "discovering" the recognition heuristic, Goldstein and Gigerenzer *start* with the proposition that it would serve adaptive ends to have "the capacity" to "merely recognize" (or fail to recognize) things, in a simple on-off binary way, very hastily. (This form of "recognition" is the adaptive tool in the Gigerenzian toolbox that people will be able to make use of.) They then assume that the capacity to make judgments about city size based on the recognition heuristic (identify immediately which of two cities one "recognizes" and then

observation echoes a perfectly common refrain in critiques of evolutionary psychology (EP) more generally: Instead, of observing a trait, say critics of EP, EP researchers selectively observe behavior and "see" the attributes that they believe they ought to find, given adaptive "needs."

Second, different people, with different cognitive abilities and "thinking styles," may systematically use heuristics differently. This fact is at least mildly incompatible with a number of aspects of the F&F view.[31] Finally, they abjure the F&F commitment to even soft versions of encapsulation: they see attributions substitution as the main heuristic mechanism, not lexical thinking

---

decide, without further reflection, that the recognized one is larger) simply builds on "this capacity". So starting with this picture of what they (probably wrongly intuit) would be a useful free-standing skill to have, they describe the (purportedly observed) mental processes that subjects solving the city-size determination task use as the instantiation of that skill. In doing so, they ignore neurobiological and experimental evidence that tells us (among many other things) (i) that what most psychologists and neuroscientists who study memory call familiarity judgments (which they call 'mere recognition' judgments) are not on-off binary judgments but (loosely) frequentist (i.e. that we encode information about roughly how often we've confronted stimuli, not just information about *whether* we have confronted the stimulus or not); as a result, many items will be very mildly familiar, but not so familiar that a person will inevitably consciously describe the item as recognized; (ii) that the city recognition task – which requires not merely recognition of the proper name but associational learning/contextual memory (what is traditionally called 'recall' memory rather than 'familiarity') – largely involves different brain regions and distinct cognitive processes than performing the simple familiarity recognition tasks they describe and claim are all that is being used in city recognition; (iii) that even the simplest familiarity tasks are not really performed solely by some isolated input-recognition module, but rather that our capacity to encode inputs as familiar is partly dependent on non-recognition cognitive capacities and that the capacity to make familiarity judgment sub-serves other cognitive tasks as well, rather than being a fully isolated task. Thus, even setting aside for now the equally profound problem that they are wrong to claim that subjects then make city size judgments without regard to further non-recognition information, what they have arguably done wrong – what H&B theorists suspect F&F researches do wrong so often – is that they have not given a more accurate picture of the cognitive process of "recognizing a city" but rather (attempted to) induce behavior by assuming that it must meet certain imputed adaptive ends. For a much fuller account, see Mark Kelman and Nicholas Richman Kelman, "Objections to the Fast and Frugal Program," in *The Heuristics Debate* 90, 93-104.

[31] This claim, initially explored largely by Stanovich and West, is detailed in *The Heuristics Debate* 34-38, 92-93,110-112. The claims that people who do better on g-loaded tests and who are more "open-minded" are less prone to make the sorts of mistakes that H&B theorists heuristics uses are prone to make are connected to several more general propositions about heuristics: First and foremost, it is thought to belie the claim that heuristics users are better-than-rational decision makers. But it also thought to be incompatible with the claim that heuristics are used not because of internal computational deficits but because they permit users to draw the best conclusions possible. Finally, it is thought to be incompatible with the notion that adaptationist pressures led to the use of the heuristics since one would expect their use to be far closer to universal if they had.

III. Moral realism (natural law v. moral heuristics)

A. Mikhail's moral grammar

1. Mikhail's account and critique of the mainstream modern legal academic view of "moral realism"

One might well rest on firmest ground if one merely noted that there are certain beliefs about morality that Mikhail thinks are both commonplace in the legal academy and highly misleading. In his view, those wedded to the wrong-headed orthodoxy reject the proposition that people are naturally able to acquire only a sub-set of abstractly conceivable moral beliefs and reject the cognate idea that there is some set of reasonably concrete beliefs about when behavior is either morally permitted, obligatory, or prohibited that are shared by all (reasonably healthy/functional) persons, without regard to either cultural background or idiosyncratic ideological disposition. The orthodox skepticism takes on several forms, and each, Mikhail believes, must be rejected.

First, proponents of the mainstream position, says Mikhail, wrongly reject claims of descriptive universality (across cultures, across classes, genders, and races) of any set of rational beliefs about social ordering and morality.[32] Mikhail believes this rejection of the existence of universals arises not so much from the discrete findings of cultural anthropology as the disposition of cultural anthropology as a discipline to both emphasize the local and particular and to avoid, over-assiduously, succumbing to the perils of intolerance or ethnocentrism that is entailed whenever one declares that any set of values that are widespread in the anthropologists' cultures of origin are universal.[33] But the skepticism also comes from the sense, more derived from liberal political theory than

[32] See, e.g. "Law, Science, and Morality," supra note – at 1062, 1064, 1066, 1087, 1106-07.
[33] Mikhail makes reference to this problem in "Is the Prohibition of Homicide Universal?" supra note – at 513-14.

anthropology, that even within relatively homogenous cultures, moral battles are ubiquitous: In fact, in this view, creating a functioning liberal state requires ordering political life so as to avoid the need to resolve the clash between people holding a host of diverse, particularized sectarian moral beliefs that, if pushed into the public sphere, would merely cause unmanageable strife.[34]

Second, Mikhail believes that the mainstream position expresses a certain psychoanalytical skepticism that moral *beliefs* are profoundly distinct from emotions, thus implicitly denigrating the "status" of whatever commonalities of reactions one might perceive. The visceral, unexamined emotion of disgust might be triggered in all people by presenting (perfectly good) food in usually toxic *forms*. (There really are experiments in which we try to get people to eat chocolate shaped like dog feces.) It is hard, though, to think of the disgust reaction as a *belief* (on what is it premised? from what is it derived?) let alone a rational belief that survives reflection or represents behavior governed by the self-reflective desire to conform one's attitudes and behaviors to any set of normative commitments. Once more, one needs to face the question of whether neo-natural law theorists, like Mikhail, either need to, or do, take a uniform position on whether "natural" morality functions more like cognition, classically understood – the simple competence to *recognize* that to follow rule X and not rule Y is moral -- or more like an emotion, classically understood, akin to sexual attraction or hunger in the sense that it engenders

[34] Mikhail makes reference to the fact that Rawls, both in *Political Liberalism* and in his revision of *A Theory of Justice*, deemphasized the search for a universal moral grammar, analogous to Universal (linguistic) Grammar, arguing that the principles of justice he extolled were largely grounded in the political needs of a particular sort of liberal state to deal with the problem of heterogeneity and ideological/religious conflict. See *Elements of Moral Cognition*, supra note – at --.

more than the *competence* to recognize the attractive or the hunger-satisfying but tends to *impel action*.[35])

Moreover, there is plainly fMRI evidence (of the usual uncertain quality) that certain reactions to "moral dilemmas" either tend to invoke emotions or be triggered by their presence (at any rate, parts of the brain associated with emotional reactions are maximally activated when these responses are given) while the opposite reaction to the dilemma is associated with activity in regions of the brain more typically associated with "reason."[36] Whether "moral universals" are best understood as cognitive or emotional

---

[35]Hauser quite plainly believes that in most situations involving what he thinks of as moral decision making, "emotions" do not so much impel what we come to see as our beliefs as they are triggered by (something that could be seen as *prior*) belief reactions, See, e.g. *Moral Minds* at 8, 30, 46, 52-53. But his position on this issue is hardly unambiguous or univocal. See, e.g. *Moral Minds* at 25, 223. Hauser further states on some occasions that while moral competence is, by and large, generated, by a morality-acquisition "module" uniquely dedicated to moral acquisition, the emotions that impel action in response to moral reactions are "general purpose" emotions. See, e.g. *Moral Minds* at 52-53. However, he notes on other occasions that there is currently no real evidence that there are *any* parts of the brain, or even capacities, which are used *solely* to deliver moral judgments so the fact that generally available emotions might be recruited into moral judgment does not seem as crucial to his argument about the "peripheral" status of emotions at some times as it does at others. See, e.g. *Moral Minds* at 221.

[36] There is a (very small) cottage industry that could be described as advancing the proposition that deontology is emotional, and consequentialism reasoned. See, for instance, Peter Singer, "Morality, Reason, and the Rights of Animals," in Frans de Waal, *Primates and Philosophers* 140, 146-151. The captain of this industry is Joshua Greene. .

Greene, working with numerous collaborators, has made a number of arguments that are purely descriptive: he posits that those giving what he (only partly correctly) characterizes as deontological responses to Trolley problems that preclude people from pushing a person to his death from a Drawbridge to save five others are using System One, emotion-based decision making processes. On the other hand, (purportedly) utilitarian/consequentialist respondents (who will divert the trolley on to a Spur Track, killing one, to save five lives) are making use of more uniquely human System Two rational processes. Greene's views can be seen in, for example, J.D. Greene, R.B. Sommerville, L.E. Nystrom, J.M Darley, & J.D. Cohen, "An fMRI investigation of emotional engagement in moral judgment." 293 *Science* 2105 (2001); .Joshua Greene and Jonathan Haidt, "How (and where) does moral judgment work?" 6 *Trends in Cognitive Science* 517 (2002).While acknowledging the broad proposition that one cannot readily draw normative implications from facts, Greene has drawn implications much like those mentioned in the texts on many occasions. See particularly, Joshua D. Greene, "From Neural 'Is' to Moral 'Ought': What are the Moral Implications of Neuroscientific Moral Psychology?" 4 *Nature Reviews Neuroscience* 847 (2003) and Joshua D. Greene, "The Secret Joke of Kant's Soul," in *Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* 35 (Walter Sinott-Armstrong ed. 2007).

Greene's view about what features of Drawbridge are critical to eliciting the different response, has changed to some extent over the years. But he remains broadly committed to the idea that the reaction to Drawbridge is more "primitive" in the literal sense (i.e. shared with other primates that evolved before humans did) and the reaction to Spur Track more reflective and uniquely human. [See Greene and Haidt, 2002, p. 519 for an early formulation.] When first presented, Greene's notion seemed to be that the

Drawbridge scenario triggered an "alarm-like" emotional reaction [Greene, 2008a, p. 41] that was both strong (precluding rational reflection) and immediate (it literally takes less time and deliberation to make) [Greene, et. al. 2001, p. 2016-17] and that it did so because the agent whose behavior was being evaluated was engaging in "up close and personal" violence. Deontology, to Greene, is typically just long-winded *ex post* rationalizations for System One, modularized emotional reactions; deliberative thinkers are utilitarians. [See, e.g. Greene and Haidt, 2002, p. 522.] The following snippet gives a good sense of his views on this: "(D)eonotlogical philosophy, rather than being grounded in moral reasoning, is to a large extent an exercise in moral rationalization. This is in contrast to consequentialism, which…arises from…psychological processes…that are more 'cognitive,' and more likely to involve genuine moral reasoning." See Greene, "Secret Joke," id. at

Greene's arguments that we can evaluate how persuasive an argument is by looking at its neurobiological provenance have been subjected to considerable critique. See, e.g.. Selim Berker, "The Normative Insignificance of Neuroscience," 37 *Philosophy & Public Affairs* 293 (2009). Berker argues not only that Greene's arguments are unpersuasive on their own terms – that he has not actually demonstrated that people reason "differently" when using deontological reasoning than they do when reasoning as utilitarian consequentialists (id. at 302 -313) -- but that it would not provide a normative reason to prefer utilitarian arguments if he had. Berker argues that this is true whether the normative argument is stated as an unadorned preference for "reason" over "emotion" (id. at 316-317); as an appeal to the notion that deontological reasoning resembles generally unreliable heuristic judgment and decision making (id. at 317-319); or that deontological intuitions were adaptive in an environment that no longer exists or covers most of the moral problems we actually now face (id. at 319-321.) He thinks even the best version of the argument that one can evaluate normative positions in terms of the thought processes that prototypically generate them -- that deontological intuitions are responding to morally irrelevant factors – fails, largely because the neurological arguments do not help us identify what factors are and are not relevant. (id. at 326-7.)

I think Berker is far too hasty in dismissing the argument from heuristics, and that he actually reveals his hesitations about his own argument by positing the plausibility of a different, more optimistic picture of heuristic reasoning, typically associated with the "fast and frugal" heuristics school that we have discussed than the more wary picture that Greene, more closely associated with the "heuristics and biases" school, endorses. The core of the problem in his argument is easily stated, even if developing the critique would go far beyond the goals of this paper. Berker has offered no reason to believe that judgment processes that lead to what would unambiguously be deemed factual or logical errors are not *prima facie* less trustworthy when they generate moral judgments, even if we lack simple criteria to identify when people are making "errors" in the moral domain. For reasons I explore in the text, Sunstein explicitly rejects the idea that deontologists alone are those stuck using all these silly, misapplied heuristics, though it would be hard to read his work without seeing that he is attracted to the idea. See, e.g. Sunstein *Moral heuristics* at 533-34.

At times, Greene seems to treat the fact that people would react emotionally to personal violence as arising in part from squeamishness at either direct physical involvement in, or some other sort of proximity to, another agent's suffering. Greene and colleagues p. 367-8 note that subjects are more willing to open a trap door through which the Fat Man falls on to the track, blocking the trolley, than to push him on to it. See Joshua D. Greene, Fiery Cushman, Lisa E. Stewart, Kelly Loewenberg, Kelly, Leigh E. Nystrom, and Jonathan D. Cohen, "Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment," 111 *Cognition* 364, 367-8 (2009). For distinct efforts to identify what it might mean to be directly physically connected to harm causing, see, e.g. Fiery Cushman, Liane Young, & Marc Hauser, "The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm," 17 *Psychological Science* 1082, 1090 (2006) (focusing on physical contact); Edward B. Royzmon and Jonathan Baron, "The preference for indirect harm," 15 *Social Justice Research* 165, 167-8 (2002) and Baron, (focusing on direct v. indirect harm); and Adam Moore, Brian Clark, and Michael Kane, (2008) "Who shalt not kill? Individual difference is working memory capacity, executive control, and moral judgment," 18 *Psychological Science* 549, 550-2 (2008) (focusing on the distinction between "direct" harms and those mediated through either other people or technological/mechanical means) Berker, *Normative Insignificance* at 323 argues that is wholly indeterminate what will trigger the sense that one has engaged in personal violence or not. Berker notes, for instance, that Greene and his co- authors do not test whether people would think it all right to get the Fat Man to jump by menacing him with a knife or

threatening his family, but not touching him. He further notes that subjects might conflate judgments that one *shouldn't* push the Fat Man with judgments that it would be imprudent to do so since he might resist; this prudential worry might disappear in the trap door case. Id. at 323.]

At other times, Greene has *defined* "up close and personal violence" more idiosyncratically. So defined, it is the sort of violence that primates generally would have developed an emotional aversion to in order to facilitate social cooperation. In this view a moral violation is "personal" if the actor does something that is likely to cause serious bodily harm to a particular person and the harm is not a result of simply deflecting an existing threat on to a different party. One can think of the criteria for personal violence as "ME HURT YOU" and as delineating roughly those violations that a chimpanzee could appreciate. The "HURT" condition identifies roughly the kinds of direct harms that a chimp can apprehend (e.g. assault, not tax evasion or pollution); the "YOU" condition requires that the victim be clearly identified as an individual (rather than being, say, the sort of "statistical life" that a polluter might "kill,") and the "ME" condition captures the idea that moral responsibility occurs only when the agent "authors," rather than "edits" the harm-causing situation, that he is the determinative agent. [Greene and Haidt, 2002, p. 519.] These two definitions are not, of course, coextensive: Findings that subjects are marginally more willing to pull a switch that makes the Fat Man on the bridge drop through a trap door to block the runaway trolley than to push him resonate in the view that people seek to avoid getting their hands dirty, but do not resonate in the idea that what people seek to avoid is violating the "ME HURT YOU" principle. Moreover, Greene now acknowledges that subjects disparage actions – like destroying a tea cup to save a greater number of tea cups, diverting a trolley car on to a sidetrack where it will be stopped from returning to the main track by a man, rather than a rock that when hit will crush the man – that implicate neither version of emotion-inducing personal violence but rather simply rely on an unwillingness to permit the infliction of intended harms. [See Joshua Greene, "The Secret Joke of Kant's Soul," in *Moral Psychology Vol. 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* 35-79. (Walter Sinnott-Armstrong ed., 2008) The tea cup experiments to which he is responding are reported in Shaun Nichols and Ron Mallon, (2006) "Moral dilemmas and moral rules," 100 *Cognition* 530 (2006) though the experiments demonstrate a gap between a certain sort of moral judgment – subjects believe pushing the single tea cup violates a rule, id. at 535 – and "all-things-considered" prudential judgments – subjects still believe the tea cup should be pushed, id. at 536, 538. The trolley on a loop track studies are initially reported in Mikhail, *Universal moral* grammar at42.]

Greene has proffered many distinct forms of evidence to bolster these basic findings: fMRI studies that purport to demonstrate that portions of the brain associated with emotion rather than reason are activated when responding to the drawbridge-like problems[e.g. Joshua Greene and Jonathan Haidt, (2002) "How (and where) does moral judgment work?" 6 *Trends in Cognitive Science* 517, 518-9 (2002)]; studies that purport to show that utilitarian, but not deontological rsponses, are interfered with by increasing cognitive load [Joshua D. G reene, Sylvia A. Morelli, Kelly Loewenberg, Leigh E. Nystrom,. and Jonathan D. Cohen, (2008) "Cognitive Load Selectively Interferes with Utilitarian Moral Judgments," 107 *Cognition* 1144 (2008); studies showing that people who are primed towards a positive emotional state are more prone to think it is acceptable to push the man off the drawbridge than those who watch a neutral clip while the clip has no impact on divert-the-trolley responses [Piercarlo V aldesolo and David DeSteno, (2006) "Manipulation of Emotional Context Shapes Moral Judgment," 17 *Psychological Science* 47 (2006); studies showing that subjects who have experienced selective brain damage, interfering with brain regions generally associated with emotion rather than cognition, are considerably more likely to give consequentialist responses than those who do not exhibit such damage [Michael Koenigs, Liane Young, Ralph Adolphs, Daniel Tranel, Fiery Cushman, Marc Hauser, and Antonio Damasio "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgments,"446 *Nature* 908 (2007)]; and reaction time studies [e.g. Joshua D. Greene, Leigh E. Nystrom, A.D., Engell, J.M. Darley, and Jonathan D. Cohen, (2004) "The neural bases of cognitive conflict and control in moral judgment," 44 *Neuron*, 389 (2004).

It is always hard for an outsider to serve as a legitimate arbiter of disputes in other disciplines, but my judgment would be that the critiques of Greene's work on reaction time distinctions are pretty devastating. The most persuasive criticism is found in Jonathan McGuire, Robyn Langdon, Max Goltheart and Caroline Mackenzie, "A Reanalysis of the Personal/Impersonal Distinction in Moral Psychology Research," 45 *Journal of Experimental Social Psychology* 577 (2008). Here is the basic problem: Greene and his colleagues do *not* find that those people who refuse to push the Fat Man respond significantly more quickly than those who decide to divert the train. What they find, instead, is that, on average, people who

depends in part, I think, on what we count as a moral universal. As I will explore, I ultimately think in Mikhail's preferred conception that the universals are most akin to parameter maps, the cognitivist account is more plausible.

Third, there is a distinct, commonplace "modernist" intuition about human nature that Mikhail also rejects, borne in significant, but by no means exclusive, part in the reaction of psychologists, historians and political scientists to the political horrors of the twentieth century. In this view, the problem is not so much that people are born without fairly strong predispositions, but that imposed positive law (and a host of other forms of counterintuitive socializing force) is needed to *overcome* the quite problematic "natural" tendencies people have to harm others.[37] This jaundiced view of the relationship between moral codes and human nature, dubbed "veneer theory" by de Waal, is that while we do indeed have many strong intuitive reactions, our intuitive reactions are to be brutishly selfish, at least outside the treatment of family, kin, and perhaps those in an ethnic group (positing that ethnicity is some form of kinship-extended grouping). Short-term helpful

---

claim it inappropriate to take a life saving step in a range of cases do so more quickly than those who think it appropriate. But that result could be – and is – driven by a handful of cases in which it takes subjects an especially long time to make a judgment that an action is appropriate or very little time to find it inappropriate. Greene and his colleagues did not look, item by item, at whether respondents to each individual question, took longer if they gave one answer rather than the other. So they could get the result they got merely by adding in a small number of questions in which most people thought it appropriate to take life-saving steps that were long, complex, and hard to work through; they could (and did) also get the result by including problems where most thought it *inappropriate* to cause a death and did so really quickly.

The issue remains contentious. Greene and his colleagues acknowledge that subjects exhibit the same reaction times to judgments that certain steps are inappropriate rather than appropriate if one merely deletes from the initial study sample two especially easy cases to find that harm-causing action is inappropriate (one in which a man considers hiring a burglar to rape his wife so that he can regain her love by comforting her after the crime and one in which a man considers killing his boss in what will appear like an accident because it will "get him out of your life"). But Greene and his colleagues seem to feel that the only reason it is appropriate to delete the cases is because the response that the actions would be appropriate is not a utilitarian response at all. See Joshua D. Greene, Syliva A. Morelli, Kelly Loewenberg, Leigh E. Nystrom and Jonathan D. Cohen, "Cognitive Load Selectively Interferes with Utilitarian Moral Judgments," 107 *Cognition* 1144, 1146 (2008). Their response seems to miss the point: These two cases are responded to so quickly because they are atypically *easy*.

[37] Mikhail hints at his skepticism about this aspect of the modernist critique of the "moral faculties" in *Elements of Moral Cognition* xv, 262 as well as "Law, Science, and Morality," supra note --, at 1077, 1079.

behavior may be "naturally" stable – either when the party helping another bears no costs as part of an immediate cooperative venture (mutualism) or when (generally near-term) reciprocity from the benefited party is more-or-less guaranteed – but if we are to expect "moral" judgments in other situations, we need to rely on counter-intuitive institutional pressures.

Fourth, Mikhail believes it is important to reject what he sees as the commonplace idea that those committed to any variety of "moral realism" must believe that people *discover* moral rules in the external world. Such "external" "rules" might be established by external, presumably divine authority or, more interestingly, be thought of as external because the rules are, akin to physical laws, the only set of rules that could conceivably govern functioning social relations. Mikhail clearly notes that he is not committed to any sort of "mind-independent moral reality." In his view, the source of whatever universalized moral injunctions we observe is within our minds, the cognitive mechanisms that permit us to acquire moral beliefs.[38] This claim is interestingly related to claims about the significance of (whatever) universalism of beliefs we might discover. One can imagine a theorist who thoroughly rejects Mikhail's claims about cognitive moral competence – believing instead that moral beliefs are some mixture of learned conventions and the product of independent reasoning performed by relatively fully cognitively flexible minds. Such theorists could still believe that we would observe at least some "universals" if they believed that only certain moral "systems" would survive and permit adequate levels of social cooperation to sustain a culture: Universalism would result, in this view, from some combination of "cultural evolutionary processes" (survival

[38] See, e.g. *Law, Science, and Morality* at 1088-89. *Elements of Moral Cognition* 220-221.

of only those *groups* following certain norms) and/or the convergence of rational cognitively generalist minds on the small set of rules that were actually workable.[39]

In terms of the normative significance of Mikhail's claims, it would obviously strengthen the idea that our "intuitions" have at least presumptive normative force if, first, the intuitions that we had developed were maximally adaptive, and second, if being maximally adaptive were either equivalent to being normatively preferable or at least a strong factor to be weighed in judging the acceptability of a norm. Neither proposition is transparently correct, either in my mind or in Mikhail's. It is a commonplace in evolutionary biology that traits that evolve need not be the best traits imaginable to solve a problem. If nothing else, there are developmental biological limits on the range of plausible mutations. And it is not clear why we should judge the ultimate moral acceptability of a belief solely in terms of the fit between that belief and an individual's chances of passing on his genes. Inclusive fitness and moral acceptability may simply be skew to one another.

2. Mikhail's positive program: moral competence and the linguistic analogy

a. The basic analogy

At core, Mikhail's view is that all people, everywhere, are born with a sort of "moral competence" and that this moral competence is fairly closely parallel to the linguistic competence that permits people to learn *a* language. While the particular languages people learn are obviously not identical (so that languages are in some sense conventional rather than universal), the capacity to learn *a* language is grounded in the

[39] I am dubious whether this sort of claim could be sustained if thinking about some of the rules that Mikhail thinks are imbedded – e.g. rules about double effect. But, of course, this skepticism would have to be extended to the implicit evolutionary claims in Mikhail's work that a certain UMG dictating or facilitating a particular set of moral rules either produced direct reproductive advantages or was an explicable side-effect of a cognitive structure that did produce such inclusive fitness gains.

competence to recognize a host of things, such as the distinctions between sentences and pseudo-sentences, ambiguous and clear expressions, whether one phrase (even a made-up one) merely paraphrases another. Naturally, in thinking about the questions that preoccupy legal theorists, questions about the degree to which any set of precise moral reactions to novel, concrete problems (of the sort that might be instantiated in substantive rules of law/morality) are determined by the limits on our moral competence are central. Still, it is clear that if the proposition that we are born with the capacity (and predisposition) to represent problems that typically raise moral concerns in a particular way is to have substantial pragmatic significance, then many abstractly conceivable moral rules must be effectively off the table, "unlearnable" as moral rules once the problems are so cognized. Certain rules ought not to match the representational structure we apply to situations raising the relevant set of moral concerns.

How might Mikhail's basic viewpoint be profitably distinguished from claims that could be described as more capacious, whether about the domain of judgments that could be dubbed moral, about the impact of moral judgments on all sorts of reactions that could be labeled as aspects of behavior, about the universality of concrete moral judgments, about the sorts of competencies that should be characterized as moral competence rather than more general cognitive competence that can be co-opted in establishing or implementing moral rules?

b. The domain of morality and moral "judgment"

i. Substantive and procedural accounts of what counts as a moral judgment

First, one must figure out what sorts of judgments – and for these narrow purposes, both concrete behaviors and emotional reactions are no different from

"judgments" -- should be called "moral" in the sense that Mikhail is interested in. One possible theory of the "moral domain" is at core "substantive" and another at core "procedural." Alas, I am not certain either account is especially stable in de-limiting the domain appropriately.

The substantive view, quite clearly articulated by de Waal and occasionally embraced, albeit more ambiguously, by Hauser, is that "moral" rules are those rules, and those rules only, that mandate that parties account for the interests of others, even when there is no clear short-term benefit to doing so.[40] de Waal is occasionally indefinite about the extent to which one can be said to make a moral judgment unless one has extended the domain over which concern for the interests of others is manifest from kin and neighbors to a wider group,[41] but prior to reaching the question of whether one has made a moral judgment unless it is directed at either helping *any* worthy person or behaving punitively towards an unworthy person regardless of her relationship to the person making the judgment, one must first decide that only judgments that instantiate Golden Rule-like obligations and some set of corollary rules about what constitutes a violation of such obligations are the stuff of morality. (de Waal is also somewhat ambiguous about

---

[40] This position is taken quite unambiguously by de Waal. See *Primates and Philosophers* 162 ("…the moral domain of action is Helping or (not) Hurting others…Anything unrelated to the two H's falls outside of morality. Those who invoke morality in reference to, say, same-sex marriage or the visibility of a naked breast on prime-time television are merely trying to couch social conventions in moral language.") At times, it appears that Hauser endorsees this view as well. See, e.g. *Moral Minds* at 290, 358, 410, but there are other times at which Hauser seems to favor a broader conception of what a moral norm is (e.g. a norm that permits a child to learn *some* sort of rule against "incest," however locally specific the content or contours of the anti-incest rule counts as a moral rule). See id. at – . At still other times he describes the moral domain *procedurally*, in ways I allude to the in the text.
Obviously, even if one decides that the "substantive" domain of morality is the domain of non-self-interested behavior, we can imagine many significant distinctions in non-self-interestedness: distinctions in intensity (how much would one sacrifice), scope (whose interests may come ahead of one's own) and skill (the ability to discern accurately the interests or needs of others.) See Philip Kitcher, "How to Get Here from There" in *Primates and Philosophers* 120, 127-28.

[41] He explicitly notes that some degree of partiality (or loyalty) is itself a moral virtue. *Primates and Philosophers* at 165, but thinks of the extension of "moral codes" to wider circles as a significant, if generally fragile, aspect of morality as well. Id. at 53-54.

whether he embraces or rejects the claim that judgments are not truly moral unless the intuitions – even if wholly altruistic – have been subject to self-critical reflection,[42] and ambiguous whether he thinks it is possible to apply "moral rules" to an adequately wide domain of people unless one is self-reflective in that way, but for now, that point is somewhat less important.[43])

If one takes this particular substantive view of what a moral rule is, rules against incest, for instance, are *not* moral rules, except to the degree that they are designed to protect the object of lust against one agent acting on selfish desires that the other agent would prefer not be manifest. They would not be aptly characterized as moral rules, I take it, even if some variety of incest prohibition were universal, even if the prospect of

---

[42] See *Primates and Philosophers* 173-175. ("The desire for an internally consistent moral framework is uniquely human. We are the only ones to worry about why we think what we think…I consider this level of morality, with its desire for consistency and "disinterestedness" and its careful weighing of what one did against what one could or should have done, uniquely human.") At the same time, he is quite suspicious of accounts of morality that emphasize self-critical reflection rather than automatic, unprocessed emotional sentiments that lead to the sorts of substantive (Golden rule altruistic) dispositions that he essentially defines as moral. See id. at 178-179. ("Philip Kitcher and Christine Korsgaard are correct to stress the importance of knowing the motives behind behavior. Do animals ever intentionally help each other? Do humans?...We are excellent at providing *post hoc* explanations for altruistic impulses. We say things such as, "I felt I had to do something" whereas in reality our behavior was automatic and intuitive, following the common human pattern that affect precedes cognition…We may therefore be less intentionally altruistic than we like to think. While we are *capable* of intentional altruism, we should be open to the possibility that much of the time we arrive at such behavior through rapid-fire psychological processes similar to those of a chimpanzee reaching out to comfort another for sharing food with a beggar.")
Hauser, too, often implies that what makes humans truly moral is a degree of self-conscious reflection and understanding that others (generally) have rights and that the agent herself is responsible if she violates those rights. He approvingly cites Rawls' argument that what distinguishes social cooperation from socially coordinated activity is that social cooperation is guided by publicly recognized rules viewed by those who follow them as appropriate. *Moral Minds* at 414-15. He further emphasizes that this self-reflective capacity may well "have played a pivotal role in our capacity to sustain large-scale cooperation with unrelated individuals." Id. at 415.

[43] I think, but am not at all confident, that de Waal believes that the extension of altruism to a wider circle of people can only occur because we have the capacity to reflect on "good" behavior and develop reasoned rules: See *Primates and Philosophers* at 54-55. ("Instead of merely ameliorating relations around us, as apes do, we have explicit teachings about the value of the community and the precedence it takes, or ought to take, over individual interests…If we accept this view…of morality as a logical outgrowth of cooperative tendencies, we are not going against our own nature by developing a caring, moral attitude…In other words, we are not hypocritically fooling everyone when we act morally: we are making decisions that flow from social instincts that are older than our species, even though we add to these the uniquely human complexity of a disinterested concern for others and for society as a whole."

violating the prohibition typically generates sentiments/emotions similar to those generated by the prospect of other "moral" rule violations, even if people could not articulate a defense for their intuitions about the impropriety of incest, and even if, developmentally, very young children manifest some of the capacities needed to implement any functioning incest taboo (e.g. the capability to differentiate siblings from non-siblings) before they had been "taught" to do so. I take it, as well, that they would not aptly be considered moral even if they were, in the other "procedural" sense I attend to soon, thought of as something-other-than-customary and therefore not capable of being waived by whatever authorities declare or articulate local custom.

Once more, I am not especially confident of my judgment in this regard, but I believe that Mikhail, following Hauser, does not limit the domain of what counts as a moral judgment in precisely this substantive way: Hauser offers only the most cursory and cryptic description of what he sees as the basic Universal Moral Grammar (UMG). His account is thus far less detailed in helping us reason about constraints on moralities than the narrower accounts of *aspects* of the UMG that Mikhail offers for the handful of moral judgments – most particularly those instantiating his particular conception of the doctrine of double effect -- whose cognitive structure Mikhail believes is at least reasonably well-understood. But Hauser's rather thinly specified UMG centers at core only on the fact that moral judgments are constrained by a set of rules that dictate that those making moral judgments of acts that lead to harms will inevitably evaluate an agent's intent and goals and the positive and negative consequences that ensue from the agent's actions[44] (Cultural variety occurs for Hauser because, for instance, the

[44] One sees this rather minimalistic account of moral grammar in *Moral Minds* in a number of places. See, e.g. 47-48, 310, 411.

"meaning" of different consequences may vary widely across cultures). Still, this account of the UMG seems to permit a broader array of "topics" to be covered as moral topics. Hauser certainly explicitly treats incest rules as part of our "moral minds."[45]

Perhaps it is perplexing that while Hauser's UMG seems thinner and less outcome-determinative than I think Mikhail's is – seeming to permit greater meaningful cultural diversity of content rules – he also seems to think there are many more near-universal content rules than Mikhail is willing to say that we can identify with confidence at this point, even when the relationship of these rules to the UMG that he does posit is, by my lights, very hard to fathom. For instance, Hauser treats it as something like a universal that people will know it is wrong to "cheat" on their primary lovers.[46] Presumably, if we follow the general argument in his work, he believes that all of the critical terms that give meat to this barebones injunction – what counts as cheating, who counts as a primary lover, what constitutes a commitment sufficient to activate the idea that one is cheating if one has sexual contact with a non-primary lover – are culturally variable. But he nonetheless does seem to believe that a rule of this broad form is universal, even though it seems hard to derive from the UMG as he, or Mikhail for that matter, describes it.[47]

---

[45] See *Moral Minds* at 22-23, 166, 301.
[46] See *Moral Minds* at 5.
[47] It is possible, of course, that a rule against cheating on primary lovers *is* universal – like a taste for sweets or an aversion to snakes – as a more particularized domain-specific adaptation rather than as a rule that can best be understood as an instantiation of our moral competence.(The typical largely unsatisfying evolutionary psychology stories would go something like the following: We are drawn to sweets because our ancestors in the Environment of Evolutionary Adaptation were rightly worried that they would be calorie deprived and not survive to pass along their genes if they didn't ingest high calorie foods. Women don't "cheat" because uncertainty over paternity makes it impossible to get child rearing resources from any particular man and men penalize cheaters because they are afraid of getting swindled and rising a kid with someone else's genes; men don't cheat despite their desire to father as many kids as possible because women punish those who can't be constrained to help protect the kids borne of their scarce eggs.) Because the anti-cheating rule *looks* like a rule that could be described as "moral" (in at least some of the many

The primary procedural view distinguishing moral from non-moral judgments is that a moral rule is merely any sort of rule that is not deemed merely "conventional" by those who follow it. A rule is conventional in the relevant sense if those who follow it believe it to be so, and thus believe that it could be waived by someone with the "authority" either to dictate the norms that must be followed or to articulate those that have spontaneously been followed in the relevant local culture.

Mikhail (echoing Hauser or de Waal in this regard) plainly believes that the *capacity* to distinguish moral from conventional rules is one of the key in-born forms of moral competence.[48] The fact that very young children can distinguish between a moral rule ("don't hurt your school mate") and a conventional one ("don't wear your pajamas to school") even when both have been articulated as rules, whose violation is subject to punishment, is a critical piece of evidence for him that we are born with significant moral abilities.[49] But it is less clear that Mikhail thinks that *only* those rules that are experienced as absolute moral rules in this fashion are rightly classified as moral rules or that anything that is experienced as this sort of non-waivable rule, regardless of its subject matter, counts as a moral rule. If this were the case, the prohibitions against incest (and perhaps what even many anti-relativists think of as unambiguously culturally contingent

---

substantive senses of the term), Hauser misleadingly thinks of it as a moral rule parallel to those more specifically acquired by the purported morality acquisition module.

[48] See, e.g. *Universal moral grammar* at 743

[49] It is by no means critical to the argument here, but I thought it worth noting a few things about the conventional/moral distinction. First, to the considerable extent that those like Mikhail who believe that the case for innate moral code-building capacity is strongly bolstered by "poverty of the stimulus" reasoning, it is dubious whether the "don't hit" v. "don't wear pajamas" example will do much work. First, and foremost, it is difficult to imagine that each rule is communicated to the youngsters who hear them with the same vocal tone or sense of dead sober seriousness by socializing parents or teachers. Second, it is not clear that even among adults, there would be anything approaching near-universal agreement about which of a set of rules were moral rather than conventional in the relevant sense: I am skeptical that those who are ardently faithful think that many rules about religious observance are subject to waiver or merely conventional, even if they recognize that not all people observe them.

prohibitions of homosexuality?) might well seem like moral, not conventional rules. Who might people think could waive them?

There are other largely "procedural" definitions of what a "moral" rule is that appear in the work of those, like Mikhail, I am describing as neo-natural law or moral realist theorists. But once more, I have very weak intuitions about anyone's level of commitment to cabining the domain of "truly" moral judgments by reference to any or all of these further procedural limitations. Still, Mikhail seems to believe quite strongly that those following moral rules are generally more capable of making judgments than explaining their judgments and that they are not really able to identify the source of the "rule" they are following;[50] that they develop (at least strong precursors) of the judgment without having been exposed to teaching of the relevant rule; and that (framed at the right level of generality), the judgments they make are universal. On balance, I think, Mikhail's definition of a moral judgment is ultimately procedural: It is a judgment acquired (like language) by a dedicated system for morality acquisition that will have the features (e.g. opacity, capacity to be learned with impoverished stimuli) that judgments grounded in other in-born competencies have.

ii. Moral judgments v. moral *behavior*

However one believes that Mikhail ultimately divides moral judgments from other sorts of judgments, it is plain, first, that he does not believe that moral *behavior* is nearly as universal as moral *judgment* and, further does not think that bottom-line moral judgments are as universally shared as the capacity to recognize moral-judgment relevant features of a situation. Mikhail is much more convinced that our initial judgments about whether behavior is obligatory, permitted or prohibited are the same (in "normal" people)

[50] See, e.g. *Law, Science, and Morality* at 1061, 1092-1093; *Cognitive Science, Ethics and Law* at 99.

than that our conduct, given this shared judgment, will be the same.[51] This is not only true in the extreme cases of sociopathy that Hauser is especially interested in. (In such cases, agents either lack the ordinary emotional responses to the judgment that one is about to engage in prohibited conduct and/or lack the empathy to compute the application of a rule forbidding harming the interests of others because they cannot perceive the interest of others.[52])

Instead, Mikhail believes, far more generally, that moral action simply need not follow moral judgment. He puts the point quite clearly: "It may be that moral *perceptions* are largely involuntary and deterministic, once the initial state of the moral faculty and subsequent experience are fixed. But it does not follow that conduct or actual physical behaviours are also deterministic. Indeed, it is an old insight that the motivation engendered by judgments of 'ought' is not an irresistible determination, but a resistible compulsion that leaves freedom intact…In addition, many factors other than moral obligation clearly play a powerful role in determining how individuals act – greed, ambition, and the pursuit of power, to name a few prominent examples."[53]

It is more ambiguous whether Mikhail thinks that bottom line moral judgments are determined, once parties exercise the more-universal capacity to represent a problem as possessing certain features. If we take the cases that have most interested him – Trolley Problems – it is possible, for instance, that the capacity to *differentiate* the classic

[51] As I note later, Mikhail can be read to be making an even less capacious claim; that "normal" people merely possess the capacity to represent events in a fashion that facilitates making moral distinctions between cases that would not readily be drawn without the capacity to make such representations.
[52] Hauser discusses variants of such sociopathic disorders often. See e,g. *Moral Minds* at 29-31,46, 232-241.
[53] Matthias Mahlmann and John Mikhail, "Cognitive Science, Ethics and Law" at 99.

Spur Track and Drawbridge problems[54] is more universal than the ultimate decision that the distinctions one might naturally draw *matter*. In my work with Tamar Kreps, we show that while subjects exposed only to the Drawbridge and Spur Track problems react quite differently to each (diverting is permissible, pushing impermissible for the overwhelming majority) these initial distinction in judgments of permissibility of diverting the Trolley on to the spur track weaken considerably when experimental subjects are simultaneously exposed to a variety of Drawbridge prompts and other prompts that emphasize problems with sacrificing a single person to help others. At the same time, we show that the commonplace intuition that it is impermissible to push the Fat Man off the Drawbridge to block the runaway train weakens considerably when subjects are exposed to cases in which they have to choose between saving more people and saving fewer when they cannot take actions that save both groups.[55] To the degree that Mikhail is making the more modest claim that people have the inborn, rather than learned, capacity to make distinct mental representations of the diverting and pushing cases, but remains agnostic as to whether those representations are made by an impenetrable moral-rule generating module, then the fact that the intuition not to divert can be readily shaken does not falsify his claims. Instead, the fact that they differentiate reactions in the single prompt setting demonstrates their capacity to do so.

[54] In the standard Spur Track Problem, subjects evaluate whether it is morally permissible to pull a switch that will divert an out-of-control trolley that will otherwise run over and kill five people on the main track on to a Spur Track where it will kill the person walking on that spur track. In the standard Drawbridge problem, it is permissible for the bystander to stop the runaway trolley by pushing a Fat Man off of the bridge on to the main track where he will block the train; the Fat Man will die but five will be saved.

[55] See Mark Kelman and Tamar Admati Kreps, "Playing with Trolleys."

I return to this issue, but I find that it is very difficult to ascertain the degree to which Mikhail is a moral-rule-modularist[56]. His claims that many laws and rules are universal as well as his claim that many moral reactions are opaque to the decision maker suggest that he is to some considerable extent a modularist about rule-generation, but one can certainly read him as merely claiming, for instance, that there is an inborn capacity to make certain representations that *could* but need not *dictate* any particular moral rules.

Mikhail does not seem to follow Hauser's (arguably imprecise) usage in describing the behavior/belief gap. Hauser almost invariably refers to those who do not act on their moral intuitions as displaying the gap we see in handling language between competence and performance, but that usage seem idiosyncratic or strained. It is quite unambiguously the case in any event that when Mikhail explores the competence/performance distinction, he is *not* merely referring to the failure to act on one's beliefs. Instead, he is exploring the degree to which people may have difficulty arriving at the moral conclusions they might otherwise arrive at because a particular problem is posed in a way that makes reaching ordinary conclusions difficult. What he wants to emphasize is that responses to moral dilemmas are frame-sensitive[57]

---

[56] That is to say that he believes that a single input or small number of inputs processed by the rule-generating module are all that can be accounted for in reaching judgments; other factors that some might think relevant simply won't penetrate the decision making mechanism.

[57] My sense is that Mikhail makes reference to the moral performance/competence distinction in a fashion that is more conventional in linguistics, but, as I have noted, I don't purport to recognize what is and isn't truly commonplace and mainstream in the linguistics discipline. Mikhail believes that there are certain judgments from among the many moral judgments people actually make that are evidential in the sense of reflecting the essential properties of an underlying cognitive system. Mikhail thinks these are the sorts of judgments that Rawls thought of as considered judgments "in which our moral capacities are most likely to be displayed without distortion." So some linguistic judgments are grammatical – these reflect competence – and others are "acceptable" – judgments about acceptability reflect the performance-based intuitions of native speakers. Moreover, there are situations in which even those with competence don't reveal it (e.g. because their attention or memory is limited); these "errors" are also errors in performance. For a good cursory discussion of Mikhail's position, see *Law, Science, and* Morality at 1094-1096. The point is discussed in far richer detail, especially in relationship to Rawls' understanding of "considered judgments", in *Elements of Cognition,* 17-19, 51-56, 342.But I am confident that Mikhail would not say that someone

But Mikhail plainly does not think that when we speak of universal morality, we need believe that everyone behaves equally morally. Acknowledging a behavior/judgment gap may render Mikhail's argument both less clear, and arguably less significant, than he believes: For those observers who believe that a genuine moral judgment *must* be a judgment that at least strongly influences conduct, that decontextualized moral puzzle-solving is not pragmatic moral judgment, claims of moral universalism simply cannot be adequately vindicated even by showing universal responses to abstract problems. Oddly, perhaps, then, while it is F&F theorists who frequently complain that H&B researchers find *in*competence by asking experimental subjects to answer decontextualized puzzles, here, it appears, the H&B theorists' argument is that we may see illusory "*competence*" by stripping moral problems of their pragmatic content. However people answer thoroughly formal, emotionally empty "trolley problems," many appear quite willing in practice to torture even innocents whose torture would move terrorists to disclose "ticking bomb" information, though such an assault is not a mere unintended side effect of action taken to generate useful information.[58]

---

who kills, despite moral judgments that killing is prohibited, is demonstrating the gap between competence and performance, and Hauser often does.

[58] More generally, as I note later in the text, one wonders whether it is possible to treat "moral judgments" as consistent across persons when moral conduct is so situationally sensitive even *intrapersonally*. It may also the be the case that in thinking about the degree to which moral judgments are *encapsulated*, it is vital to recall that the capacity to make an action-relevant bottom line judgment may be less encapsulated than the capacity to make something that looks more like a judgment of "grammaticalness."

The notion that (purportedly) "moral judgments" may be highly sensitive to the narrative form in which they are encountered is made explicit in Richard J. Gerrig, "Moral judgments in narrative context," 28 *Behavioral & Brain Sciences* 550 (2005). While Gerrig is emphasizing the degree to which Sunstein's moral heuristics may operate differently depending on the narrative setting in which a moral dilemma is encountered, the conceptual point fits more comfortably both with the tradition within H&B research in which we emphasize evaluative elicitation-sensitivity and with the further tendency within H&B thought to be wary of the temptation of domain-specific modularists and quasi-modularists to believe they can readily identify a problem's "natural domain." Events may not come pre-packaged in this sense as "double effect" problems but rather as problems involving certain forms of heroes or villains facing much more

iii. Specifically moral capacities v. capacities useful in reaching moral judgments

Finally, I take it that Mikhail's view of the morality-acquiring cognitive structure is somewhat less capacious than Hauser's in the following sense. Hauser is prone to emphasize all the features of cognition that he believes are necessary to implement any rules that he defines as moral rules, while Mikhail seems more prone to emphasize our need to understand the more precise algorithms that govern the narrower morality-acquisition module. Thus, for instance, Hauser is more prone than Mikhail would be to emphasize that it would be impossible to make judgments about others' intent without, say, having a developed capacity to distinguish inanimate from animate objects or to develop a more general theory of mind, while at the same time recognizing that the development of the capacity to make the animate/inanimate distinction or to develop a theory of mind has multiple uses beyond inferring intent when making moralistic judgments.[59]

Mikhail appears to be considerably more interested in the competencies that are unique to morality acquisition, and constrain the sorts of moralities we are likely to be able to learn, rather than those that may be more generally useful in implementing a wider range of moral schemes than he believes we actually observe. Now, of course, it is

---

situationally-precise action-dilemmas.

[59] Similarly, Hauser believes that people could not make judgments about whether they had been treated equitably – and resistance to inequitable treatment is one of many critical moral universals for Hauser -- without having a certain degree of natural numeracy that permits people to compare how well distinct subjects have been treated in distinct situations. (I should note that I find his precise arguments about numeracy extraordinarily under-developed): Still, he does not imply that numerical competence is merely one feature of a "bundled" equity-perception module rather than a more all-purpose cognitive trait that can be utilized in making equity judgments. In this sense, Hauser's argument is reminiscent of Gigerenzer's metaphor of the adaptive toolbox. Looked at this way, "numeracy" is just one of a set of basic cognitive capacities that can be appropriated for multiple uses.

possible that Mikhail could be persuaded that the content of the UMG is itself responsive to limitations dictated by some secondary set of human competencies.[60] But Mikhail tends to be preoccupied with the more direct moral computational grammar. Thus, for instance, he believes that people inevitably compute moral events with regard to intent, but he seems considerably less preoccupied with the underlying cognitive correlates that permit the apprehension of intent to be computable.

c. The content of Mikhail's Universal Moral Grammar

i. Thin features of the UMG and the debate over non-tautological universals

What then *is* the content of Mikhail's UMG? Well, at this point, I would say it still seems rather meagerly specified, and I surmise Mikhail would agree that cognitive scientists are still in the very early stages of discovering the range of UMG governing principles. I don't think it would be terribly unfair to say that Mikhail's UMG has, at this point in time, several fairly thin features, and one thick one. Here is my best understanding of the "thin" (or not obviously enormously constraining) ones: First, all moral systems are built on the idea of ascribing responsibility to agents for causing results, and they must distinguish between results caused intentionally, knowingly and accidentally. I am more than a bit unclear on whether he thinks that all moral systems must further distinguish between involuntary "action" and voluntary action that the party is not subjectively aware will cause harm.[61] I am also not entirely clear whether he

[60] Here is a purely hypothetical case I have in mind to help illustrate my point: I surmise, if Mikhail could be persuaded that our "numerical" computational capacities dictated that we could make only certain sorts of judgments about equity of treatment or proportionality in punishment, he would likely believe that people would have developed only those modes of computing "moral" events that allowed us to "plug in" the sorts of "numbers" we could conceptualize.

[61] I am also unclear on whether he thinks the UMG dictates an answer to the following sort of typical "hard" legal problem: does a person act voluntarily at point 2 in time if he takes some voluntary action at point 1 that creates a risk (or certainty) that he will act involuntarily later? While falling to the earth, a skydiver *might* be said not to be acting/to be acting involuntarily, but we might say that his fall is voluntary

believes that all moral systems must distinguish, or that people must have the computational capacity to distinguish, within the domain of unforeseen and unintended harms, negligent from non-negligent causation of harm, i.e. have some category of culpable carelessness, indifference, and/or inattention.[62] Second, all moral systems require maintaining idea that there are important distinctions between moral and conventional rules. Third, all moral systems require that all action could be classified as obligatory, permissible, or forbidden. Fourth, there are some basic content-based prohibitions (against murder, rape and other similar types of aggression) that are not only universal as content-rules but inevitably take the same structural form: murder, for instance, is intentionally causing death without justification; rape is forced sex etc.)

Obviously, it is possible to be skeptical about the significance of the claim that these are meaningful universals, or that observing these sorts of universals would lead us to believe it more likely that there is some morality-acquisition module. The question is whether these "grammatical features" constrain the development of moral systems or merely provide analytical categories that we can use to describe, essentially

---

because it is a certain consequence of the decision to skydive. Is a person driving (a voluntary act) when he runs over pedestrians while in an unconscious state from an epileptic seizure if he drove aware of the risk that would occur? For a discussion of such time framing issues, see Mark Kelman, "Interpretive Construction in the Substantive Criminal Law," 33 *Stan. L. Rev.* 591, 601-603 (1981). One of the many reasons I am so skeptical of universalist accounts is that a classroom of students scarcely ever agrees on when it is appropriate to characterize "conduct" that follows such risk-of-later-involuntary-movement as voluntary or not. If the UMG simply suggests that we treat involuntary conduct as less problematic than voluntary conduct, it seems nearly empty to me unless it helps us establish which conduct is voluntary. I return to this point in the text, particularly in the discussion of Mikhail's discussions of the distinction between permissible, mandatory and forbidden conduct and his discussions of the universality of norms forbidding homicide and rape.

[62] Once more, it is difficult to take claims of universality on such issues seriously: Many Anglo-American commentators believed that criminal liability/moral blame for negligence was wholly inapt while others believed it perfectly sensible. For an excellent discussion of many of the repeated themes in this long-standing debate, see Kenneth Simons, "When is Negligent Inadvertence Culpable?" 5 *Crim. Law & Phil.* 97 (2011).

retrospectively, the structure of any rules.[63] Naturally, there are those who will question the claims that there are non-empty moral universals. Such skeptics will think instead that the "rules" Mikhail articulates are essentially observer-imposed analytical categories than can be applied, more-or-less tautologically, to any disparate set of rules, rather than internally-generated limits on the structure of moral cognition.[64] In this sense, they will further argue that the key terms are not computable observables but the culturally contingent *real* rules that govern morality.

In this view, it is empty to say that everyone believes and can easily learn that killing without justification is morally prohibited murder. In this view, any set of rules regulating killing could be logically parsed as-if it had that structure in the sense that one would merely define the sort of justification needed to help define a wrong as the sort of thing that negated wrongfulness.[65] Instead, those suspicious of Mikhail's claim will argue that all of the action comes at the level of distinguishing what is and is not an operationally adequate justification (finding cheating spouses, infidel detection etc.). Similarly, it may simply be a matter of definition that everyone believes that "forced" sex is morally prohibited rape -- we probably could not use the word/concept "forced" unless it was defined by easy contrast with something more acceptable. If there is no agreement, across or within cultures, on the sorts of constraint that are compatible with an adequately

[63] Skeptics will likely level the same accusation about the "observation" that moral systems all distinguish the prohibited, permitted, and obligatory. Once more, they are likely to argue that this is an analytical truth, not a synthetic one.

[64] The claim that all conduct must be (in the final analysis, in terms of a final authoritative decision) obligatory, forbidden or permitted is almost surely a tautology. Perhaps Mikhail means to say that all behavior is so classified without any borderline cases or sense that there are cases that pose dilemmas. Such a claim would not be tautologically true, but then again it would not obviously be descriptively plausible either.

[65] This is essentially the argument that Posner made that Mikhail attacks. See, e.g. Richard Posner, "The Problematics of Moral and Legal Theory," 111 *Harv. L. Rev.* 1637, 1640-41 (1998).

free choice to assent to sexual contact, one wonders what it means to say that "rape" is universally prohibited.[66]

Finally, it is plain that we can (quasi-tautologically) describe any moral judgment in terms of a judgment about an agent's morally causal responsibility for a (culturally relative) bad consequence. But for us post-Coaseans, the claim that the mind has a very limited range of algorithms to process, represent and compute *factual* unidirectional causal relations remains what could generously be described as a puzzling one

This point requires a fairly substantial digression. For most academic lawyers, the digression will be needlessly long (because the point has become reasonably commonplace), but it is enormously significant, in large part because both Hauser and Mikhail assert that a substantial aspect of the UMG is that it involves the purported capacity to identify when one party has caused another harm. The demise of the idea within the legal academy that people can make such stable unidirectional causal inferences can fairly be traced back to the analysis of legal relations provided by the Legal Realist, Wesley Hohfeld, in the early part of the twentieth century, and restated in somewhat different terms by Ronald Coase in 1960. Briefly, the argument is this. All disputes between two or more parties basically involve conflicting desires about how to deploy scarce resources. The principle that we may do as we wish, provided we do not cause harm to others, is useless in deciding which set of desires should prevail, because, however we decide the dispute, one actor will be "harmed" as a matter of fact, in the sense that he or she will be told he or she cannot do as he or she wishes with impunity. Thus, we cannot resolve the dispute as a matter of "fact" as to who has harmed whom; we

[66] For one (of many) detailed discussions of the indeterminacy of judgments about when women's choice to engage in sexual contacts is adequately unconstrained, see Mark Kelman, "Thinking About Sexual Consent," 58 *Stan. L. Rev.* 935, 964-971 (2005).

can only decide it based on some normative commitments (implicit or explicit) that lead us to favor one side's interests over the other side's interests.

For lawyers, that way of looking at the problem of harm-to-others traces back at least to Hohfeld's early twentieth century reconceptualization of legal rights. Looked at functionally, Hohfeld argued, rights are constituted by the reciprocal duties that others bear to the rights holder. Thus, any expansion of one owner's protected sphere of unfettered action (the would-be user's use privileges) *must* come at the cost of contracting the rights of those adversely affected by her growingly protected uses to be immune from the adverse effects of such uses. To the extent that we privilege the owner of parcel #1 to barbeque freely, the owner of parcel #2 loses immunity from the losses that come from dealing with smoke. To the extent that we allow the owner of parcel #1 to build a tall building, the owner of parcel #2 stands to lose her view. Thus, however we resolve the dispute between the two owners, we must either interfere with the would-be user's "autonomy" (in the sense that we will limit her free action) or the "autonomy" of the party seeking to be protected from loss (in the sense that we will compromise our ordinary commitment to protecting people from non-consensual shifts in their baseline position.)

Fifty years after Hohfeld, the economist Ronald Coase made essentially the same point, phrased in economic rather than legal terms, in his famous article, "The Problem of Social Cost," arguing that, as a descriptive matter, all social costs represent the joint costs of conflicting desires in a world of scarcity. The but-for cause of harm is inevitably the actions of two interacting parties: a railroad emitting sparks does not cause a fire unless

the farmer chooses to place his crops next to the train line. Causation in this sense appears mutual, not unidirectional.

There is a more sophisticated reformulation of Coase's basic point, which better reveals the inadequacy of the unidirectional harm principle, uninformed by covert perfectionism or welfarism, to resolve disputes between conflicting uses. The *cost* we are concerned with in the typical case of interactive harm, need not, as a matter of actual social practice or normative theory, take the form of the physical damage that would occur if each party acted without either legal restraint or knowledge of the other's activity. Look again at Coase's famous example of interactive harm: A railroad wishes to run its trains on tracks that lie alongside a farm; if it does so, though, the sparks emitted by the train will set fire to the farmer's crops. It is true that if the railroad proceeds unimpeded, the consequence will be physical damage to the crops in the form of a fire, and we might intuit (in Mikhial or Hauser's sense?) that the railroad "caused" the fire. But "damage" to the parties' interests need not take that form. The relevant damage – the true subject of moral judgment -- that concerns Coase is what he labels the *social cost*, by which he means the difference in the sum of the values of the two parcels or activities in a world in which they did not interact and their value in the world in which they do interact. The social cost *might* take the form of a crop fire, but it might also take the form of added spark suppression costs for railroads, if they are forced to take preventive measures to prevent the fire in the first place, or lost profits for the farmer, if the farmer is forced to take preventive measures (for example, by ceasing to plant on land adjacent to the tracks). If farms were nowhere near trains (the non-interaction situation), the parties would not have to bear the cost of their interaction, in any form (fires, reduced

crops or spark suppression costs). But it is equally the case that if the owners were near one another, but had different desires – if the farmer had no desire to use his land for flammable crops or the train company's proprietors for spark-emitting transportation -- there would also be no social cost. The social cost is thus the product of clashing desires, given interaction. Naturally, not all uses of one's self or one's property compromise the interests of others, whether we measure the adverse effect by diminished property values, or by decreased subjective utility. But the only time courts are asked to adjudicate or we are asked to make moral judgments whether or not the would-be user is privileged to act is when *someone* is adversely affected by the proposed use. To put it another way, the problem of law (like the subject matter of economics, morality, or distributive ethics) depends on the existence of conflict (or scarcity). Unless we decide (as perfectionists), that one desire is more legitimate, more expressive of preferred human values, or decide (as welfarists) that one desire is *stronger*, we appear stuck.

In legal circles, a few people have tried to salvage a determinative role for an unmoralized notion of unidirectional harm-in-fact by limiting (legal or moral) liability to a subset of acts that cause harm-in-fact, with the subset defined by the objective (unmoralized, natural) properties of those acts. (In essence, both Hauser and Mikhail believe the UMG dictates that we compute harm-causation in this way.) Two notable efforts in this vein have been Richard Epstein's argument that only physically invasive causes are culpable,[67] and Hart and Honore's argument that only unusual causes are culpable.[68] .

[67] Richard Epstein, "A Theory of Strict Liability," 2 *J. Legal Stud.* 151 (1973) Richard Epstein, "Defenses and Subsequent Pleas in a System of Strict Liability," 3 *J. Legal Stud*. 165 (1974).
[68] H.L.A. Hart and Tony Honore, *Causation and the Law* 26-94, 109-129 (2d. ed. 1985).

Epstein starts with a universalistic (strict liability) view of culpability for harm-in-fact: that you, as a matter of fact, harmed me is, prima facie, a sufficient reason to hold you liable for that harm. But he immediately limits the reach of this broad principle by invoking a limiting (and quirky) definition of cause: only those supplying, threatening or forcing others to employ invasive force are deemed to cause an event. Epstein's position is neither plausibly complete, descriptively, nor morally compelling. Its incompleteness is clear. If the Defendant leaves his unlit car on a busy highway, it is the Plaintiff's car that forcefully invades when it smashes into Defendant's car. It is unlikely that Epstein really wants to suggest that the Plaintiff caused the crash, though. Similarly, Epstein acknowledges that a hypersensitive plaintiff is causally responsible for the atypical damages he suffers from defendant's routine invasive activities, even though the plaintiff plainly does not forcefully invade himself. The more troublesome problem with Epstein's scheme, however, is the normative one: why should we care about the physical properties of a causal factor at all, in determining culpability? It is easy to concoct examples that make that fixation seem absurd: One defendant erects a mirror which deflects sunrays that thereby "invade" a neighbor's property, causing it to be too bright, while the other builds a high fence, blocking light and thereby causing it to be too dark. Epstein would have us conclude that because only the first defendant's actions are invasive, only the first defendant is liable, although clearly the plaintiff is equally harmed, in similar ways, by either action. But even in the run-of-the-mill case (my car hits yours, stopped at a stoplight), where intuitively one feels that liability should attach to the cause that is physically invasive, that intuition cannot really ultimately rest on physical causation itself; it must come from some external value(s) that turn out to be

served (on some occasions) by distinguishing among causes based on whether they are physically invasive or not. It is hard to imagine any theory of rights—other than the purely tautological claim that we have a right to be free of physical invasion—that would do the trick.

In another attempt to isolate some subset of causes-in-fact as morally culpable, Hart and Honore argue that liability attaches to those causes that cannot be taken for granted—that are unusual in the normal course of events. If Epstein's account forces distinctions that seem singularly unappealing, this account is unhelpfully circular. The question of which actor is simply "going about his business" and which one is interfering with the ordinary flow of events cannot *determine* the legal cause of the injury (or, in Mikhail and Hauser's sense be computed as a basic cognitive primitive) as it will be significantly determined by decisions about legal and moral cause. If the railroad is entitled to emit sparks, the farmer's decision to snuggle his crops right up to the tracks will disrupt the ordinary course of events and appear to be the fire's cause (just as a decision to park on the highway would doubtless routinely be viewed as disruptive by American jurors). If the farmer is entitled to use all of his property for crops, and others must take steps to insure those crops remain undamaged, the spark-emitting train will look disruptive (and therefore cause the conflagration). We cannot resolve these issues without recourse to extolling existing social practice /custom; perfectionist dialogue about morally preferred uses; or welfarist dialogue about subjective gains inherent in each use. Moreover, just as with Epstein's physically invasive causes, the normative appeal of this distinction is unclear: why do "ordinary" acts have a higher moral standing than more unusual ones? And as with Epstein's physical invasion test, whether an act is

usual or unusual may be a tolerably good proxy for something else we might care about, on welfarist or other grounds- e.g., whether those harmed by the act should have anticipated it and thus protected themselves; the likelihood the act is socially useful. But it is hard to see why we would care about it in itself.

b. The thicker feature of Mikhail's UMG: normative and positive implications of experimental findings about Trolley Problems

At the same time as we might ultimately question whether there is much bite to any of Mikhail's "thin" rules (about the importance of the capacity to focus on intent, the need to separate the moral from the conventional, and the ability to identify the party who causes harm), it is important to recognize that the vast bulk of Mikhail's experimental and detailed theoretical work has been focused on making one "thicker" (i.e. more obviously content-rule constraining) claim about the contours of the UMG: The most critical thicker claim is that the principle of "double effect" is a by-product of the action representational features of the UMG. Thus, without regard to culture, learning, or particular predispositions, all "normal" subjects will distinguish cases in which an agent acts impermissibly because that agent commits one or more distinct intentional batteries prior to and as a means of achieving his good end from those in which an agent's conduct is permissible because the violations are (merely?) subsequent foreseen side effects of an action taken for clearly beneficial purposes. Mikhail clearly believes that the distinct cases trigger distinct mental representations and seems to believe, albeit far less clearly, that once represented as they naturally are represented, the evaluation of the overall action follows inexorably.[69]

[69] For a description of the relevant representations, see *Moral heuristics or moral competence* at 557-558 or *Universal moral grammar* at 146.

It is not precisely clear what if anything Mikhail thinks follows normatively were we to accept the idea that a fuller UMG might someday be specified and understood. One wonders, after all, whether one could ever derive any normative principles from factual observations. I think, at this point, any accounts are largely speculative, but the speculations seem to me worth making.

First, it is possible that what Mikhail might ultimately argue that the question of whether a moral principle or practice generated by the UMG is by virtue of that a "good" or "acceptable" argument is in some ways senseless, in much the same way as it would seem rather senseless to ask whether the linguistic structures we can generate given the linguistic competence represented by the universal grammar are "good" linguistic structures. In a sense, if we believe that the moral domain is *defined* as the product of the morality-creating module, the products of the UMG *are* human morality.

There is a second argument that I tentatively feel comes closer to Mikhail's view. If it were the case that only certain moral rules could be readily learned and readily attract agreement, it would be of some weight in evaluating those rules. The degree of weight is quite ambiguous: it is possible that the fact that a rule is acquired effortlessly and attracts consensus is, in Mikhail's view, merely a mild factor in its favor or a near-killer argument for its acceptability. And it is also the case that Mikhail seems more committed to the idea that certain representations useful for reaching a certain judgment are in-born than that the judgments are inevitable, given the representational capacity. Teasing out the precise basis for this sort of claim is not easy though. Certainly, there are familiar functionalist arguments that all things equal, a set of *legal*, governing rules that citizens readily learn and commands ready assent from most subjects has advantages over rules

that are hard to teach and hard to legitimate. It is not clear, though, either whether these arguments apply (and if so how?) to moral rules, or whether these functional advantages, largely in terms of compliance levels and/or enforcement costs, outweigh whatever disadvantages one might say arises if the readily-learned rules are "bad" in some other sense.

Third, as I mentioned (and criticized) earlier, it is possible to link an evolved capacity to learn a certain sort of rule to the claim that the rule is adaptive and to link the claim that the rule is adaptive to the claim that it is, by virtue of its adaptive power, superior.

Fourth and finally, it is possible to make a somewhat weak argument that widely shared intuitions are presumptively valid (because people are "smart" and discern the "truth" in some sense) or, less affirmatively, that there are simply no sources of judgments about morality other than intuitions. But even if this were the case, it is not clear whether we are best off relying on *preliminary* intuitions rather than intuitions that are tested and pressed in ways that Mikhail explicitly disclaims. What he thinks the UMG most clearly dictates is unself-conscious, hard-to-explicate reactions rather than judgments that can be defended, generalized, and articulated.

B. Sunstein's moral heuristics

1. Heuristics as "rules of thumb" of uncertain origin

Sunstein basically views heuristics as short cuts or rules of thumb. These short cuts produce the same bottom line substantive answer that a fuller exploration of the whole range of potentially relevant information would produce in the typical case, but it

is cognitively simpler to use the short cut.[70] In each sort of case in which a subject makes use of a heuristic, the agent has some limited set of goals – to assess the probability of an outcome, to judge the permissibility of certain conduct – and the rule of thumb will be "accurate enough" to assess probability or to make the appropriate judgment about permissibility most of the time. Thus, the agent substitutes the "heuristic attribute" for the true "target attribute" in making a judgment. What can often be problematic about the use of heuristics is that these heuristics are, like any rule, inapt to the full range of situations in which they may apply: Just as it is *usually* the case that one will judge probability accurately if one follows the "rule" that events that are readily available to memory have occurred more frequently than unavailable events, but sometimes one will not (because one occasionally recalls events because they are salient rather than frequently encountered), so will it be the case, for instance, that omissions are usually less culpable than commissions (because, for instance, those who omit to take steps are less likely to intend harm and intending harm is relevant; because one cannot discern whether or not omissions are deliberate or not and blame without "proof" is a poor idea) but sometimes they are not.

Sunstein is not especially clear whether he thinks of the moral heuristics as consciously adopted rules designed to meet known ends;[71] whether he believes instead that they are essentially the sort of cognitive routines that we develop because in trying consciously to meet a particular end over a range of situations, we sub-consciously

[70] Like H&B theorists generally, Sunstein emphasizes the internal limits of the decision maker, not the possibility that those using a rule of thumb that fits the actual environment would reach a solution superior to those who considered and tried to weigh more aspects of the problem.

[71] When he refers to law students who use the "Scalia heuristic" as a moral heuristic -- believing that anything Justice Scalia believes is either automatically right or automatically wrong – he is plainly referring to a consciously adopted rule of thumb. See *Moral heuristics* at 533.

develop a habit of substituting one or a small number of attributes that we see often in analyzing a situation for a fuller analysis of the situation; or whether individuals are predisposed to process the simpler heuristic cues, completely unaware that they might be predisposed to do so because, somewhere in our evolutionary history, processing these simple cues was sufficient to meet our ends in most "similar" situations.[72] He is simply not clear whether people consciously know the "target attribute" at all, and if they do, whether they are consciously aware that they are substituting a "heuristic attribute" for the target attribute because they know it is easier to do so.

Thus, for instance, it is very difficult to determine whether Sunstein believes that those who distinguish omissions from commissions ever think that, say, discerning the intention of the party whose course of conduct they are evaluating is their true purpose. It is just as hard to tell if he believes they did so before developing a conscious rule of thumb that they would merely ascertain whether they were dealing with an act or failure to act or whether the "rule of thumb" simply developed as an increasingly automatic reaction because in a large number of situations in which they were attempting to discern intention, they unconsciously noted that they so rarely found it in omissions cases that the simpler search for an act became a habit. Because he plainly states that people use heuristics when employing the more automatic, cognitively intuitive System One,[73] it might seem that he is more drawn to a different story: people have no conscious idea of why they distinguish omissions from commissions. Making this distinction is not a *method* of meeting a purpose (e.g. to ferret out intentional actors) that they have consciously developed nor is doing so the residue of having gone through many iterations

[72] The idea that the use of the heuristics was implanted, through selection pressure. in unaware brains somewhere deep in our evolutionary history appears, e.g.. in *Moral heuristics* at 534.
[73] See *Moral heuristics* at 533.

of trying to ascertain intent and settling on an economical way of getting there. Rather, it is a "rule" whose connection to its original purpose is unrecoverable. But I don't think his work is ultimately all that clear on that point, and I suspect that the debate over whether and how to minimize the use of "moral heuristics" generally depends in some substantial part on sorting this sticky issue out.

This primary ambiguity in identifying the origins and basic nature of the moral heuristics is at least partly responsible for a secondary ambiguity in his argument: To the extent that the heuristics were either consciously adopted as short-cuts to meet known ends or developed as habitual solutions to a problem whose significant meaning-parameters were unambiguous, it would be easy to make sense of the claim that the heuristic *misfired* in certain settings. If the *goal* of a rule of thumb is to get its user to make some judgment J when and only when the situation warrants that judgment, it is inapt when it fails to do so. But to the degree that these moral heuristics are not precisely short-cuts to reaching an articulated goal, but rather modes of cognition that simply meet some end that Sunstein (or any other observer) ascribes to them, Sunstein will always face the criticism that those using the heuristics *are* meeting some actual free-standing end that Sunstein simply fails to see, rather than failing to meet the end that he has identified.[74]

An analogy might help. Think about an employer who statistically discriminates, e.g. by using cues that mix information about the race and class of job applicants to make judgments about workplace dependability and discipline. One can imagine an employer

[74] This criticism is raised in Barbara Fried, "Moral heuristics and the means/end distinction," 28 *Behavioral and Brain Sciences* 549 (2005). I return in the text to the ways in which Fried suggests this problem might be answered. It is also raised in Ilana Ritov, "Cognitive heuristics and deontological rules," 28 *Brain and Behavioral Sciences* 559 (2005) (noting that subjects who appear to fail to meet the weak consequentialist ends Sunstein ascribes to them might instead be meeting deontological goals.)

who says, "My conscious goal is to find a worker with qualities Q(a)…Q(z)" and I believe that if I look at the address the person grew up at, his race, and the high school he attended, I can assess the presence or absence of these hard-to-observe qualities." Such proxies (consciously adopted rules of thumb) can be unquestionably inaccurate if someone whose proxy traits lead the decision maker to infer the applicant has quality Q(a) when she actually has quality not-Q(a). One can imagine, too, that statistical discrimination, though not a consciously adopted decision rubric, arises through unconscious *learning*: In this view, the statistically discriminating employer still has a clear aim: he knows he is looking for certain employee qualities. But he does not know that he has come to use race, address, and high school as proxies, though he has done so because he has so frequently come to associate facts about these variables with his ultimate judgments. Once more, the use of the unknown rule of thumb could still be revealed to be troublesome and inaccurate, given that the goal he was seeking when he developed habitual rules of thumb was to assess worker quality and the rule of thumb may at least sometimes fail to do so. If, however, one imagines instead that negative judgments based on race and class are simply automatic (more-or-less modularized cognitions) made by those of another race and/or class, it is no more than *possible* to say that these judgments *must* have developed because they were proxies for worker quality judgments and could therefore be unambiguously mistaken when they lead the agent to believe that a qualified person is unqualified. It is also possible to say that they meet some entirely separate goal (for instance, increasing in-group solidarity or the self-esteem of the evaluators.)[75]

[75] If making the omission/commission distinction is automatic, inaccessible to judgment and justification, and was formed prior to and without regard to any individual decision maker's goals or experience, it is

Ambiguous conceptions of the nature of heuristics arguably creates a third ambiguity in the general argument as well: To the degree that Sunstein is simply noting that rules can be inapt, one might argue that his argument is directed just as much at any reflective Kantian who believes in the possibility of general principles as it is directed at those using heuristics. My point is not to enter the debate about whether there are anything that can be described as non-empty rules or principles that can be stated in such a way that they are not over and under-inclusive,[76] nor to deal with the significance of Sunstein's recognition that it might be *systemically* better if people stuck to using inapt heuristics (or principles) rather than to try to solve each case on its own merits, given the usual host of problems with case-by-case decision making.[77] It is simply to note that it is sometimes difficult to tell whether Sunstein is trying to identify a particular set of

---

always a bit of a 'just-so' guessing game to ascertain what purpose is served by making it. But if we don't know why it develops, it is difficult to know what it means to say that it can get applied when it is "inapt." Inapt to meet what end? We know that in most situations in which it can be said that experimental subjects find an omission less culpable (or more permitted) than an act which (at least) arguably may have the "same consequences," the cases can be differentiated in a host of ways. There may be actual distinctions in intention, presumed distinctions in intention, ease of ascertaining intention,[75] certainty that the agent is central to the consequence arising, distinct levels of ambiguity (both factual and conceptual) over the causal role the agent played, distinct degree of direct involvement in the harm, distinctions in the degree to which one course of behavior maintains the "status quo" and another alters it. There may also be a distinction in the degree to which one course of action is more ordinary/normal/expected, or differences in the relative ease of focusing our attention on the consequences of our course of behavior. For a brief discussion of how difficult it is to say what distinctions those who draw an act/omission distinction might be drawing, see Christopher J. Anderson, "Alternative perspectives on omission bias," 28 *Behavioral & Brain Sciences* 544 (2005).

It is difficult to say that those who make the distinction are serving none of *these* possible ends in drawing the contrasts they draw, assuming it might be as reasonable to distinguish punishment or permissibility based on these distinctions as it is on intention, let alone some other ends that come less readily to our minds as observers. Obviously, H&B theorists frequently had to struggle with the question of defining error; Sunstein's struggles merely recapitulate the broader problems. It is considerably easier to find error when subjects reach logically incoherent or plainly factually incorrect judgments; a bit easier when they (at least) readily disclaim their preliminary judgments. But where, as in many of these cases, none of those easy critiques is available, judgments that heuristics are troublesome are themselves troublesome.

[76] Sunstein begins *Moral Heuristics and Moral Framing* by describing heuristics almost wholly in terms of the overgeneralization of intuitions that work well in ordinary life. See id. at 1557.

[77] See, e.g. *Moral heuristics* at 534-35 (defending the use of heuristics on just such rule-utilitarian grounds), 541-542 (noting that it might be better if people adhered rigidly to rules that generated bad judgments in rare and exotic cases rather than be tempted to depart from the rule mistakenly believing they are justified in departing.)

troublesome heuristics (defined above all, from my viewpoint, in terms of unself-conscious attribute substitution by those who are neither aware of the "target attributes" they actually seek to ascertain nor of the relationship between the "substitute attributes" and any plausible set of "target attributes") or to restate, in one particular context, a more general critique of the use of rules or proxies.[78]

Ambiguity about the source of moral heuristics creates yet a fourth, arguably less significant, ambiguity as well: To the extent that the heuristics are consciously chosen simplification strategies, one would not expect them to be universal. To the extent that they are unconsciously chosen cognitive strategies, they might well be just as universal (as first-line intuitions, if not as ultimate "solutions" to moral problems) as Mikhail's UMG-based algorithms. To the degree that they arise from incorporating unconsciously learned patterns, they might or might not be close to universal. But Sunstein never really tells us whether or not he thinks the moral heuristics are universal in part because he does not sort through carefully how we have come to use them.

3. The content of Sunstein's moral heuristics and his theory of "bias" or error

Ultimately, though, what Sunstein seems most certain of is that, as a result of some unspecified processes, certain reactions to "moral problems" are typical System One fast-and-simple, automatic judgments. It is even less clear analyzing Sunstein's work than working through Mikhail's what sorts of issues should be said to raise the relevant

[78] This problem is noted in Karen Bartsch and Jennifer Cole Wright, "Towards an intuitionist account of moral development," 28 *Behavioral and Brain Sciences* 546 (2005.) It is explored further in Urike Hahn, John-Mark Frost, and Greg Maio, "What's in a heuristic?" 28 *Behavioral and Brain Sciences* 551 (2005) (noting that all legal norms are defeasible, so that stated at some level of generality, all norms, heuristic or not, are over-general in the first instance). They argue that Sunstein ought to foreswear the idea that the moral heuristics are best understood as particular, inevitably over-general *content rules* (e.g. omissions are not so bad, betrayals are especially bad, things that are natural are good) in favor of a view that heuristic reasoning refers to a particular *process* of reasoning. I agree with their claim that Sunstein has not adequately identified what process those reasoning heuristically on moral issues are using and return to this issue in the text.

sort of "moral" concerns. I am quite confident that he does not limit the domain of moral judgments to judgments on a certain class of substantive topics (e.g. harming and helping). I am (marginally) confident that he does not draw Mikhail's strong procedural distinction between, say, moral and conventional judgments (or judgments that are universal and those that are local, or judgments that can be understood and defended and those that seem to persist even when they cannot be readily rationalized, or judgments that we can make even if we have not been "taught" to make them.) I think, though again I am not confident in this view, that he simply treats any judgment that action is worthy of condemnation or praise or any judgment that an action should be thought of as properly permitted, mandated or punished as a moral judgment. (He is just as ambiguous as Mikhail about whether abstract judgments count as moral whether or not they are connected to moral action decision making, but, more than Mikhail, he tends to analyze actions rather than stated reactions.) By and large, the moral heuristics that he looks at seem fairly content-specific.[79] Plainly, though, we can always imagine moral heuristics that worked at an even finer grain in terms of content (e.g. treat those who have compromised ones' safety when they have made *explicit* protection contracts as betraying in an especially bad way) or a broader grain (treat intentional harm as worse than harm that occurs as a result of recklessness or carelessness, assuming the harm the actor causes through his conduct is otherwise the same.[80])

[79]For instance, I would describe the following heuristics he discusses as "content-specific" rules rather than as "modes of construal:" The fact that we employ a betrayal heuristic means that we will judge betrayals – harm-causing by those who have specifically promised to protect the ultimately injured party from harm – more harshly than action taken by others that causes equivalent harms. The use of a heuristic that one should not knowingly cause a human death lead us to treat those who directly calculate the costs and benefits of action that harms others as more morally reprehensible than the actions of those who never make such explicit calculations, though they impose equal or greater risks).

[80] That this is from some viewpoints merely a "generally accurate" rule of thumb could be seen if we consider the possibility of arguing that provoked killers (or those said to be acting under the influence of

What I therefore must return to in discussing the relationship between the Sunstein/Mikhail debate and the more general H&B/F&F debate is that Sunstein does not seem to be discussing mental *capacities* so much as content-specific *rules*, while Mikhail is essentially doing just the opposite. Of course, when Mikhail posits the existence of certain capacities, he does so believing that the presence of these capacities may well either typically generate or perhaps even guarantee the presence of a certain delimited set of content-specific rules. And, on the flip side, one can describe a person using one of Sunstein's content-specific rules as merely demonstrating the capacity to process information in accord with the rule, but I think we will see this distinction in approach has genuine bite.

Sunstein ultimately both catalogues a set of moral heuristics and attempts to establish a general method to assess the argument that the heuristics lead to something that could best be thought of as error. Not surprisingly, he recognizes that the claim that subjects are making errors in these cases will be more contested than parallel claims that Kahneman and Tversky made in discussing the basic *cognitive* heuristics: The H&B school's experimental subjects made judgments about facts that were sometimes simply logically impossible (e.g. there are more earthquakes in California than natural disasters West of the Rockies; more words ending in 'ing' than '-n-') and sometimes merely wrong

---

extreme emotional disturbance for which there is a reasonable explanation or excuse) are still *intentional* (if partly excused) killers while those who do not desire to kill, but calculate that it is too costly to take steps that would reduce the risks that their conduct will indeed kill when we believe the risks they have taken are substantial and unjustified are morally more problematic, even though *most* intentional killers are "worse" than most "reckless" killers. Similarly, one can readily construct arguments that rapists who are merely negligent as to consent – systematically unaware of whether women are consenting or not because women's sexual agency is of so little moment to them that they are utterly inattentive to its expression – are at least as morally problematic as those who, from sadism or explicit swelf-conscious misogyny, harm women certain that they are doing so.

(e.g. there are more words beginning with ‘r’ than words whose third letter is r; more deaths from airplane crashes than household falls.)

Sunstein’s catalogue consists of four categories of moral heuristics: First, there are heuristics that he describes broadly as involving risk regulation. The first two particular instances he has in mind, though – a punitive reaction to those who engage in *explicit* cost-benefit calculation when deciding to take actions that will impose risks on others and a resistance to establishing markets in emissions that permit people to pay for the right to emit pollutants – might be catalogued in a somewhat distinct functional fashion from the one Sunstein uses. They seem to me, at core, to involve what he sees as confusions between our generally valid moral reactions to situations in which the optimal harm level is zero[81]and situations in which the optimal harm level is plainly positive. It is valid in cases in which the accepted optimal harm level is zero to condemn efforts to balance gains to “perpetrators” against losses to victims, but if one extends this anti-balancing “intuition” or heuristic to situations in which risk is inevitable and/or desirable, one will make bad policy. There is a distinct class of heuristics involving risks grounded in the “betrayal” heuristic that I have already mentioned: Instead of evaluating the overall risk of a particular outcome arising from the use of a particular product, agents will overweight the bad outcomes that come from the harms directly caused by a safety device – even though that safety device prevents a good deal of harm from secondary causes – because getting injured by a good that “promises” to protect you is seen as a betrayal of trust.[82]

[81] Or, to put the point more modestly, these are cases in which justification defenses to prima facie wrongs are exceptional rather than routine.

[82] Think in this regard about people failing to take vaccines that prevent disease because the vaccine itself has dangerous side effects, even when the disease reduction outweighs the side effect risk in terms of

Second, there are a series of what Sunstein sees as "biased judgments" associated with the use of the "outrage heuristic" in punishment.[83] What these cases have in common is that those using the heuristics seem insensitive to the consequences of punishment generally or the particular form or level of punishment they are considering imposing.[84] If adequately outraged by behavior that has endangered consumers, for instance, they seem unconcerned whether or not punishment of the misbehaving entity will meet the goal of increasing the long-run safety of available products. Similarly, they are prone to demand that a company expend its funds to clean up the toxins it has illicitly generated or disposed of rather than to use the funds to clean up other toxic sites at which the expenditure of funds would generate greater health benefits. Finally, in this regard, subjects are insensitive to the probability of detection in assessing punitive damages.

Third, Sunstein believes that there is a "moral heuristic" against "tampering with nature" or "playing God" that leads people to over-value outcomes they see as more natural. They will misperceive, for instance, the gains and losses associated with "natural" and "artificial" additives though of course all additives are simply organic compounds with whatever set of good or bad effects such compounds might produce when ingested. Similarly, they will over-demonize novel technologies that seem to substitute for existing natural processes (hence "irrational" resistance to cloning, stem cell research, even IVF) and misestimate the relative risks associated with "natural" and "man-made" events.

---

expected mortality and morbidity, or failing to install air bags that prevent far more deaths in accidents than they cause because the fact that the air bags themselves sometimes cause death is viewed as a "betrayal."

[83] Sunstein explores this class of moral heuristics further in Cass R. Sunstein, "On the Psychology of Punishment," 11 *Supreme Court Econ. Rev.* 171 (2004).

[84] Naturally, this class of cases raises most cleanly the possibility that the subjects have retributive goals distinct from the weak consequentialist ones that Sunstein attributes to them.

Fourth, and finally, Sunstein believes that both the distinctions made between acts and omissions, and what he sees as modestly related distinctions made by those seeking to follow the "double effect" principle are at core heuristics that poorly meet our considered ends in minimizing bad outcomes and condemning those worthy of condemnation in particular cases.

How does Sunstein establish that those making moral judgments consistent with these heuristics are making *mistakes*? The self-conscious answer he explicates most clearly in his work is at core substantive, grounded in a particular theory of the nature of rational thought. In this view, the subjects are making mistakes if their conclusions are inconsistent with what he calls "weak consequentialism," defined as a framework that takes account of consequences, including the violation of imperfectly constraining deontological principles when evaluating action. He acknowledges (too weakly, I am sure, to meet fully the objections of readers committed to many forms of deontological reasoning) that to the degree that a party seems irrational only because his response pattern seems oblivious to consequences (recall the punishment examples), this will not seem like an error to some deontologists.[85]

Sunstein's responses are also, at times, seemingly "procedural" but even in the situations in which this seems to be the case, he may be unduly suppressing substantive controversy over the ends he has unself-consciously ascribed to the "mistaken" agents. Thus, at times, it appears that he believes that two judgments are inconsistent given what he sees as the metric the agents must, transparently, intend to apply: Take the betrayal heuristic. If one thinks that subjects *must* be trying to compare the wisdom of safety devices by looking at bottom line aggregate risks, then those using the heuristic are

[85] *Moral heuristics* at 534

reaching judgments that are not consistent. In cases in which no betrayal effects are present, they prefer to accept a 1% risk of death rather than a 5% risk while they prefer the opposite when the betrayal heuristic is activated. But such decisions could seem "procedurally" suspect in at least two distinct ways: They might be *unstable* (subjects would renounce the decision if its features were pointed out to them) or they might simply be *inconsistent* in respect to the metric the researcher believes must be in play.

The argument that choices that do not survive reflection are irrational is one with a substantial pedigree in the H&B literature generally. It is worth recalling then, the debate over the claim in the traditional judgment and decision making literature, and merely noting that the arguments about the persuasiveness of the claim are likely to be raised here. F & F researchers are likely to argue that it may well be true, but trivial, that subjects will regret or disown judgments that are reinterpreted for them in formal and abstract terms that make the judgments transparently flaky. Recall the argument in the context of F&F critiques of typical H&B findings: Once one explains the Linda problem in terms of logical conjunctions, those who have said she is more likely to be a feminist bank teller than a bank teller can see that they are wrong. But the judgments may have been correct to meet the organism's real pragmatic ends. Believing Linda is a feminist bank teller meets our pragmatic need to treat conversational cues as relevant and demonstrates our capacity to read sub-text as well as text into statements we hear. In the context of the betrayal heuristic, rejecting safety products with bad side effects *could* be a fast and frugal strategy, grounded in a "betrayal detection" device that could well be a close kin of the "cheater detection device" in cementing social exchange, that leaves parties safer than they would be if they tried to make multi-cue based decisions about

aggregate risk, perhaps by creating incentives for "protectors" to do better. Still, of course, the H&B counterarguments to these sorts of F&F objections are powerful as well: Subjects vulnerable to perceptual illusions who renounce their views of the relative size of two circles once they measure each circle seem to have been mistaken not only because we treat the size of a circle as a brute external fact, but because we trust the judgment they make after what they view as appropriate reflection more than we trust the one they make without such reflection.

Arguments from "inconsistency" seem to take two forms. In the less controversial form, a response is inconsistent when the same outcome is evaluated differently merely because it is described in a different fashion. Again, as I noted earlier, this sort of frame/elicitation sensitivity is frequently highlighted by H&B researches arguing that heuristics lead to "mistaken" judgments of expected value not only because subjects make factual errors in judging probabilities but because their evaluation of outcomes is frame sensitive, violating principles against accounting for the presence or absence of irrelevant alternatives or violating principles that the value of an end-state does not depend on how that end-state is named or described. And Sunstein at times simply imports, wholesale, from the conventional H&B literature examples in which judgments that he calls moral are frame sensitive: He reports for instance that judgments about the propriety of adopting a vaccination program are sensitive to whether subjects are told about how many lives will be saved or about how many will die, even when the bottom

line in terms of mortality is identical[86]; obviously, this merely restates the classic H&B gain/loss aversion asymmetry experiments.[87]

Preferences may be inconsistent in a second, more capacious, and arguably more controversial sense as well, though. They may be described as inconsistent simply because they cannot be justified by a reflective principle that allows the decision maker to explain the dimension or dimensions along which cases judged distinct were really distinct, or articulate a principle that could be applied across cases.

Consider, in this regard, the standard responses to standard stand-alone Trolley Problems. I think, for Sunstein, that a subject is inconsistent in his responses in this sense if the only decision principle he can articulate is that he should maximize the number of lives saved (in an act-utilitarian sense? given rule-utilitarian qualification?) but then makes distinct judgments in situations in which the number of lives lost is identical.

---

[86] See, e.g. *Moral heuristics* at 535,

[87] At the same time, he notes that the answers subjects give to significant (moral?) questions about the degree to which we should trade off future deaths for current deaths are irrationally sensitive to elicitation method. Thus, one set of subjects is asked questions in a way that suggests that they would choose a program that saves 100 lives now rather than one that saves as many as 7000 lives in 100 years. (As one might imagine, one can argue that subjects given the problem in this form imagine – rightly or wrongly – that other technological changes will occur in the next century that will save the 7000 without the program; this point is merely an extension of the critique of the original loss/gain asymmetry studies that I noted were grounded in the claim that programs that were described as saving 200 of 600 people might *actually* be better than those described as resulting in the deaths of 400 of 600 because more than 200 might live, as a result of supplementary programs, if a program that saves 200 is adopted.) Surprisingly, perhaps, subjects were typically indifferent between programs saving 55 lives now and 105 lives in 20 years and those that saved 100 now and 50 in 20 years, suggesting that the trade-off is 45saved current lives for 55 saved 100 years hence, not 100 for 7000). More surprisingly still, they preferred programs that showed a steady increase in life-saving efficacy over time to those that seemed to reveal a gradual worsening of our capacity to control the environment or a breach in their view that human history should be progressive (people prefer a program that saves 100 lives this decade, 200 the next, and 300 the next to one that saves 300, then 200, then 100 lives.)

What is clear in all these cases is that the preferences are simply inconsistent if in all cases we have done nothing other than alter the way in which the same outcome is described or elicit responses by adding or omitting irrelevant information (e.g. that one program will save an invariant positive number of people now and in the future should not change judgments about trade-offs, compared to situations in which only the trade-off is presented alone). Similarly, if all we have done is highlight a feature of the decision-making environment that may well have been present when it was not made salient – in setting one (the 100 for 7000) trade-off, we do not highlight the ongoing technical regress, while we do in the last experimental setting), it is troublesome that our evaluations should shift so radically.

Though his discussion of the Trolley Problem strongly suggests that the target attribute that he believes that subjects are (should be?) trying to identify is the attribute that would be accepted by a pure act-utilitarian,[88] I take it as well that he might also describe the subjects as inconsistent if their true, "target" judgments were grounded in a particular form of moralistic retributivism that they failed to apply consistently across cases. Thus, imagine that Sunstein believes that the subjects were committed to distinguishing those more culpable "killers" who actively *desired* that the victim die (even if they desired it as a means to some further end) from those who merely *accepted* the death of an innocent, and even took steps to minimize the likelihood of that death. Subjects would be inconsistent in this view if they blamed some, but not all, who merely accepted death.[89]

Assume I am right that Sunstein is not clearly correct to attribute the goals he attributes to subjects in these settings so that it is simply unreasonable to complain that they are being inconsistent if they don't meet the attributed ends. What if his point, instead, were merely that they could not articulate *any* other principle (or worse still, accept any one they might be offered) that would render their judgments consistent-in-

---

[88] See *Moral heuristics* at 540-541

[89] It is a subject for a far richer debate than I need to detail for now whether the desire/acceptance distinction is truly stable as a moral distinction or whether it is rather always nothing more than a factually contingent one. Thus, the standard case in which the murderer ostensibly desires death though he has a secondary and purportedly irrelevant motive – he kills his uncle to get the inheritance – might better be described not as a case in which he desires his uncle's death, at the moral level, but one in which he recognizes at the contingent factual level that though he would like to get the money while Uncle lived out his natural life, that is just certainly not going to happen.
Mikhail obviously believes that these sorts of temporal ordering sequences (contingent though they may be) are critical UMG building blocks. Thus, it is clear that in each of the following two cases, the "defendant" merely accepts the victim's death but only in the first case is it clear, in temporal ordering terms, that his death *necessarily precedes* the "good" (desired) result. Case One: D1 diverts a trolley so that it hits a large object that will slow the trolley down giving those on the track time to escape. The large object is a person. D2 diverts the trolley so that it hits a large object that will slow the trolley down; the large object is an inanimate weight, and it is the weight that will slow the train. But there is a victim standing next to the weight who will be killed if D2 diverts the trolley in this way.

relationship-to-that-principle.[90] Is *that* a critique of moral heuristic-based thinking? Perhaps not. Perhaps a focus on what appears to be purely procedure-focused consistency surreptitiously imports an undefended substantive bias towards non-deontological schemas in which consequences are judged in relationship to relatively readily commensurable consequence-describing metrics (utils, dollars, lost lives, whatever.)[91] It is certainly far *easier* to make more transparently consistent judgments if they merely must be consistent in the sense that the agent accords equal treatment to all situations in which he discerns that readily observed, readily measured outcomes are the same.

C. Further reflections on the Sunstein/Mikhail debate informed by the broader debate over heuristics

My goal is not so much to resolve the debate between Sunstein and Mikhail as to press in a particular way on the claims that each is making.[92] My real hope in this section though is to demonstrate that we can illuminate this debate a good deal by seeing it in significant part as just one instantiation of the broader heuristics debate I tried to set out earlier. To put this point a bit more narrowly, I believe that it will help to see that some of the critiques of H&B theory I articulated briefly are just the sorts of critiques one should be especially sensitive to in looking at Sunstein's work. Similarly, the sorts of critiques of

[90] I don't think it is true in the Trolley cases that the cases *cannot* be distinguished in terms of a general principle or trait: While the *death* of the victims in *each* case is a known side effect rather than intended, the *battery* in the push case is *intended* while it is not in the divert case. The point in the text though is that it might not matter even if no such principle could be adduced.

[91] This point is emphasized in Barbara Fried, "Moral heuristics and the means/end distinction" at 549-550.

[92] I hope that my efforts to summarize their work expressed a particular sort of criticism that I won't dwell on further in this section: Each of them seems considerably less clear in articulating the precise nature of his claims than I think would be ideal. And I suspect some of my sensitivity to what I perceive as the ways in which each theory was inadequately specified comes from focusing on the heuristics debate: For instance, my sensitivity to Sunstein's failure to distinguish individually developed rules of thumb from general features of domain-specific cognition is grounded to a considerable extent on recognizing how that issue plays out in thinking about the nature of heuristics.

F&F and MM theory that I raised are among the sorts of critiques that ought to make us most wary of Mikhail's claims.

1. Interrogating Sunstein

While I think the most commonplace and most telling criticism of H&B theory generally is that H&B theorists at least arguably identify judgment processes and problematic performances that are unlikely to occur in naturalistic settings, I do not believe that those who would criticize Sunstein's work on moral heuristics from an F&F vantage point would argue that he has identified judgment patterns that are unduly lab-specific or unduly sensitive to the elicitation procedures used in the laboratory setting. In fact, as I mentioned, it may be the case that it is Sunstein who would argue that Mikhail's universal moral competence is unduly restricted to an odd, and uninteresting, set of laboratory settings that may not demonstrate true pragmatic moral competence but merely a form of abstract problem-solving ability.

Instead, I think, critics attempting to extend the general critique of the H&B literature to Sunstein's work would focus on what I described to be the second sorts of critique: First, I strongly suspect that they would argue that the heuristics he identifies are under-specified and inadequately tethered to identifiable human capacities. Because of this, it is difficult to identify in any particular case when or how the heuristic will operate. Perhaps worse still, it is difficult to ascertain what positive role the use of the heuristic might serve, except by reflecting on the general advantages of rule-utilitarian judgment metrics, advantages that have nothing to do with identifying any particular short-cut, proxy-based cognitive mechanism. And yet there is no clear argument that one could

really see the judgment as resulting solely from what could best be seen as limits in our cognitive capacities.

Second, I think they would argue that he fails to explore the possibility that the heuristics produce "better-than-rational" results, given the information available in the decision-making environment, not just most of the time, as a rule of thumb might, but all of the time (because multiple cues generate noisy, non-recurring patterns; because multiple cues generate intractable problems or generate judgment outcome sets with incommensurable competing concerns.) In this sense, the problem is that he contemplates only two of the three possible ways of looking at the heuristics: Sunstein certainly contemplates and emphasizes the view that they generate mistakes, and we should try to correct these mistakes.[93] He further contemplates the view that while they generate mistakes in individual cases, we might make more mistakes overall if we did not use them all of the time and instead tried to pick out situations in which it would be helpful to drop them.[94] But what he does not contemplate is the possibility that the heuristics do not simply generate fewer errors, used systematically, but that there is at least a sub-set of cases in which they systematically outperform non-heuristic reasoning in each case in which they are used.

Recall the criticism that H&B theorists generally neither adequately specify the cognitive processes that "biased" subjects purportedly use nor do they attempt to lodge

[93] We might correct them at the individual level, by developing better System Two oversight techniques to check System One reactions. We might correct them at the institutional level, by shifting the locus of decision making from those more likely to act on the basis of System One intuitions to a set of decision makers less likely to be in a position to react quickly and automatically.

[94] That is to say, the error *rates* created by foreswearing rules of thumb are higher than the rates we see if we use them, whether this is a result of untoward, biased motivation when we depart from universal rules or because we are too cognitively limited to make use of information outside-the-heuristic box.

the heuristic in a well-defined cognitive capacity.[95] I strongly suspect most F&F theorists would find Sunstein's heuristics equally under-specified and unduly detached from identified cognitive capacities. Take, for example, Sunstein's (extraordinarily interesting) "betrayal heuristic." I think it is actually quite hard to determine the situations in which he believes it should operate because it is unclear what "betrayal" really is in his view. Does the heuristic operate only when safety devices harm or kill? Would it extend to finding annoying aspects of vacations much more unpleasant than similar annoyances in daily life because vacations "promise" pleasure? Does it matter if one is taking a vacation package arranged by a purveying, quasi-intentional entity like a travel agency or does one treat "the vacation" as a pseudo-animate source of "betrayal?" How does the mind distinguish betrayals from situations in which the putatively "betraying" party has promised a mix of favorable and unfavorable outcomes that the promised party deems beneficial on the whole and then has delivered on that linked set of promises: Is there (merely) some class of cases (and how would that class be identified?) in which the mind (irrationally?) refuses to comprehend the existence of complex, fulfilled promises with negative and positive features?

At the same time, one reason it might be hard to figure out what the betrayal heuristic entails is that Sunstein makes no real effort to figure out how, given a plausible account of the set of cognitive capacities we might have that would be implicated in "betrayal situations," we might develop one, but not all, versions of a "betrayal

[95]The first illustration I offered was that F&F theorists complain that "availability" was neither adequately defined – in the sense that it was not clear what it would mean to say that a class of events was more available than another class -- nor carefully described as an aspect of memory retrieval. The second was that one could not determine when parties would be subject to the gambler's fallacy – negative recency – rather than the hot-hand fallacy – positive recency – because neither was defined or lodged in what the F&F theorists saw as the relevant capacities to make judgments about animate, intentional actors and inanimate, unintentional action.

heuristic." He does note, at a fairly general level, that it makes sense that people would feel especially aggrieved by breaches of trust. He points out in that regard that when trust is breached, those who are betrayed lose not only what they would lose to anyone who injured them but lose their faith that they can rely on the sort of trust-based relationships that are central to social cooperation. But the picture of trust and social cooperation is not even sketchily developed, nor does he argue that we have developed either an unreflective System One "emotion" (betrayal aversion) or an automatic "cognition" (atypical capability to identify the factual risks imposed by those one trusts to protect you) to facilitate the maintenance of trust.[96]

Don't get me wrong. While I find these typical F&F hesitations about Sunstein's heuristics (like the "betrayal heuristic") quite compelling, it is by no means the case that I find that current F&F efforts to overcome these problems are persuasive. The truth is, we may simply be in a position where we don't yet or won't ever identify the precise nature of and scope of the cognitive short-cuts we use or understand how they build upon a well-specified set of capacities. It is plausible to me, for instance, that the experiments and surveys demonstrating that people would rather accept a higher overall risk of death from a car accident in a car missing air bags than a lower one from a malfunctioning or otherwise-fatal airbag reflects a (more basic? more cognitively explicable?) omissions bias rather than a betrayal heuristic. And it is just as plausible to me that F&F (or MM) theorists will someday come to believe that sensitivity to betrayal arises from the same sort of evolutionary pressure as the purported Cheater Detection Module (that I

[96] Even in terms of the standard H&B "attribution substitution" view of heuristics, the betrayal heuristic seems poorly specified. I am not utterly confident on this point, but I don't think that Sunstein is actually arguing that subjects' target attribute is "aggregate risk reduction" and that they mistakenly believe that if they reduce betrayal based risks that they will actually serve the end of reducing aggregate risk. The theory does not appear cognitive in that way, but, as I said, I am just not sure.

discussed), and has the same sort of purported adaptive impact. Just as we solve seemingly cognitively identical rule-violation identifying tasks more readily when they involve situations in which rule violation could be described as cheating, so might we solve risk-assessment tasks more readily when they involve "betrayal detection." (And that they will argue that betrayal detection and aversion each serve the same broad sort of adaptive purpose as does cheater detection in making social cooperation possible.) But I would almost surely have doubts about whether betrayal aversion or detection is decently understood, or represented at the apt level of generality, once we tied it into an adaptive capacity, just as I remain skeptical not only that there is something like a cheater detection module that solves problems drawing on a few non-dedicated general cognitive mechanisms but that we could possibly identify precisely what aspects of the "cheating detection" problem are the ones that characterize it as a salient, differentiable sort of problem.

More generally, think about another problem I adverted to in trying to describe what Sunstein means when he speaks of moral heuristics: Sunstein is quite casual about drawing the possible distinctions among conscious rules of thumb, unconsciously developed judgment-pattern recognition in situations in which goals remain conscious, and general human-capacity based cognitive mechanisms. This failure may well be grounded in the more general failure to specify carefully both the nature of a heuristic and the particular capacity and/or capacity limitations it draws on. If we identified a set of moral heuristics that were typically developed as conscious rules of thumb by individuals, it might make sense to think about correctives at the individual level, or if they were *learned* by individuals, we might think of "educational" reform, broadly

construed.[97] If we identified a set of moral heuristics whose developed use was invisible, there might be distinct ways of bringing the existence of the heuristic to consciousness.

If, though, the heuristic is neither taught nor developed, but lodged in deeper, unconscious cognitive structures, it is first more plausible that it would be useful to figure out in order to help us make a normative judgment about the heuristic's likely utility, if and why there is a gap between the environment in which its use evolved and current environments. We must figure out as well whether the environment really provides us, on any occasion, with cues that would permit resolutions of problems that seemed superior. Even if we decided that the heuristic was dysfunctional in these sorts of ways, we would almost certainly be more skeptical of the possibility of individual-by-individual reform, at least in the absence of a more developed theory than Sunstein even hints at of how System Two "oversight" thought can be activated when it does not spontaneously work. The failure to work through the underlying cognitive mechanisms leaves us with relatively feeble and under-theorized reactions to the relative recalcitrance of distinct heuristics: we know that most of the time, most of them are relatively immune to education, incentives, and efforts to force focusing, but there is little that H&B theorists offer us to tell us why that is not always the case.[98]

Naturally, the same difficulties that Sunstein faces in attempting to convince his readers that the moral heuristics lead to *bad* outcomes will complicate efforts he (or the

---

[97] As a descriptive matter, that there would be no reason to think that the use of this class of heuristics is anything close to universal, while the use of heuristics in the third class likely is. If, to draw on a prior example, white males engage in certain forms of statistical discrimination against women or Blacks because they are taught to or consciously develop it as a strategy, it is more plausible that it might be overcome by certain forms of reflection and simple rational argument than it would be if it were an unconscious universal response to some biological need to show certain form of in-group favoritism.

[98] Because, for instance, we don't have a clear capacity-based picture of what hindsight bias *is*, it is difficult to say why it seems to be diminished by forcing people to construct counterfactuals but not by incentives.

F&F researchers more typically absorbed by this project) might make to interrogate the possibility that they instead lead to better-than-rational results. Gigerenzer typically uses very general techniques to identify the sorts of problems that are solved poorly by those attempting to be "fully rational" – observing, for instance, that the subject faces the sort of moral problem in which he would be prone to over-fit regression equations to non-recurring data, identifying that he is facing the sort of moral problem in which he is likely to need to sum incommensurable outcome variables. Concluding that any effort to use these sorts of techniques would prove helpful seems to me, at this point, as much a matter of taste and faith as anything else. But it is still worth noting two important points: First, we should acknowledge the fact that Sunstein has paid little attention to the possibility that he has identified super-rational heuristics. Second, though, as I discussed earlier, Mikhail's claims that the "heuristics" might be the inevitable product of a morality-acquiring module (and may give rise to universally held judgments) does not really tell us whether the outputs of that module are superior, along any imaginable dimension, to the products of some other "reflective" or "classically rational" process that is considerably harder to learn or generates some set of reactions we instinctively find far more jarring.

2. Interrogating Mikhail

Once more, it is important to recall why H&B theorists were so wary of the F&F school's accounts: First, they typically flipped the accusation that H&B researchers under-specified both the nature of, and mechanisms behind the heuristics they identified. Instead, they argued, F&F researchers purport to describe basic features of cognition, but do so not by examining cognition carefully but by assuming that certain features of thought *must* exist because it would make some sort of theoretical sense that they

should.[99] In the typical case, H&B critics suspect that the basic cognitive features are distorted to fit some just-so adaptationist story. I don't think that those who worry that Mikhail has fit his UMG to an adaptationist story so much as he has tailored his story of a "moral module" to resemble the language acquisition capacities broadly posited by Chomsky that are certainly, if not uncontroversial, more accepted than any accounts of moral competence. But one ultimately sees the same sorts of worries: is Mikhail distorting the definition of moral competence and moral judgment to make it look more like linguistic competence than it really does? Distorting data to make moral judgments seem "universal" in the same way that grammatical judgments are? Making unsupported claims that "moralities" have the same finite number of significant parameters as grammars purportedly do?

Second, the H&B theorists are invariably highly suspicious of claims that significant cognitive processes – including moral judgment making – are highly encapsulated.[100] It turns out that the question of whether Mikhail thinks of moral judgments as encapsulated depends on resolving definitional questions that I noted are quite thorny: To the degree that we believe (as a matter of definition?) that a moral judgment is not a true "judgment" unless it is instantiated in moral behavior, or at least in some sort of reasonably potent urge to engage in moral behavior or to feel some sort of disquiet if one does not, then there may be lots of evidence (much of which I suspect

[99] Recall one aspect of the discussion of the recognition heuristic: the critique of the F&F work was that that the F&F researchers had not accurately portrayed the sort of memory that their experimental subjects were actually demonstrating, but merely imagined that they were manifesting the sort of capacity that, first, they thought would be adaptive to have developed, and second, would permit subjects to make use of the sort of fast and frugal heuristic they imagined people use.

[100] Thus, recall from the critique of the existence of the recognition heuristic the claim that subjects appeared to account for compensatory information, secondary cues beyond the single cue (is one but not both of two cities in a pair "recognized") that Gigerenzer and Goldstein assumed (wrongly in my view) was used lexically in making judgments about the relative size of two cities.

Mikhail would accept) that moral judgments are not especially encapsulated. Similarly, if we believe that moral judgments are only those judgments that are "considered" in certain fashions (adopted as categorical imperatives? embraced as fitting some consciously desired life plan? stable when considered alongside multiple moral problems), then the cognitive processes that permit the development of those sorts of judgments may not be (even in Mikhail's view) especially encapsulated. But what is less clear is whether Mikhail thinks that even initial (unreflective, unacted-upon) bottom-line judgments are (strongly, modularly) immune from reflection or even that they are (weakly, with stopping-rule like features) prone to be made on the basis of just one or a few features of the problem.

What might be helpful is to think about both these issues in relationship to the judgments on Trolley Problems that have most preoccupied Mikhail, particularly those Trolley Problems that do not involve "personal violence" – throwing someone from a drawbridge to block the trolley v. diverting the trolley[101] – but those that merely alter

[101] It is worth noting that in the standard, personal violence, throw the victim from the drawbridge version of the Trolley Problem, experimental subjects almost surely resist to some extent the precise instructions they are given. As a result, their condemnation of the actor is even more over-determined than Mikhail believes (it comes not just from using the person as a means to an end, and not just from resistance to "personal violence.) Once more, I understand that it is usually F&F scholars who criticize H&B researchers for posing problems that savvy subjects refuse to answer on the experimenters' terms, but the critique may well be one that the H&B researchers would throw back at Mikhail here. Though subjects are explicitly *told* that the person who is to be thrown at the trolley not only can stop the runaway train but is *uniquely* big enough to do so, that instruction is factually inane: As a result, it is almost surely the case that one reason experimental subjects condemn the person P who throws X at the trolley to save five lives is that P could have, just as easily (and more heroically), jumped in front of the trolley himself if he wanted to save multiple lives or. Even more plausibly, most respondents almost surely believe that the Fat Man will not block the runaway train so that pushing him in front of it results in six, rather than five deaths, not one rather than five. Diverting the trolley sounds like something that could efficacious in saving lives in real life; blocking runaway trains with a single human body does not. More generally, this raises the problem of whether we can readily comprehend precisely how subjects will represent scenarios with innumerable potentially relevant features: MM and F&F researchers tend to force scenarios into pre-packaged domain-specific boxes.

whether the victim is killed because he is standing next to the heavy object that stops the runaway trolley or whether he *is* that heavy object.

The first thing to note is that the analogy to linguistic competence seems to falter badly on the numbers: It is difficult to understand exactly why Mikhail reads his own data as supporting his claims that moral judgments are strongly determined by a morality-acquiring module. While it is indeed the case that a statistically significantly higher proportion of respondents do indeed believe it improper to "use the man as a means to stop the train" rather than "know the man will die because he is standing next to the heavy blocking object," the truth is that the results reveal nothing like the sort of consensus that we see in using basic grammatical rules. A quite substantial 48% of respondents think it is permissible to use the man to stop the train (vs. merely 62% who think it okay to kill him as a side-effect.) It strikes me that the "linguistic analogy" is being strained past the breaking point if it relies on judgments that are this weakly shared (imagine 48% of native English speakers deciding to reverse noun/verb order). Mikhail is not so much observing a capacity as imagining one he either thinks "fits" human needs or closely resembles the best-understood knowledge-acquisition "module."

More striking perhaps, claims that the Spur Track and Drawbridge problems generate truly distinct (or opposite) responses turn out to be extraordinarily difficult to sustain when looked at in detail: As I noted in my work with Kreps, Drawbridge, in the first instance, tends to elicit a far more mandatory side-constraining judgment that pushing is impermissible, even when, for example parties to be saved by pushing are related to the putative pusher or are more thickly identified. Spur Track tends to generate judgments that diverting is merely permissible – not mandatory, which is more clearly the

conceptual opposite of impermissible – and even judgments that it is permissible are held far more weakly than judgments that pushing is forbidden. Permissibility judgments are also altered significantly by identifying the putative victim on the Spur Track or tweaking the facts by stating that he is related to the person contemplating diverting. Furthermore, it proves to be the case that initial responses are unstable in the presence of prompts that tend to push against either the intuition that killing one to save others may be impermissible or that push against the tendency to ignore the aggregate numbers of lives that will be lost when contemplating a particular action.[102] These shifts do not seem to me plausibly described as competence errors, but rather as reflections of the fact that whatever representational capacity people might have to differentiate Spur Track and Drawbridge cases play a modest role when they need to make bottom-line moral judgments in a variety of contexts.

Worse still, the "granularity" problem[103] and encapsulation problems that beset all modular and "softly modular" theories are enormously bothersome here. Mikhail is confident that he is observing nothing but "double effect" reactions, but I am puzzled by why he thinks this to be the case (this is the granularity problem) or whether he thinks the

[102] When experimental subjects are simultaneously exposed to Drawbridge and Spur Track prompts at once, along with prompts that highlight problems with sacrificing one person's interests so long as it will benefit a greater number of people than are harmed, responses to the basic Spur Track prompt become much more like responses to the Drawbridge prompts while Drawbridge responses do not change at all. If subjects are asked to respond to moral dilemmas as they think a morally admirable person would and simultaneously see Drawbridge and Spur Track prompts alongside prompts in which they are prone to try to minimize lives lost – prompts in which they must allocate resources to save more or fewer people from harm -- then opposition to pushing in the Drawbridge case drops substantially.

[103] All theorists committed to domain-specificity run into the problem that a domain can be specified at broader or narrower levels of generality: is a dedicated cheater detection mechanism best described as a sub-set of a mechanism devoted to reasoning about deontic conditionals or as a mechanism devoted solely to social cooperation-protecting cheater detection? Do those who do standard domain-specific evolutionary psychological work on female sexual desire think that the "capacity" to pick out and be attracted only to mates who will care for the kiddies is its own domain, a sub-set of a far larger one (sex without material support is just a form of cheater detection?) or too large a domain (there are actually different attraction rules for the range of distinct situations in which sexual choices might be made)? For a fuller discussion, see *The Heuristics Debate* 61-2, 76-9.

"double effect" reactions are just one input into fuller "moral judgments." (This is the encapsulation problem and it occurs in part because we have no firm idea what a bottom line moral judgment really is.) Jack Bauer – the hero of the once-popular TV show *24* – violates Mikhail's "universal injunctions" not to commit batteries merely because doing so has subsequent good impacts as often as most of us change socks. (For instance, he elicits confessions by starting to torture a suspect's innocent sister in front of him.) But he is just a hero willing to make the hard choices to virtually all of his audience.[104] Do the *24* viewers represent the choice in a fashion distinct from the fashion that they should if simply manifesting Mikhail's UMG (i.e. are there other features of the situation that are represented that he is just missing?) Do they not "stop" in making a judgment once they have computed a single cue ("I've got a double effects problem here") even if the cue is significant (i.e. is the judgment non-lexical?)[105] Or, as social psychologists have long

[104] I do not mean to claim that the viewers would all be comfortable with such torture in real life: their "support" for Bauer may well be sensitive to the fantasy context. But then again, Mikhail's experiments simply create a distinct fantasy context.

[105] The encapsulation problem is impossible to divorce from the problem of whether one can derive a (certain form of) "ought" from a certain form of "is" as well. Assume for argument's sake that that one was convinced that certain moral rules are radically more easily learned, acquired with very few stimuli by people in early stages of development. As I have noted, Mikhail doesn't give us much guidance about the question of whether "easily acquired" moral rules are superior (along any dimension) to rules that are difficult to acquire.

Assume that we start by thinking about perceptual *illusions* (e.g. a circle C of equal size to a circle C' will look smaller against a backdrop of large circles): Let's further say that even those generally skeptical of Massive Modularity as an across-the-board cognitive theory think the perceptual input systems might well be modularized. It might be true that it increased reproductive fitness to make this particular judgment and to have the *sight-based judgment* – if not the ultimate judgment on size – be fully recalcitrant to separate knowledge that one's judgment was wrong – i.e. even after you *know* the circle is smaller or bigger than you thought, you can't *see* it that way. It might also be the case that the perceptual illusion was better described as an evolutionary by-product (i.e. given the way our perceptual system works – largely developed for other reasons – it is likely that we'd make this mistake though making the mistake is not itself adaptive). Still, people *can* plainly measure the two circles and avoid making bets on relative circle size based on their perceptual systems. There are plainly perceptual *illusions*, and it is not obvious that anything that Mikhail tells us about competence necessarily contradicts Sunstein's notion that initial moral judgments are also often (in some less clearly specified sense) illusory as well.

So one question is whether Mikhail thinks that the "moral intuitions" are *merely* like initial perceptions, but not stable sources of ultimate judgment. It is simply not clear whether he thinks judgment involves the combination of many inputs, or whether he resists or accepts the basic H&B view that we typically do not use the full range of inputs, that we might be predisposed to use few cues – e.g. only the visual cues in

suggested, are all *real* pragmatic judgments heavily *situation-determined*? If, as is the case, seminary students are less likely to help a needy homeless man if in a hurry to deliver a sermon on the Good Samaritan or if subjects are far more likely to behave altruistically when they've just gotten a dime back from a pay phone, what might it mean to think that there are any interestingly universal judgments about things like the apt level of morally compulsory altruism?

IV. Conclusion

It is not at all clear that reactions to moral quandaries are truly widely, much less universally, shared, without regard to cultural and ideological distinctions. And even if reactions are widely shared, it is not clear how to *interpret* whatever universality we observe. One might argue, as Mikhail does, that we can figure out a mind-dependent universal moral code by analyzing the representations of moral quandaries that people naturally generate, unself-consciously, without learning how to make them. But one might also argue, as Sunstein does, that all we can conclude when we observe universal reactions is that all of us people are cognitively limited creatures who make use of simplifying strategies to deal with hard issues that work out pretty well, generally, but which misfire in significant numbers of cases. We need something besides our intuitive reactions to tell us whether, and if so when, our intuitions are serving us poorly.

---

making circle size judgments – because these are usually sufficient and we've got limited time and energy, in the H&B internal limits sense). It is possible instead that the thinks (like the F&F people) that the use of the single quasi-perceptual cue (I've encoded this problem as an omissions problem or as raising double effects issues) will do the equivalent of outperforming the use of a ruler (though what will it mean to "outperform" in the moral domain is not clear). Finally, it is conceivable that, like the MM people, that there are no ruler-equivalents that can change a moral judgment, just as there might conceivably be a world in which there were no rulers that can penetrate one's size judgments, which are made, once and for all, by the Circle Size Judgment Module

The truth is, though, that we can interpret the way people make factual judgments and reach decisions with little or no apparent moral content in much the same way. It may well be the case that there are certain heuristic "techniques" that all people use to reach judgments and make decisions, making use of just a sub-set of data that might seem germane and just a sub-set of analytical techniques that we seem able to use on some occasions. Once again, there is controversy over whether there are any such universally used cognitive short-cuts, and, once more, there is controversy over how one ought to interpret the use of heuristics. Are they, at core, generally useful effort-reducing short-cuts, bound on some occasions to lead a decision maker astray, whose efficacy in particular cases must be assessed by non-heuristic cognitive assessment techniques? Or are they, at core, the best source of our practical intelligence, evolved adaptive mechanisms that generate better judgments and decisions that serve our ends better than do techniques more consonant with high effort, "rational" choice? In the judgment and decision making literature, each of these distinct interpretations is associated with a school of thought. Heuristics and biases (H&B) scholars emphasize that people substitute an easily processed judgment for a more complex one, and that this serves their ends well-enough most of the time, especially given limitations in time, attention and computational power, but that there are predictable problems that flow from this. Fast and frugal (F&F) theorists emphasize both that most judgments and decisions are made on the basis of a single or small number of cues – that the mind has few general computing capacities that permit people to take account of further facts or considerations and balance them against the single decision-cuing trait of a situation – and that the judgments and decisions that are made by such lexical decision makers are optimal, not

just given internal processing limits, but optimal because they have typically evolved to solve the finite number of differentiated recurring problems the organism faces.

What I have emphasized in this piece is that the two literatures overlap considerably more than people writing in either tradition have implied, so that the long-standing debates between H&B and F&F scholars are likely to illuminate the debates between those like Mikhail who seek to extol a natural law lodged in shared intuitions to represent moral problems in a particular way and those like Sunstein who note the existence of a rather disparate set of biases.

Many of the things that are attractive in Mikhail's work echo what is attractive in the work of F&F scholars, but, more importantly for my purposes in this paper, much of what seems puzzling or unconvincing is his writing is what is puzzling and unconvincing in F&F work more generally. F&F scholars purport to observe carefully and precisely the features of cognition – and sharply criticize H&B authors for giving accounts of particular features of cognition that are both under-specified and inadequately theorized (in the sense that there is no decent evolutionary account of why the feature might have developed). But in many cases, they seem not so much to describe a cognitive trait as to describe a trait that somehow *must* exist, given their guesses about what adaptive pressures would have created. Similarly, Mikhail does not seem so much to describe the Universal Moral Grammar or cognitive processes that actually exist as he describes a set of moral representational capacities that would exist if there were some moral capacity that worked in much the same way that the capacity to learn language worked. Because of this, perhaps, he actually has a remarkably thin picture of what judgments are best described as moral and how morally relevant representations do or don't determine

bottom-line moral judgments (let alone determine emotions that might be associated with the judgments or actions that might or might not be triggered by judgments and emotions associated with judgments). Worse still, perhaps, he shares the F&F school's unwarranted belief that, as a descriptive matter, people invariably make lexical judgments or that, as a normative matter, they should: Just as there is little or no credible evidence that people make factual and evaluative judgments disregarding factors other than factors that might at first seem salient, or might dominate decisions in ordinary cases, the evidence that moral judgments are anything but immune to a range of competing considerations is powerful and the claim that single-factor judgments would normatively dominate multi-factor ones is puzzling.

At the same time, Sunstein's work is frustrating in many of the ways that much of the H&B work on non-moral judgments and decisions proves frustrating. One can rarely ascertain when he believes people are using conscious rules of thumb, learning recurring patterns that would aid them in making decisions when consciously trying to achieve a particular goal over and over again, or making use of once-adaptive judgment mechanisms whose goals are fully opaque to consciousness. Yet deciding which of these sorts of decision-making heuristics is being used matters a good deal, both in figuring out whether or not we are dealing with a problematic bias and in figuring out how to remedy anything we decide is problematic. And like H&B theorists more generally, he tends to report on rather particular findings of poor judgment – for instance, people subject to a "betrayal heuristic" irrationally accept higher risks of death if they can avoid being killed by a protective device – without giving enough detailed description of the content or origin of the heuristic to permit us to predict what its domain would likely be.

I am probably more sympathetic to H&B work generally than to F&F work. I suspect I would rather that work be vague than wrong. I think that is especially true when the scholars making claims that seem just plain wrong to me are, like F&F scholars and like Mikhail, much more prone to believe that they have discovered, and must act as proselytizing messengers, of an especially important Single Simple Truth, whether the Truth is the F&F truth that people are lexical decision makers who do better than those who account for more features of a situation do or Mikhail's truth that we are all born with the capacity to learn just one answer to some non-trivial set of moral problems. Sunstein's edifice does not collapse if we decide he has, for instance, misconstrued when and how people may make self-destructive decisions because they overestimate the harms caused by programs or devices that on balance increase safety. That is true, of course, largely because there really is no edifice. But Mikhail's view of how we do and should evaluate, say, rules about torture or killing innocent civilians in military raids (if not his views of how every significant aspect of a criminal code is determined) really do depend on being convinced that there are stable answers, at least to paper and pencil Trolley Problems if not to cognate problems that people confront in distinct real social situations, and there is not a tremendous amount of give in his views that permits him to account for what strikes me as significant discordant data.