

THE HEURISTICS DEBATE: ITS NATURE AND IMPLICATIONS

Mark Kelman – Stanford Law School

Condensed portions of book draft for 9/6/07 NYU Colloquium

Not for Quotation or Attribution without Author's Express Written Permission

Part One: The heuristics debate

At some (high) level of generality, there is considerable overlap in the way pretty much everyone interested in heuristics at all thinks about heuristics: At some level of generality, there is widespread agreement that people are employing heuristics whenever they make a judgment without making use of some information (that could be relevant) or some computational abilities (that at least some people possess). Again, there is agreement as well that using strategies that are plainly not formal optimization strategies is, sometimes, absolutely necessary. Many of us can “know” enough about the flight of a fly ball in baseball to catch a ball hit quite far from us even though there is lots of (potentially and actually) available information available 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 (apparently) 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 our eye on the ball. They adjust our running speed so that the angle of gaze – the angle between the eye and the ball – remains constant or within a small range.

At a high level of generality, too, everyone agrees that heuristics are often “functional” – they produce 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 (again, given at least temporary consensus on the definition of dysfunctionality) Moreover, there is widespread agreement that in a multiple actor setting in which one agent 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 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 some) about both advertisers and sneaky lawyers.

At a high level of generality, all agree that it is often easier or preferable to change the environment in which decision makers function 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: In that sense, the disposition to use heuristics may (at times) 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 more, or even give negative or positive 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 what is often labeled the

“heuristics and biases” school (H&B), most associated with the Nobel Laureate, Daniel Kahneman and with Amos Tversky. What I explore in this book is not so much the impact of that literature but the *debate* between proponents of the heuristics and biases school and those associated with the “fast and frugal” heuristics school, (F&F) most associated with Professor Gerd Gigerenzer. Those in the “heuristics and biases” 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 miscalculate probabilities and miscalculate end states....[It might, at very first blush, be described as partly gloomy, because attuned to the many errors we might make in meeting our ends because we use certain simplifying “tricks” to make judgments of fact and value that mislead us, in different ways, in many settings. It might also be viewed as reformist/meliorist, because grounded in the supposition that conscious efforts can move individuals and institutions closer to meeting their stable aims.]

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, are considerably less interested in “biases” or errors than in *achievements*....[They] emphasize the degree to which the heuristics that we use will far more typically (though not invariably) be either adequate to the decision-making tasks at hand, or 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. ..¹

¹ 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?)-- 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

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 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”?

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 they 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 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.

- 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?
- Somewhat (but not entirely) 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 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 unmotivated to solve problems “well”? Do individual differences (if real) suggest that heuristics are a response

(largely) to internal limits, not features of the external environment? Does the (purported) existence of such limits imply (instead or additionally) that we have different capacities to “override” heuristics? If so, is it wrong to characterize heuristics as strongly informationally encapsulated cognitive responses to inputs? Finally, does the fact that some people “avoid” 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)?

- Do people (often, rarely, or never) consciously employ...heuristics? Are heuristics (at least sometimes) the deliberately chosen strategy of a cognitively-generalist mind or do people use them without being aware *that* they are using them or why it might be advantageous to be using them in a particular setting?
- To what extent should we expect significant problems to arise from the use of heuristics? To what extent should we encourage the use of new heuristics (assuming that heuristics can *ever* be adopted consciously)?....

a. *Brief 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 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...²

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 immediately, rationally relevant one (its probability.)...[For instance, when using the availability heuristic, individuals estimate the frequency of an event or the likelihood of its occurrence (or recurrence) “by the ease with which instances or associations come to mind.”]³

² 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 book, 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...

³ I hope at the colloquium we set aside debates (included in the book) between those who view probability (rather than sampled frequencies) as subjective and those who view probability judgments as objective. The controversy is relevant to some of what I find the least interesting disputes between F&F and H&B.

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 generally easier to detect incoherence than inaccuracy, of course, since assessing inaccuracy requires that the experimenter herself knows the actual probabilistic distribution of the phenomena at issue.)⁴ 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 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....

Not only do H&B researchers detail ways in which people fail to assess accurately (or coherently) the probability that certain outcomes will arise if they choose a particular option, they also attempt, not surprisingly, 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 for 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

⁴ 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.

frequently attempt to demonstrate that the use of heuristics generates intransitive preference orderings or violations of dominance rules.⁵

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 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.⁶

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 of “substantively bad choices”, 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” (or...is troublesome) is hardly obvious...[for a slew of reasons not worth belaboring at the workshop.]

⁵ Thus, for example, when presented the (rather transparent) choice between (1) a 25% chance to win \$240 and a 75% chance to lose \$760 and (2) a 25% chance to win \$250 and a 75% chance to lose \$750, subjects know to choose (2) (which dominates (1), promising *both* higher gains and lower losses at equal levels of probability.) However, if asked to make the following pair of concurrent decisions between (3) a sure gain of \$240 and (4) a 25% chance to gain \$1000 and (5) a sure loss of \$750 and (6) a 75% chance to lose \$1000 and a 25% chance to lose nothing, the vast majority of respondents choose the combination of (3) and (6) over (4) and (5), though combining choices (3) and (6) yields dominated option (1) while combining choices (4) and (5) yields the dominant option (2).

⁶ 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...

[Since they believe that people will frequently fail to behave... “rationally”...the question arises: Why?]....I think 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.... Many (but again, 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.⁷

System Two thinking is, in this view, pretty much the opposite: It is at core rule-based, analytical, conscious and 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, though like many of the F&F people, many H&B theorists seem to assume that the choice to use a heuristic is sometimes conscious and deliberately processed rather than automatic.⁸

⁷ 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.

⁸ Most H&B “two system” – or dual process -- models assume that System One thinking inevitably (or automatically) occurs and is simply (sometimes) overridden by System Two. It appears that most H&B

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. As a result 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....

b. 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 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....⁹

researchers assume that when we “cognize” a problem, the systems engage in something like temporal order. Accounts of when, why and how “correction” occurs are not terribly lucid or developed in the H&B literature, in precisely the same way that we will come to see that accounts of the “domain” of domain-specific responses is poorly specified in F&F (and massive modularity theory.) Many H&B theorists seem to believe that System 2 permits us to solve the many novel problems we face and to establish more individualized goals and plans.

⁹ Of course F&F people frequently and forcefully emphasize that optimizing is not feasible because of limitations that could best be described as internal – that full-blown optimization would require that persons had the minds of super-computers. In fact, F&F researchers are (even more) prone (than H&B theorists) to disparage the way in which economists and other full-blown rational choice theorists understand bounded rationality. Economists typically understand heuristics as consciously chosen rules of thumb to be employed

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. [In ways we might discuss at the session, I find all three of these points problematic: problems are not intrinsically intractable or tractable; 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.] 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... permitting “less is more” effects (superior performance based on less information)....

when the “costs” of gathering and processing more informational cues would outweigh the expected benefits of increased precision in decision making. If that view were correct, say the F&F theorists, then people would require even more super-human powers, because knowledge of the expected value of additional information requires truly enormous amounts of information and processing capacity.

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....

F&F researchers not only posit that boundedly rational thought arises in a particular fashion – the organism “fits” its adaptively evolved capacities to environments in which the use of a particular capacity will meet its proximal needs -- 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 that I will explore in detail (in Chapter Eight) in the context of making judgments about relative city size. 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 – 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”...¹⁰) 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 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 one that I will continue to explore (mostly in the omitted material on massive modularity theory)... 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

¹⁰ I explore in detail in sections not included in this excerpt 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, e.g. Mark Kelman and Nicholas Richman Kelman, “Revisiting the city recognition heuristic” (2007);

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 heuristic-based biases) it is because subjects need not generally *choose* what sort of decision-making process (or how many cues) they will use....

c. Cross-cutting critiques: what the debaters emphasize

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....Further implicit in suspecting that there is likely a disjunction between “poor” laboratory performance and true human capacity is the commitment to the idea that capacities are (generally? always?) fairly (entirely?) domain specific. F&F theorists, who believe that cognition is significantly domain-specific wouldn’t expect people to get better at solving problems that impact inclusive fitness simply by developing more generalized cognitive capacity that they could utilize to solve any set of novel problems,

but rather by utilizing narrower algorithms that solve a narrower set of problems-they-really-face...

Whatever its ultimate *origins*, 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. [I will discuss this point at the presentation if people are interested. For some of you, it might merely remind you of a familiar debate.]¹¹

¹¹ 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

- *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 EEA), even though solving the problem seems to involve no more formal math skill than solving 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 [and once more I will discuss this more if people are interested]..¹²

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?”

¹² 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

- *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 the laboratory game. Fundamentally, they do so, however, because they ignore the instructions they have been given – they have confronted them 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 [and we might discuss the debate if some find it helpful to do so.]¹³.

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.

¹³ 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-

- *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, 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 explanation for them to interpret the questions as they do) and 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

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 $.7 \times 7 + .3 \times 3$ 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 might note that I remain utterly befuddled by the claim that experimental subjects should be expected to “confuse” these two games.)

proceed. One can probably understand this particular general controversy well by reflecting on... certain F&F critiques of the conjunction fallacy experiments.¹⁴

- *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 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 and think we are most likely to be able to identify mental capacities/mechanisms not simply by observation, but by reasoning

¹⁴ 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”).

backwards from the “need” (in inclusive fitness terms) that the organism had to meet to the capacity it must have developed.

Because, for example, H&B theorists 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 (and these capacities are either used in settings in which they were originally adaptive or co-opted – often with good results, sometimes not – when it is possible to use the capacity to cope with an environment that is novel from the vantage point of those who utilized capacities only in the EEA).

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....

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 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. It is vital to recognize that this derogatory 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”¹⁵

¹⁵ In the book, 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 on “recognizes” and then 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

Second... different people, with different cognitive abilities and “thinking styles”, may systematically use heuristics differently... are at least mildly incompatible with a number of aspects of the F&F view... Systematically differential use of heuristics is hard to square with the claim that we should (almost invariably) understand heuristics far less in terms of computational deficits and far more in terms of locating apt fits between computational capacities and the external environment... [Furthermore] non-universality of usage of heuristics... is not especially readily reconciled with the claim that using relatively mandatory heuristics serves important adaptive functions....

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 [which they see as incompatible with the “imperfect” commitment to rationality...]

Part Two: Moral heuristics v. moral competence

At times, legal theorists have invoked the debates I have alluded to rather directly. ... Cass 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

stimulus or not); as a result, many items will be very mildly familiar, but not so familiar that a person will even 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

many of our intuitions are merely ‘rules of thumb’ that are inaptly applied to some of the situations in which we apply them. His argument to this effect quite explicitly draws on the “heuristics and biases” tradition.¹⁶... At the same time, John Mikhail’s reply to Sunstein¹⁷ (and others who either question the normative validity of moral intuitions or argue that no such strong intuitions exist that are not culturally contingent and learned¹⁸) draws, if somewhat less explicitly, both on aspects of massive modularity theory and aspects of the “fast and frugal school”. Once more, Mikhail is quite explicit that he believes that we are readily able to make non-reflective judgments that we cannot readily explain that are in fact grounded in our capacity to process a quite small number of features of a decision situation. He implies, albeit less explicitly, that these specific-cue responsive judgments fit 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

¹⁶ Here are the basic texts by Sunstein addressing this issue that I will be examining: 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).

¹⁷ The texts by Mikhail that I see as most central are: 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 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 difficult to understand Mikhail’s argument without comprehending Hauser’s, and difficult in turn to understand Hauser’s without understanding de Waal’s.

¹⁸ 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. John Mikhail, “Law, Science, and Morality: A Review of Richard Posner’s *The Problematics of Moral and Legal Theory*,” 54 *Stan. L. Rev.* 1057 (2002).

existence of a Universal Grammar... One might further say that Mikhail implies (less directly still) 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) adaptive and entrenched in something that could be described as “human nature.”¹⁹ 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 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.²⁰ ...

a. *Describing Mikhail’s 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...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

¹⁹ Hauser seems to me considerably more committed than Mikhail to making an (inexplicit) 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)...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.

²⁰ Plainly, the Sunstein/Mikhail debate has some immediate practical relevance: 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 unalterable. Moreover, many questions that might otherwise seem morally (and legally) vexing – the permissibility of various forms of more-and-less active 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 considerably less vexing than they first appear if certain responses 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.

concrete beliefs about when behavior is either morally obligatory or prohibited that exists across all (reasonably healthy/functional) persons, without regard to either cultural background or personal ideological idiosyncratic 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... beliefs about social ordering and morality. 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, the possibility of intolerance or ethnocentrism.... But the skepticism also comes from...liberal political theory...that even within relatively homogenous cultures, moral battles are ubiquitous: In fact, in this view, [creating] a functioning liberal state [requires] avoiding 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 strife....

Second, Mikhail believes that the mainstream position expresses a certain psychoanalytically-rooted 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 with any set of normative commitments. Once more, I need to come back to the question of whether "moral realists", 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 more than the *competence* to recognize the attractive or the hunger-satisfying but tends to *impel action*. 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, but his position on this issue is hardly unambiguous...

Third, there is a distinct, commonplace "modernist" intuition about human nature that Mikhail also rejects....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 to harm others. This jaundiced view of the relationship between moral codes and human nature, dubbed "veneer theory" by de Waal, is that while...short-term helpful behavior may be "naturally" stable – either when the party helping another bears no costs as part of an immediate cooperative venture (mutualism) or when reciprocity from the benefited party is (more-or-less) guaranteed – if we are to expect "moral" judgments in other situations, we need 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 (either because the “rules” are established by external, presumably divine authority or 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”; the source of (whatever) universalized moral injunctions we observe is within our minds, the cognitive mechanisms that permit us to acquire moral beliefs.

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 competence to recognize a host of things, such as the distinctions between sentences and pseudo-sentences.... Naturally, in thinking about the questions that preoccupy legal theorists, questions about the degree to which any set of precise moral reactions to novel 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 to learn certain sorts of moral rules is to have pragmatic significance, then many abstractly conceivable moral rules must be effectively off the table, “unlearnable” as moral rules...

How might Mikhail’s basic viewpoint be profitably distinguished from claims that could be described as more capacious...First, one must figure out what sorts of judgments (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 ambiguously embraced 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. de Waal is occasionally ambiguous on 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, 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 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 reflection²¹, and

²¹ 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.”)...

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...)

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 their violation 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...thought of as something-other-than-customary and (therefore) not capable of being waived by authorities (who 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 UMG (Universal Moral Grammar). His account is 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 the doctrine of double effect -- whose

cognitive structure he believes is at least reasonably well-understood. But Hauser's rather thinly specified UMG – focusing 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 (cultural variety occurs for Hauser because, for instance, the “meaning” of different consequences may vary across cultures) -- seems to permit a broader array of “topics” to be covered as moral topics. Hauser certainly explicitly treats incest-avoidance rules as part of our “moral minds.”²²

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 “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... plainly believes that the *capacity* to distinguish moral from conventional rules is one of the key in-born forms of moral competence: The fact that very young children can distinguish between a

²² 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 there are 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. 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 not to cheat – 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. 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. Because it *looks* like a rule that could be described as “moral” (in at least some of the many substantive senses of the term), Hauser misleadingly thinks of it as a moral rule parallel to those more specifically acquired by the morality acquisition module.

moral rule (“don’t hurt your school mate”) from a conventional one (“don’t wear your pajamas to school”) even when both have been (at least ostensibly been) articulated as rules, whose violation is subject to punishment, is a critical piece of (in my mind quite weak) evidence for him that we are born with significant moral abilities. 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 a moral 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 prohibitions of homosexuality?) might well seem like moral, not conventional rules. Who might even be around to waive them?

There are other largely “procedural” definitions of what a “moral” rule is... 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; 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 in-born competencies have.

However one believes that Mikhail ultimately divides moral judgments from other sorts of judgments about behavior that is thought of as undesirable (without regard to one’s self-interest in engaging in it? without regard to the fact that it seems appealing or

tempting to some significant degree?) or justly punishable, it is plain that he does not believe that *moral behavior* is nearly as universal as moral *judgment*. Mikhail is much more convinced that our initial judgments about whether behavior is obligatory, permitted or prohibited are the same (in “normal” people) than that our conduct, given this shared judgment, will be the same.... 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.²³...

What then is the content of Mikhail’s UMG? Well, at this point, I would say it still seems rather thin (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 recognize that there is a rich debate in linguistics not only about whether there is a universal grammar, but also about which rules such a grammar really contains if it exists. But my point for now is a far narrower one: there are many, many more rules, described at a much higher level of detail, that purport to describe the features all languages must possess than there are equivalent UMG rules at this point.

²³ Oddly, while it is F&F theorists who frequently complain that H&B researchers find incompetence in answering decontextualized puzzles, here, it appears, the H&B theorists’ argument is that we may see illusory “competence” by stripping 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 (clearly violating the double effect principle). It may also 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”.

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 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. I am also not entirely clear whether he believes that all moral systems must distinguish, or 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.) Second, all moral systems require the maintenance of the 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 is intentionally causing death without justification; rape is forceful sex etc.)

Obviously, it is possible to be skeptical about the significance of the claim that these are meaningful universals, or that observation of 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

retrospectively, the structure of any rules...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. 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). 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 adequate justification (finding cheating spouses, infidel detection etc.)...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.²⁴

At the same time, 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 subjects will distinguish cases in which an agent acts

²⁴ There is a monstrously long note explaining (or refining) Coase’s critique of the feasibility of making factual unidirectional judgments of harm-causing, but I hope most academic law sorts at the colloquium won’t need it.

impermissibly because that agent commits one or more distinct 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 subsequent foreseen side effects of an action taken for clearly beneficial purposes. Mikhail clearly believes that the distinct cases trigger distinct mental representations and that once represented as they naturally are represented, the evaluation of the overall action follows inexorably.

b. Exploring Sunstein's moral heuristics

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. (Note, that 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.) In each case, 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 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 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 recalls

events because they re salient, not frequently confronted), 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 ends²⁵, 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 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. (Another way of putting this point is that he is 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.) ...

This primary ambiguity in identifying the origins and basic nature of the moral heuristics is at least partly responsible for a secondary ambiguity in the argument: To the (uncertain) 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

²⁵ 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.

heuristic *misfired* in certain settings. If the *goal* of a rule of thumb is to get its user to make some judgment J 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 no more than arguably meet some end that Sunstein (or any “observer”) ascribes to them, he 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....²⁶

If (for instance) 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 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? ...

Ambiguity about 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, 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

²⁶ If people want to explore this tri-partite division further, it might help to think about ways in which employers statistically discriminate as an analogy...

problems with case-by-case decision making. It is simply to note that it is sometimes difficult to tell whether Sunstein is trying to identify a particular set of troublesome heuristics... or to restate... a more general critique of the use of rules or proxies.

Ultimately, though, what Sunstein seems most certain of is that, as a result of some unspecified processes (that either have occurred to us as individuals, through self-developed processes or explicit learning or that have occurred to human beings more generally), certain reactions to “moral problems” are (at this point in time) 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 sort of) “moral” concerns. I am quite confident that he does not limit the domain of moral judgments to judgments related to a certain class of substantive topics (e.g. harming and helping)...and I am (marginally) confident that he does not draw Mikhail’s strong procedural distinctions... I think, again with very little confidence, 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....) What I 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 guarantee the presence of a certain de-limited 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.²⁷

The 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 (or, to put it more modestly, in which justification defenses to prima facie wrongs are exceptional rather than routine) and situations in which the optimal harm level is (plainly) positive. It is valid in cases in which the routine 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...: Instead of evaluating

²⁷ 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 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 wrong (e.g. there are more words beginning with ‘r’ than words whose third letter is r; more deaths from airplane crashes than household falls.)

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. (Think in this regard about 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 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”.)

Second, there are a series of what Sunstein sees as “biased judgments” associated with the use of the “outrage heuristic” in punishment²⁸: what they 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. (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.)...

Third, Sunstein believes that there is a “moral heuristic” against “tampering with nature” or “playing God” that leads people both to over-value outcomes they see as more natural... to over-demonize novel technologies that seem to substitute for existing natural processes (resistance to cloning, stem cell research, even IVF) and to misestimate the relative risks associated with “natural” and “man-made” events.

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

²⁸ Sunstein explores this class of moral heuristics further in Cass R. Sunstein, “On the Psychology of Punishment,” 11 *Supreme Court Econ. Rev.* 171 (2004).

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” (taking account of consequences, including the violation of imperfectly constraining deontological principles when evaluating action). He acknowledges (too weakly, I am sure, to meet the objections of readers committed to 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.

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 transparent metric the agents must 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 reaching judgments that are not consistent (in cases in which no betrayal effects are present, they prefer a 1% risk of death to a 5% risk while they prefer the opposite when the betrayal heuristic is activated.) But the decision could be “procedurally” suspect in at

least two distinct ways: It might be *unstable* (subjects would renounce it if the decision's features were pointed out to them) or it 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 history in the H&B literature generally. It is worth recalling then, the debate over the claim in the traditional context... 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. (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.) 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....

Arguments from "inconsistency" seem to take two forms. In the less controversial form, a response is inconsistent when the same outcome is (morally) evaluated differently merely because it is described in a different fashion....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

(moral?) 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; obviously, this merely restates the classic H&B gain/loss aversion asymmetry experiments. 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....

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 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.

I take it as well (though his discussion of the Trolley Problem strongly suggests that the target attribute that he believes that subjects are trying to identify is the attribute that would be accepted by a pure act-utilitarian) 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 or (though this second inconsistency is less plausible) failed to blame some who desired it. (It is a subject for a far richer debate than I detail in this excerpt 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....)²⁹

I set aside for now the obvious problem 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-relationship-to-that-principle? 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.) 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

²⁹ Mikhail 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 the battery leading to 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.

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. (I hope that my efforts to summarize their work expressed a particular sort of criticism that I won't dwell on much 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.) 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 in the first three parts of the book...

a. 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 (for a variety of reasons I tried to summarize earlier), 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 uninteresting, set of laboratory settings that may not demonstrate true pragmatic moral competence but a form of abstract problem-solving ability.)

Instead, I think, they would focus on the second two 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. Worse still, it is difficult to ascertain either 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) or whether one could really see the judgment as resulting 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) 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 contemplates the view that they generate mistakes, and we should try to correct these mistakes. (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.) He further contemplates the view that while they generate mistakes in individual cases, we would make more mistakes overall if we did not use them all of the time and tried to pick out situations in which it would be helpful to drop them. (That is to say, the error *rates* created by forswearing 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.) 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 what could best be pictured as broad classes of cases in which they systematically outperform non-heuristic reasoning in each case.

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 the heuristic in a well-defined cognitive capacity.³⁰ 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.

³⁰The 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.

Does the heuristic operate only when safety devices harm or kill? Would it extend to finding annoying aspects of vacations much more painful 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 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.³¹

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 utilize 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 discussed), and has the same sort of (purported) adaptive impact. Just as we (purportedly) 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

³¹ 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.

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 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 sort of problem....

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 F&F researchers) 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, Mikhail’s claims that the “heuristics” might be the inevitable product of a morality-acquiring module (and may give rise to universally held judgments) tells us nothing at all about 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.

b. 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 they 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*. In the typical case, H&B critics suspect that the basic 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 gets 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.³² It turns out that the question of whether Mikhail thinks of moral judgments as encapsulated depends on resolving definitional questions that I noted are

³² Thus, recall from the critique of the existence of the recognition heuristic in chapter 8 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.

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 (some of which I suspect 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?), 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) judgments are (strongly, modularly) immune from reflection or that 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, for our discussion, is to think about both these issues in relationship to the judgments on Trolley Problems that have most preoccupied Mikhail (particularly those that do not involve “personal violence” – throwing someone from a drawbridge to block the trolley v. diverting the trolley – but those that merely alter 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: while 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. 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; Mikhail is not so much observing a capacity as imagining one he thinks “fits” human needs.

Worse still, the “granularity” problem³³ and encapsulation problems that beset all modular and “softly modular” theories are enormously bothersome here: Mikhail is confident that he is observing “double effect” reactions, but I have no idea why (this is the granularity problem) or whether he thinks the “double effect” reactions are just one input into “moral judgment.” (This is true in part because I am not sure what he thinks a moral judgment really is.) Jack Bauer – the hero of TV’s mega-hit *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. Do the *24* viewers represent the choice in a fashion distinct from 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, even if the cue is significant (i.e. is the judgment non-lexical?) Or, as social psychologists have long suggested (to a degree that I

³³ All theorists committed to domain-specificity run into the problem that a domain can be specified at broader or narrower levels of generality: is cheater detection a sub-set of reasoning about deontic conditionals or is it 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 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).

confess I often find excessive), 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?)