THE HEURISTICS DEBATE: ITS NATURE AND IMPLICATIONS
Mark Kelman – Stanford Law School
Condensed portions of book draft for Columbia Law Workshop – February 16, 2009
Not for Quotation or Attribution without Author's Express Written Permission

NOTE: pages 1-22 of this draft are an effort to summarize the first seven chapters of the book, which are designed to both summarize and interpret the "heuristics debate". The section on criminal punishment is a condensed portion of one of the five "application" chapters (the application chapters are on "moral universals", Langdellian formalism, incommensurability, criminal punishment, and information disclosure.)

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 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 their eye on the ball. They adjust their running speed so that the angle of gaze – the angle between the eye and the ball – remains constant or within a small range.

At 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 actor may not treat another's interests as if they were her own, the fact that we employ heuristics can be *exploited* by those who have the capacity to manipulate an environment so 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, or even give incentives to do a better job…

The vast bulk of the literature in both law and the policy sciences that has made use of the concept of heuristics has been literature drawing on 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 miscompute probabilities and misevaluate 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… 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 even 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]

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

[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 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.

which we must solve problems, so that our task is first to understand the *fit* between cognitive capacity and environmentally-established problems?

- What criterion does each school use in evaluating whether a judgment or decision-making process is “rational”?
- To what degree do theorists in a particular school believe that judgment and decision-making is (mildly, substantially, or absolutely) “informationally encapsulated”? Are people capable of “overriding” heuristics... when they make a judgment, using cues beyond the informationally limited ones that would trigger a particular judgment outcome if they simply employ a particular heuristic?
- 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 frequently if we were (somehow) “smarter”?
- Does the theorist assume that all (functional) adults are equally likely to use both useful and dysfunctional heuristics? If some people with particular traits (e.g. higher intelligence, conventionally defined; certain personality traits that are

generally associated with “open-mindedness”) are less prone to use.. [some dysfunctional] heuristics, does this imply that we use heuristics because some, but not all of us, are computationally limited or inadequately motivated to solve problems “well”? 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

[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…

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.”]

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.)[3] 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

[3] 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.

obviously troublesome when they violate coherence rationality – they are, for instance, intransitive or violate dominance rules. Not surprisingly, then, H&B researchers frequently attempt to demonstrate that the use of heuristics generates intransitive preference orderings or violations of dominance rules.

Further, and more significantly, the H&B theorists typically argue that they need not have substantive views on what tastes are "objectively preferable" to argue that people are not evaluating end-states properly if the evaluation of such end-states is frame-sensitive. H&B theorists have been especially adept at exploring situations in which some end-state X is evaluated as better than Y if the outcome X is described in one fashion but not another or if X is evaluated as better than Y only if there is some irrelevant third 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.[4]

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 probably not worth belaboring at the talk.]

[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

[4] 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…

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.[5]

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…

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

[5] 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.

individual distinctions in the *capacity* to engage (in more situations) in System Two thinking.[6]

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 nor to optimize in the fashion H&B theorists suggest we should even if we had limitless computational powers….[7]

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 conceivably discuss at the session – though I find this set of points pretty uninteresting -- 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,

[6] Thus, people who are trained in statistics are (modestly) more likely to override the use of (many) heuristics. Similarly, people who are more "intelligent" (in the sense measured by traditional "g-loaded" tests, like IQ tests or the SATs) use many of the heuristics less frequently. The point, for this group of H&B theorists, is not that the "sort" of intelligence that g-loaded tests measure is the only sort of relevant intelligence (or even the most important), but that it is a genuine measure of *something*. That something appears to be the capacity to manipulate non-contextualized formal symbols in accord with the dictates of conventional rational choice theory….

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

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)….

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… 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 explore in (what can only be described as ungodly) detail 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"…[8]) The interesting point for now is how F&F researchers have reacted to H&B findings that

[8] 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);

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 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) to 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…. 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.][9]

[9] 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]… [10]

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?"

[10] 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.][11].

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.

[11] 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 X 7 + .3 X 3 or 5.8 rather than 7.
One *could* figure out what choices to make using some (undefined) general cognitive mechanisms (that permit the calculation of expected values in all sorts of situations). Alternatively, one might have developed (at least relatively) domain-specific cognitive mechanism to solve the problem of picking an optimal mix of distinctly risky gain-seeking activities from a small option set that dictates that one will engage in probability matching. F&F theorists, echoing evolutionary psychologists prone to believe that people have developed narrow domain-specific "answers" to problems that presented themselves to our ancestors facing evolutionary pressures argue, for instance, that the "cognate" problem to the urn problem in a natural environment is to pick between foraging sites with distinct probabilities of finding food. The optimal strategy in that setting may not be to maximize expected value, though, but to both get more food and to learn more about unexplored environments, at least when one is satisfied that one has gone to enough high-odds sites to insure that one will be a bit flush with food. (I 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.[12]

- *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

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

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

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[13] 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"….[14]

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

[14] 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 one "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 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

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…

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…

Part Two: An Example of efforts to look at legal policy implications: criminal punishment…

It is important to recognize, right from the start, the considerable limits on our capacity to use insights from cognitive psychology to "frame policy." When I discuss deterrence in this chapter, for instance, I do not mean to imply in any way that we could ever hope simply to *induce* the empirical relationship between a particular policy interventions and the level of offenses by better integrating psychologically-based theories of how people process cues and/or make decisions. What I think we can hope to get from psychological theory – from all theories of human behavior – are two things: First, I believe, the relationship between empirical social science data and theory is invariably dialectic. Empirical studies are never fully persuasive on their own (*every* econometric study has some degree of insoluble problems of omitted variable bias and

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 co-linearity, every “natural experiment” is imperfectly controlled); to some degree, then, we “test” the plausibility of data by whether it makes sense given a convincing underlying theory of behavior. (At the same time, the plausibility of theory is “tested” by its fit with available data.) Second, I believe that theory helps us generate testable empirical hypotheses that we might not otherwise generate. (Without thinking, for instance, about the “theoretical” possibility of hedonic adaptation and peak-end hedonic reporting that I will discuss in some detail later in this chapter’s text, my claim is that one would not think to investigate the empirical possibility that longer sentences might *diminish* specific deterrence.)

The problems of using the insights gleaned from the heuristics debate to help frame criminal law policy – even in the very limited ways I have just described “using theory” – are even more severe, though. The applied psychological literature is still quite undeveloped, and, at this point, rarely tethered to empirical studies (either at the macro data level or the micro experimental level). So, in a sense, all I can hope to do in this chapter is describe the broad sorts of policy-relevant suggestions that would grow out of the “heuristics and biases” and “fast and frugal” literature respectively.

The most significant point I want to explore... relates to what is traditionally referred to as the “deterrence” function of criminal law… {T]hose associated with the heuristics and biases school…believe, like rational choice theorists generally, that the behavior of those deciding whether to violate criminal statutes is (at least to some significant extent) grounded in calculations about the expected value of offending,[15] but that these

[15] Deterrence theorists typically focus on state-influenced, price-based movements along a fixed curve – the degree to which given any level of background taste for both the positive results of crime (how much the defendant “enjoys”, say, vandalizing or assaulting or values the goods he steals) and the negative ones

calculations of expected value (the probability of punishment, the value of what they will gain from the commission of the offense, and the "disvalue" of whatever punishment they receive) are all likely to be systematically distorted. The level of effective deterrence is plainly a function of the *perceived* expected value of offending, and we should not assume that the perceived value is congruent with some sort of "objective" expected value ("objective probability" times the invariant, frame-independent subjective evaluation of gains and losses from successful or unsuccessful criminal efforts). Scholars associated with the "fast and frugal" school, though, are likely to be highly skeptical of the idea that would-be offenders calculate the expected value of offending at all. Thus, decisions about how to behave (to comply or not to comply with law) are likely to be based on responses to an extremely delimited set of cues. Given this view, manipulating either objective expected punishment or perceived expected punishment is skew to the goal of increasing compliance with law. Instead, we need to manipulate the signals that people actually use in making fast and frugal judgments about how to behave.

[In the book, but not in the material I am sending you all for the workshop, I also discuss briefly how distinct ideas about the use of heuristics play out in thinking about retribution (the notion that punishment is justified as an apt response to prior wrong-doing, regardless of the forward-looking consequences of exacting punishment)…. in the context of assessing how we might determine whether a particular punishment is proportionate to the criminal's wrong-doing, rather than excessive or unduly lenient. I also explore, briefly, some problems that those with particular views about the use of

(how bad is any particular fine or term of imprisonment or guilt pangs), the amount of crime will decline if expected punishment rises. Naturally, though, it is also possible to focus on the social and legal forces that shape the relevant tastes….[This is largely outside this chapter's purview…. Further, it is possible to focus on changes in the expected costs and benefits of crime that occur because of private action: e.g. private counter-violence or anti-theft devices.

heuristics in decision-making are likely to see in implementing criminal punishment systems designed to "incapacitate" offenders (to prevent those deemed prospectively dangerous from harming the non-incarcerated population by isolating them from that population while they are particularly dangerous).]

A. Diminishing the crime level

1. The "fast and frugal" approach

At core, "fast and frugal" theorists are extremely skeptical that people would make a decision about how to behave by engaging in the sort of conventional cost-benefit analysis that rational choice theorists assert that those contemplating committing offenses engage in… Their skepticism about the capacity of actors to engage in such analysis is not grounded in the (familiar) idea that the sub-set of the population seriously considering criminal activity is especially unlikely to calculate: the failure to calculate is not in their view dominantly a function of internal limits at all, let alone internal limits that vary (significantly) across persons.

So the fast and frugal critique of the descriptive realism of deterrence theory is not grounded in the commonplace ideas that those who commit (at least some sub-set of "conventional" common law, if not white-collar) crimes might be atypically unintelligent and bad at calculation, atypically likely to be using intoxicants that compromise both the capacity to calculate and the motivation to attend to anything but the satisfaction of impulsive desires, or atypically flooded by the sorts of emotion (like rage) that interfere with the capacity to calculate.[16] Similarly, it is not grounded in the possibility that rational choice theorists would acknowledge that actors (sometimes? often?) lack the

[16] For a very good, analytically mainstream, account of these sorts of hesitations about the "realism" of deterrence theory, see Paul H. Robinson and John M. Darley, "Does Criminal Law Deter? A Behavioural Science Investigation," 24 *Oxford J. Leg. Stud.* 173, 179-180, 194 (2004).

information necessary to make even decent calculations about expected punishment levels, both because they do not even know what nominal punishments are imposed for convicted violators and because there is little information available about objective risks of apprehension, prosecution, and conviction.[17] (In this regard, I also set aside the concern that will clearly preoccupy those influenced by the heuristics and biases school – i.e. that people will systematically misuse available information so that perceived probabilities will diverge from the best estimates of objective probability that could be made using the information that they do have access to.)

Instead, it is grounded in the more general claim that (all) people (typically) use lexical, not compensatory, decision making processes. They will act as if there is a single best cue to search for in deciding how to behave – and the decision whether to commit a particular offense (speeding, failing to recycle, whatever) is such a behavioral decision. If there is no information about the value of the dominant cue (that permits them to stop searching for more information and to make a decision), they will go through a second search/stop/decide process, looking for the presence or absence of a second cue. (And so on.) But the cues they use need not be "substitutes" for expected value calculations (the best method of approximating what they would learn using a fuller cost-benefit measurement procedure were one feasible). Instead, following these heuristics in making decisions might simply meet some other ecologically rational goal (e.g. permit the maintenance of reciprocal relationships that best protect one's offspring from harm.)

In a sense, the particular hypotheses that particular "F&F" theorist make about the compliance decision are less significant to me than the *form* of the hypotheses: Thus, one

[17] The idea that many actors contemplating crimes neither know anything about the content of the law nor about apprehension and prosecution practices is raised by Robinson and Darley, id. at 175-178

can imagine that a particular F&F theorist will believe that the typical search order for determinative, lexical rules that an actor uses (at least in some domains) is first to figure out what one has done before (habit). If the actor has never confronted the choice before, she might then check whether the behavior violates known, strong internal norms. [There is a section in the book at this point about whether these internal norms are frugally processed deontological "intuitions" that resonates in part in the discussion in an earlier chapter on "moral realism.".] Failing to find an answer to that question, she might look to see what others' around her are doing and imitate social practice. Note that there are a variety of forms of "imitation" heuristics – one can imitate the first person or most proximate person one sees, imitate the majority, imitate those one thinks are "successful" or do the opposite of those one has observed when there is feedback that they are unsuccessful. What should be obvious is that as these imitation "heuristics" require more and more complex evaluation of the results of the conduct one is (purportedly) merely mimicking in a fast and frugal way, the distinction between these (purportedly) heuristic strategies and rational choice strategies grounded in analyzing actually available information get very slim. If I only imitate those I deem successful, I need to figure out what I mean by success: presumably, I will imitate (something like) those persons who behaved similarly who got good outcomes (gains without losses) most often. How this differs from doing my best to make an expected utility calculation is, to put it mildly, murky. Finally – if there are no others to observe – the agent might (then and only then) look to see what the formal legal norm demands.

Or, in some settings – perhaps in making decisions about how fast to drive --- one might first look to imitate others (match the speed of prevailing traffic). Then, only if that extremely fast and frugal strategy is unavailable, the decision-maker might look to follow some other fairly simple habit-based norm (e.g. go 15 k.p.h. faster than the posted speed limit.)

The most prevalent F&F argument that people must make these sorts of decisions heuristically is grounded in what strikes me, and others coming more from the rational choice tradition, as a *non sequitur*: The F&F theorists (rightly) note that full-blown calculation (e.g. of the optimal speed to drive) would require the decision maker to have both too much (unavailable) information and to make (overly difficult) computations. To calculate the optimal speed as rational choice theorists (purportedly) hypothesize, the decision-maker would need to know how much faster one can get where one is going if one drives faster and the value of the "saved" time, the number of speeding tickets one would expect to receive if driving (each amount) faster, the cost of the speeding tickets, the number of extra accidents one would cause and the costs of those accidents (to the degree they are financial costs only, one would need both to know the pecuniary damages one would suffer uninsured and the injuries to others one would generate, the likelihood of being caught and sued,, and the range of jury verdicts as well as know about insurance coverage), the subjective value of the "thrill" of fast driving. But rational choice theorists do *not* require that actors fully calculate the expected value of each action in that way. Rational choice theorists acknowledge that actors may well use any of a variety of starting places (rules of thumb) – and past action or action of those around them are perfectly good candidates to serve as rules of thumb.

All that rational choice theorists assert that is inconsistent with the F&F picture is that the actor will be sensitive to *shifts* in expected value and make decisions in a compensatory, non-lexical way: Thus, he will slow down (all else equal) if he sees a police car up ahead (changing the probability of detection) even if no one else sees the police car so that they all keep driving fast; he will speed up (all else equal) if he is rushing to an important meeting rather than driving to the airport to catch a plane that won't take off for hours; he will slow down (all else equal) if his windshield wipers are working poorly during a storm, and he is more than typically scared that he will get into an accident if he is driving more quickly. Those who make use of encapsulated heuristics lack the capacity to integrate any such facts into judgment (because they are not themselves judgments generated by in domain-specific algorithms.) When one thinks about "decisions" to commit serious crimes – robbery, murder and the like – the F&F idea that they occur as responses to single lexical cues seems especially implausible: people seem to assault or kill when they are especially enraged (the gains of the behavior are atypically high), not because they do so habitually or imitating others (no one kills that often) and commit property crimes when the chances of detection are atypically low (or when they face immediate need for resources, as drug addicts might.)

It seems far more plausible, though, that the changes in expected punishment that the legal system can generate either by shifting nominal penalties or altering global, but not locally and immediately perceived, rates of apprehension, are especially unlikely to influence behavior even if it were utterly daffy to think that actors are unable to use compensatory information in making compliance decisions. It is not nearly so obvious that the fact that Town T catches 15% of speeders and City C only 12% will influence speeding decisions in the same way that seeing a police car around the bend will, or that levying $250 fines rather than $100 fines for speeding will have the impact on expected costs that realizing one can't see very far through the storm has. This may be true because of the barriers to perception that H&B theorists specify (that I will discuss shortly), but it may simply be true that within a wide margin, shifts in expected punishment are simply of very limited moment. If that is the case, though, then we should use alternative strategies to increase compliance, if increasing compliance is our goal, rather than to fuss with rules that would abstractly appear to shift expected punishment.

Again, studying and evaluating the precise mechanisms that scholars influenced by F&F theory have suggested might help increase compliance are less important at this point than thinking about the types of strategies they are considering. The truth is that the literature applying F&F theory to legal compliance is essentially too new and too sketchy to have generated well-developed or empirically tested suggestions. Still, there are a few points worth noting:

First, F&F theorists are likely to share with a variety of legal scholars the view that positive law works best when it mimics pre-existing social norms. For these (limited) purposes, what drives this belief is not the idea that social practice tends to evolve in functional ways, but simply the idea that if law is to be effective, it must be obeyed. For law to be obeyed, it must either be known (and people can readily learn law if they have already learned most of its content in the course of ordinary moral education) or obeyed without reference to law, but by reference to the observed behavior of others, behavior that likely tracks social norms….[There is a discussion of "mistake of law" doctrine at this point and the F&F defense of broadening the scope of such defenses.]

Second, F&F theorists are likely to believe that one can generate compliance with new laws only by investing heavily in making compliance habitual: In this regard, the success of Germany in generating high levels of waste separation by households for purposes of facilitating recycling was *not* grounded in setting penalties for non-compliance or rewarding those who did what was desired. (Even if "expected cost" altering mechanisms were generally effectual, they were impractical strategies in this context, given the high transaction costs of imposing fines or rewards and the incongruence with norms of household privacy that would have been entailed in observing each household's behavior to reward or punish it.) Instead, household behavior was initially changed from the ground up – there was enormous effort spent indoctrinating school kids on the necessity to separate waste from kindergarten on with the hope and expectation that parents would not disappoint their children's moralistic expectations, and a fairly high level of advertising aimed at adults. The need for making these start-up efforts has gradually eroded, though, because waste separation has become so habitual that Germans merely follow the first lexical rule ("what have I done before?") in making decisions about how to handle household waste.

Third, compliance may often best be induced by altering the potential violator's immediate *capacity* to violate norms. While it is possible to think of "barriers" to non-compliance as merely raising the costs of non-compliance, F&F theorists seem to think that those who comply because it is immediately difficult not to do so are making decisions based on a single cue – is action A readily achieved? – rather than "costing out" the difficulty of taking the non-preferred action as a prelude to making an overall calculation of the costs and benefits of non-compliance. In this regard, one should think of using speed bumps, rather than higher fines, to restrain speeding or seat belt interlock systems rather than laws against driving without belts. More subtly, though, one might imagine that the use of certain sorts of road markings and curves in the road make people perceive themselves as going faster than they actually are and thus induce people to slow down.

Fourth, and finally, F&F theorists typically do not disdain the use of either formal state punishment mechanisms (fines, prison) or informal social control mechanisms (gossip, negative reputation, guilt-producing educational messages). But they typically argue that these are effective in the long run only if they help implant habits (that individuals follow, using the "what have I done before?" search rule) or "social norms" (that work if people follow some sort of imitation heuristic.) I am puzzled by why these establish habits or norms in the first instance (except through rational choice mechanisms), and, more importantly, why they cease to generate rational choice-based compliance over time unless they become habits or norms if they once generated compliance on that basis….

2. The "heuristics and biases" approach

Rational choice theorists, and H&B theorists, agree that what is relevant to a decision maker is the perceived expected value of a decision, even if the perception is inaccurate for some reason. Thus, for conventional rat-choice theorists, for instance, a would-be violator will not be deterred by the prospect of punishment unless she knows that she might be punished. H&B theorists focus less on the external impediments to accurate perception – the lack of available information – and more on internal barriers to processing available information. The expected value of offending depends both on the perceived probability of various outcomes (how likely is one to succeed in committing the crime; how likely is one to be apprehended; how probable is it that one will be subjected to each level of punishment) and on the perceived value of each possible outcome (how good will it feel to "succeed", how bad will the experience of each sort of possible punishment be)…

[There is a brief discussion of conventional "external" barriers to receiving information and on the distinctions between perceiving the *existence* of punishment rather than the precise level of expected punishment…]

While the H&B literature on how would-be violators may actually both assess probabilities and evaluate outcomes is a bit more developed than the F&F literature on the roots of legal compliance, it is one again more important for my purposes to look at the *forms* of argument that H&B theorists have suggested than to assess the persuasiveness of any particular proposition. My sense is that the possible distortions in probabilistic thinking are both more obvious to those who have thought about the standard heuristics than the problems of end-state evaluation, and hence, less thought-provoking (if no less important for policy formation purposes.) At the same time, there are a standard set of arguments grounded in the H&B literature to suggest that those implementing the law will be misled by the usual biases into giving unintended signals about what behavior is appropriate and inappropriate.

a. *Expected punishment miscalculation: misestimating the probability of punishment*

Here are some exemplars of the ways in which H&B theorists suggest that would-be violators may misestimate the actual probability of (each particular level of) punishment…

First, there are H&B stories to suggest that people will systematically *over-estimate* the risk of punishment because people generally exaggerate the possibility of low-probability events. At the same time, if people suffer from an "optimism bias" – as most H&B researchers suggest they do – they might underestimate the probability of *bad* events. Since punishment for offending is both (rather) rare *and* (plainly) bad, it is indeterminate, *a priori*, whether people typically under or overestimate true punishment probabilities. Whether conscious state action (other than making instances of punishment more salient and hence available) can mute the optimism bias – thus magnifying perceived expected punishment -- is open to question. We know, at the level of fairly broad theory, that it will diminish (to some uncertain extent) if actors believe that bad events are thoroughly out of their control, so that the (well-publicized?) use of enforcement techniques (like random audits or searches) that diminish the capacity of the genuinely skilled to evade punishment (so that those wrongly believing they *are* the genuinely skilled will assess their chances too optimistically) may be effective.

Second, unless the probability of punishment is fairly high, there will be little deterrent effect because low-probability events have little impact on behavior, even if people accurately cognize the probability of such events. (We need not imagine that there is some threshold of probability below which people will be thoroughly undeterred: we need only imagine that the marginal efficaciousness of punishment falls rapidly as the probability of punishment drops.)

Third, people may (falsely) believe that the probability of getting punished is lower for them than it would be for a random violator if they have been recently punished: Their (biased) perception (resembling the gambler's fallacy)[18] is that is it highly improbable one will be "caught again" if one has just been punished. To the degree the view is true, it seems especially factually irrational since it is considerably more plausible that ex-offenders will be apprehended by the police *more* frequently than random violators, given that they are on suspect lists and have had criminal confederates who might gain by informing on them.[19] While it may be the case that a number of other biases I will soon discuss suggest that convicted criminals (especially, perhaps, those recently released) will *over*estimate the probability of detection and punishment, the presence of (something like) the gambler's fallacy might help explain (purely from a deterrence vantage point) why we aggravate punishment for recidivists. If those who have been punished (especially recently) systematically underestimate the probability of detection, their perceived expected punishment will be atypically low unless we aggravate punishment levels.

Fourth, at the same time, one might expect those who have been (recently) punished to overestimate rates of punishment because the possibility of punishment is atypically salient and available for those who have been punished, while those who have not been might well under-estimate the probability of punishment because they cannot so readily bring to mind instances in which committing crimes has negative consequences.

[18] For the classic discussion of the gambler's fallacy (ordinarily manifest in something like the false belief that one is less likely to toss a heads, even using a fair coin, if the coin had shown several heads in a row), see Amos Tversky and Daniel Kahneman, "Judgment under uncertainty: Heuristics and biases," 185 *Science* 1124 (1974). In the book's final chapter, I return discuss the "fast and frugal" take on the gambler's fallacy in the context of optimal provision of consumer information. I should note, though, that F&F theorists would be skeptical of the claim that recently released criminals would be swayed by the gambler's fallacy rather than the (opposite) "hot-hand fallacy" (that would lead them to *over*estimate the probability of detection if they had been detected in the past since they would believe that intentional actors – in this case, the law enforcement community – would be responsible for apprehending them and people typically believe intentional act ors will exhibit "positive recency" (the belief that they will repeat what they have been doing).

[19] See David A. Dana, "Rethinking the Puzzle of Escalating Penalties for Repeat Offenders," 11 *Yale L.J.* 733, 742-53(2001)

One might think the tendency to discount true punishment probabilities would be especially pronounced among those who have committed crimes but not been punished (since non-punishment is then the most available outcome). Whether this tendency can be overcome by making the punishment of *others* more salient (without the use of morally dubious and constitutionally impermissible punishment spectacles like the stockades and public hangings) is obviously open to question. Not surprisingly, those influenced by H&B try to find ways to increase the salience of law enforcement without crossing these sorts of moral lines. They might, for instance, argue that it is sensible to make sure that parking tickets be large, bright and visible so that by-passers see that others have been ticketed or like the idea of having visibly marked police cars regularly appear in neighborhoods, even if officers riding in marked cars actually apprehend fewer criminals and thus lower the objective expected punishment level at any given level of spending on police. (Here, of course, the under-specification of the traditional H&B heuristics makes it difficult to generate policy applications: we know at some level that we overestimate the probability of salient/available events but have a weak idea about what the roots of salience or availability might be….)

b. *"Frame sensitive" end-state evaluation*

At the same time as H&B theorists develop both descriptive accounts of the divergence between the objective probability of punishment and its perceived likelihood, and suggest some policy tools that might help overcome underestimation, they argue that we need to be more attuned to the ways in which potential violators are likely to evaluate both the rewards of successful criminal activity and the pain of the punishment they do receive and/or expect they could receive. (We should be alert in thinking about the evaluation of punishment to distinguish general deterrence effects – in which would-be violators evaluate projected punishments that they know about but have not directly experienced --from specific deterrence effects for those who have experienced punishment in the past and are now considering whether or not to violate the law again.)

Once more, I am more interested in the sorts of arguments that H&B theorists have made that emphasize that there may be no context-independent, frame-independent method of ascertaining how people will react to a particular fine or term of imprisonment than I am interested in the persuasiveness of particular accounts. There is somewhat less developed H&B-influenced literature focusing on how labile evaluative reactions to punishment may be than there is literature on biased probability estimation, but it is worth sorting through some of the arguments that have been made, directly, as well as those that one might think follow from the literature on biases.

First, there is no doubt that the decision about whether or not to commit a crime (if made in significant part in response to calculations about the costs and benefits of disobedience) might well be different if the (expected) punishment were more immediate, or the benefits of offending more delayed. The disvalue of punishment is obviously lower because it occurs in the (relatively distant) future (after apprehension and a fairly drawn out criminal process that is not likely to resolve itself, whether through plea bargain or trial, for quite a while after the initial offense occurred) than it would be if it occurred (if at all) right after the crime were committed.

But it is obviously not clear that we should describe such discounting as "frame sensitive" decision making. It would plainly be frame sensitive if an actor reevaluated expected punishment that would occur at the same time differently depending on how the punishment was *named* or reevaluated it because of the presence of absence of an irrelevant alternative. It is not at all clear that the timing of a good or bad event *should* be described as an irrelevant fact about the event, though. It is not at all clear that it is merely a way of "framing" the *same* event. So, if we are to describe would-be violators as irrational or biased in some fashion when they think about the timing of either the rewards of criminality or the punishment, we must mean that they are using discount rates that are inconsistent or irrational, not merely that they are using discount rates. [There is then a discussion of some of the H&B-influenced literature on hyperbolic discounting)…

Second, there is some suggestive evidence – looking at the punishment-imposition patterns of experimental jurors – that judgments about how serious a penalty is *are* frame-sensitive. [I then discuss at some length evidence I gathered with Tversky and Rottenstreich some years back that people may evaluate a punishment as more harsh if it was the extreme option available rather than an intermediate one and if it were "contrasted" with a similar but milder punishment.].

Third, the simplest assumption that legislators, members of sentencing commissions, or judges imposing imprisonment terms might make is that the value of each month or year spent in prison would be relatively invariant. Were that the case, doubling sentence length would impose twice as much pain or displeasure. If it is not the case – if the marginal level of either experienced or remembered disutility either declines or rises or changes in more unpredictable ways – then sentencing policy must be more nuanced. [Obviously, too, the degree to which we can measure the *experienced* disutility associated with punishment may be relevant in making judgments about whether punishments for distinct crimes are proportioned correctly, in retributive terms.] It might also, if actual disutility levels are perceived by would-be violators *ex ante*, affect the level of deterrence associated with each punishment. Questions, though, about how prisoners themselves will experience long terms (and whether their perceptions are in some sense "biased"), how would-be violators will perceive how they would perceive such terms, and whether perceptions of either group are policy-tractable are among the most complex questions in this area, yet are not especially well-addressed in the existing literature.

It is probably more plausible to argue that *specific deterrence* (the impact of experiencing punishment on the calculations of potential recidivists) is affected by either experienced or recalled punishment disutility than to argue that general deterrence signals depend on the experience (or memories) of those who have been punished. (Actual or perceived disutility of past punishment matters to the generally perceived "price signals" only if those who have been imprisoned effectively communicated their views to would-be violators.) It is important to note, as well, that F&F-influenced theorists might be especially skeptical of the claim that actual disutility levels impact the perceived severity of punishment. Although I have found no writing to this effect, I am certain that F&F theorists would believe that to the degree that would-be violators are sensitive to punishment at all, the would-be offenders perceive punishment levels in terms of a single, readily processed cue – punishment length – rather than trying to figure out, more precisely, using compensatory information of various sorts, how they would likely feel about punishment of different lengths. In a "Take the Best" cognitive universe, the single cue that would best signal severity is length of sentence, and additional cues (e.g.

discounting the increased severity of longer sentences because of projected hedonic adaptation) will simply never be processed.

There is nothing in rational choice theory, though, to tell us whether or not the marginal disutility of prison declines (as, say, the marginal utility of income ostensibly declines). What the heuristics and biases literature might do is to identify several reasons why we should expect both sharply declining marginal disutility (hedonic adaptation) and even more sharply declining marginal disutility for those who have had prior experience in prison (a form of desensitization.) At the same time, H&B researchers argue that we should expect people who have been punished to misperceive and misremember (and most plausibly to *under*estimate) the displeasure associated with punishment: If that is true, longer terms of imprisonment may "really" hurt offenders (and perhaps be justified as appropriate in retributive terms for more serious crimes) but they are unlikely to pack additional deterrence punch. Oddly, perhaps, I think the H&B literature simultaneously suggests that those who have not experienced prison will *over*estimate how aversive they will find it. It might also be the case that short terms of imprisonment "immunize" those who have experienced them to longer terms (desensitizing people to variations in punishment).

If all of these propositions are true, the ordinary American pattern of punishment for "professional criminals" – widespread use of low sentences in the early portions of their criminal "career" (as juveniles) followed by harsher penalties for multiple recidivists – could scarcely be worse from a deterrence vantage point. Few people considering committing serious offenses do so from a naïve position in which they overestimate the pain from punishment, few have served long enough sentences that the fact that they underestimate the pain associated with their prior punishment is salient in their decision-making, and many have been “immunized”.

Conceptually, it is clearer that misperception and distorted memory are *biases* or errors than that those who hedonically adapt are either irrational or that they are “misperceiving”.[20] Instead, observing hedonic adaptation simply reminds us that reactions to abstract “goods” (or “bads”) are context-dependent….

“True” hedonic adaptation occurs when people’s “actual experience” of the events in their lives are (highly) path-dependent in a particular way: Events that are at first bad seem less bothersome if they become routinized and expected, and events that were once good seem less pleasurable once taken for granted. The most famous (albeit highly problematic) finding on hedonic adaptation is that those who win lotteries are far less happy than one would think (after the small initial bump of excitement) and that those who sustain serious injuries that leave them paralyzed are not nearly so unhappy as one might expect (once they get over the initial shock.)

It would seem to make substantial sense that one’s affective reactions would diminish once any state – bad or good – were stable. To the degree that the primary role of emotions is to induce action, to impel efforts to change the situation one is in when there is pain and to maintain states that produce pleasure, it makes sense that one would have strong emotional reactions only to *shifting* circumstances, because once a state is persistent, it appears likely that there is little we can do to counter it.[21] Hedonic

[20] It is not entirely clear – in ways that I touch on in the text – that it is even sensible to talk about misperception or distorted memory of hedonic states. The question, in a sense, to which I return, is whether one can say one has been unhappy (or in pain) without knowing it or recalling it. I suspect that giving a decent answer to that question requires both complex philosophical inquiry and a great deal of neurobiological sophistication about how pain (and pleasure) are processed….,

[21] See Shane Frederick and George Loewenstein , “Hedonic Adaptation” in *Well-Being: The Foundations of Hedonic Psychology* 302, 303, Daniel Kahenman, Ed Diener & Norbert Schwarz eds. 1999) (“Because

adaptation to the immutable leaves us room to be sensitive to small, incremental local changes that are now most likely to be action-relevant. Looking forward, I may dread, roughly equally (and quite substantially in each case), confinement in a nine foot or a seven foot cell. If I hedonically adapt to my actual ongoing state – say, confinement in a seven foot cell – I may work to achieve a realistic goal, moving to the nine foot cell by behaving well in prison.

"True" desensitization occurs if, over time, one becomes less sensitive to distinctions in end-states than one once was. That one might adapt without becoming desensitized can be seen if one considers that a prisoner might hedonically adapt to life in a seven foot cell, while becoming *more sensitive* to the distinction between being confined in a nine and seven foot cell. The opposite is possible too. One could find it harder to distinguish between ten and fifteen year jail terms once one has been in prison while finding the average day in prison *increasingly* painful.

Policy makers influenced by the hedonic adaptation literature have tended to argue that increasing prison terms is likely to be ineffectual (as a method of increasing punishment, in retributive terms, or as a method of achieving specific deterrence) Longer terms may even, when coupled with duration neglect, peak/end reporting effects that I discuss below, perversely *decrease* specific deterrence (and punishment) by making long prison terms seem *less* painful than shorter ones.

At the same time, there is a particular form of effect (that arguably mixes desensitization with adaptation) that could have profound impact on optimal punishment practices. Assume that a person would, at first exposure, find a "short" (e.g. six month)

the persistence of an aversive state is an indication that it cannot be changed, hedonic adaptation may prevent the continued expenditure of energy in futile attempts to change the unchangeable and redirect motivation to changes that can be made.")

prison term fairly bearable and ineffectual as a deterrent, but would find a moderate term (e.g. two years) to be significantly painful enough to deter crime. If she *first* experiences the non-deterring six month sentence, though, she may no longer be deterred by a two year sentence; the response to that marginal cue is simply dampened. These sorts of 'adaptation to intensity' desensitization effects have been shown in animals – e.g. pigeons may be deterred from seeking a reward by a shock of 80 volts if that is the first shock that is ever administered but if they are first administered a sub-deterring jolt of 60 volts, they may seek the reward even when shocks go as high as 300 volts. As I noted a bit earlier, our most conventional punishment practices – very low punishment for very high numbers of first time offenders, particularly youthful offenders – may only serve to make actors indifferent to a broader range of punishments, even when they would otherwise find somewhat longer punishment noticeably different and worse.

Researchers like Kahneman associated with the H&B tradition have expressed a certain degree of skepticism about whether hedonic adaptation is "real" or merely a reporting artifact. The view that it is a reporting artifact is a complex one that I will explore at some length, but at core, the idea is that (seeming) hedonic adaptation actually resembles typical cognitive heuristics. There is a true variable we seek to identify – e.g. the probability of events in the prototypical H&B scenario, one's "true" hedonic state in this case – and that one instead substitutes more readily cognitively accessible attributes for the hard-to-discern real attribute. So the (proto)typical H&B subject tries to identify whether more words end with "-n-" or "ing", but answers illogically that the second ending is more common because words with that ending are more available and she substitutes easily-made judgments of availability for a difficult, reasoned approach to assessing probability. The person seeking to know how she reacted hedonically to her time in prison cannot readily summarize her experiences: She thus substitutes various "happiness reporting" conventions. For instance, the person who might appear to adapt hedonically may merely figure out if her most recent or salient experience is better or worse than she expected -- and expectations are readily established by experience – and declare herself happy if they meet or exceed these diminished expectations.

This is not an easy argument to sustain, though it may well be at least modestly persuasive. It is easy to know what Kahneman means when he says that people misestimate the number of words ending in distinct letter combinations or misestimate the probabilities of available and unavailable forms of accidental death. Moreover, it is important to note that there *are* situations in which *reported happiness* is plainly influenced by factors that could not possibly impact experienced happiness: in those cases, the traditional heuristics and biases account seems most persuasive. People *mean* to report how happy they have been over some extended period but are heavily influenced by extremely temporary states that are much more readily recalled – e.g. whether they have recently found a dime, whether the weather is good. Or their responses are highly sensitive to the way in which questions are framed or whether they are asked to construct negative or positive counterfactuals before answering questions about their hedonic state.[22] But it is not nearly so clear what Kahneman means when he says that those who appear to hedonically adapt have merely changed their aspiration levels, that they have not changed their hedonic experiences: How do we know that "true" happiness

[22] There are a number of discussions of this sort of transparent misreporting. For a summary, with references to the primary literature, see Mark Kelman, "Hedonic Psychology, Political Theory, and Law: Is Welfarism Possible? 52 *Buffalo L. Rev.* 1, 19-27 (2004)

is not happiness relative to (some sort of) expectations? What does it mean to have a true hedonic experience that we not only do not, but seemingly cannot, know?

It does mean something to Kahneman: To him, hedonic reactions to life events are essentially involuntary binary responses, with a natural zero point. People either like the situation they are in (and wish it would persist), dislike it (and wish it would end) or are indifferent. Even if they cannot readily make cardinal judgments about precisely how good or bad an end-state is, they *cannot help but make* this sort of good/bad/indifferent judgment. What Kahneman is claiming is that long-term prisoners continue to find their daily situation just as full of aversive situations as the newcomers do, but simply misreport that fact because they are able only to *report* satisfaction relative to ever-shifting expectations. (Whether this is consistent with robust findings that almost all prison suicides occur in the first few days of imprisonment[23] is hardly clear, but for now, I am assessing the theoretical point more than the empirical one.)

[There is then a fairly long, theoretical discussion of Kahenman's distinctions between satisfaction treadmills and reporting treadmills, as well as a discussion of the distinction between adaptation and developing a "broadened" indifference point because of desensitization. It is of considerable interest to me, but may well drop out of the ultimate text because I have yet to find another sentient being who finds the discussion of any use or moment, even using the rather forgiving standards I use for judging the academic arguments I make.]

Whether hedonic adaptation is best described as diminishing something we would consider "real" negative hedonic reactions to long-term imprisonment or merely

[23] See L.M. Hayes, "'And Darkness Closed In': A National Study of Jail Suicides," 10 *Criminal Justice and Behavior* 461 (1983).

"reported" negative reactions, it is important to recognize, when thinking about the *ex ante* impact of future punishment, that would-be violators will almost surely discount the actual level of hedonic adaptation. When making affective *forecasts*, subjects typically underestimate the degree to which they will get used to bad situations and stop appreciating good ones (whatever it means to "get used to" such situations). People naively believe they will be persistently happier if they get tenure or win a prize or persistently unhappier if they endure a tragedy (even losing a child or being sent to a concentration camps) than they have been in the past when they have had positive or negative experiences. Broadly speaking, H&B researchers who have been especially interested in the failure to predict adaptation think that most people remain "relatively happy" most of the time and external events do little to change that for long. They think that subjects overestimate the durability of the affective shifts that accompany negative events for six broad reasons: First, they often misconstrue the events, unable to imagine them with an adequate degree of specificity (e.g. they think broadly about "going blind" but cannot imagine that they will have gone blind in a particular context, e.g. slowly as a result of a congenital disease or as part of a heroic effort to save a child). Second, they embrace inaccurate general theories about the sources of happiness and attach them to their hedonic forecasts (e.g. they wrongly think that "money is the key to happiness" and prospectively overestimate the impact of financial success or failure). Third, they are "defensively pessimistic" about bad events so that real life will be better than they expected. Fourth, because they can most readily imagine the strong immediate reaction they will indeed have to a negative event, and then anchor their estimates of the hedonic impact of the event to the anchoring point, they overestimate long-term consequences.

Fifth, undue "focalism" – the tendency to forget that many other things, good and bad will happen after the salient event on which they are focused – tends to distort one's impression of the impact the salient event will have on overall happiness (even if one wins the lottery, your kids might get sick; even if one gets in an accident, your kids may marry really great people). Finally, and perhaps most significantly, there is a residual sort of "immune neglect", a tendency to underestimate the degree to which organisms do not maintain a perpetually gloomy state once they recognize that the gloom cannot produce action that will alter the environment.[24]

Hedonic adaptation may be relevant to punishment practices in a further way. There is substantial reason to believe that people may adapt better (and hence "feel" better, at least in some meaningful sense) when they are *sure* something bad has happened than when they're 95% sure but still can imagine that there might be a way out. In this regard, those who receive bad HIV reports or a firm diagnosis that they have Huntington's disease seem better off than those not yet informed, but highly suspicious, of their status.[25]

This may seem counterintuitive at first glance. Under one view, of course, the person with a 100% chance of something bad happening is just like the person with a 95% chance except that as to the last 5%, he is in worse shape. One way of putting that is that state X is being 95% certain that one is HIV positive, and that state X is present both

[24] For a far fuller account of this "durability bias" – including experimental studies that both demonstrate its general existence and attempt to demonstrate which of the causal explanations for the tendency to discount the "return to equilibrium" are most invariably operative – see Daniel T. Gilbert, Elizabeth C. Pinel, Timothy D. Wilson, Stephen J. Blumberg, and Thalia P. Wheatley, "Immune Neglect: A source of durability bias in affective forecasting," 75 *J. of Personality and Social Psych.* 617 (1998).

[25] See Jason Brandt et. al. "Presymptomatic Diagnosis of Delayed-Onset Disease with Linked DNA Markers: The Experience in Huntington's Disease," 261 *J. Am. Med. Ass'n.* 3108 (1989) and Jeffrey M. Moulton et. al., "Results of a One Year Longitudinal Study of HIV Notification from the San Francisco General Hospital Cohort," 4 *J. Acquired Immunity Deficiency Syndromes* 787 (1991).

in those who are certain that they are HIV positive and those who are not yet informed but are fairly sure that they are HIV positive. If state Q is “a lottery ticket with a 5% chance that I learn information about my HIV status and the information that I learn is that I am HIV positive” and state R is “a lottery ticket with a 5% chance that I learn about my HIV status and the information that I learn is that I am HIV negative”, one plainly prefers R to Q, but one may not in fact prefer the sum of X and R to X and Q. Alternatively, one could frame this point by noting that it is clear that the subject prefers state E, HIV- status, to D, HIV+ status: but that ordinarily implies that the expected value of a 95% chance of D plus a 5% chance of E is valued more than a 100% chance of D.

The fact that this set of preferences seemingly violates ordinary rationality conventions can be seen if we imagine the state E being “winning” $1 million (parallel to HIV negative status) and state D being winning only $100 (the lower valued HIV+ status). It would plainly be better to have a 95% chance of D and a 5% chance of E than a 100% chance of D, and we would think we would be able to infer that from the fact that E is preferred to D. The conventional rationality principle that those who prefer “certain” bad news violate is Savage’s “sure thing” principle – the principle that if one is offered a lottery X and lottery Y which differ only with respect to the fact that X contains prize A as one prize and Y contains B, that one rationally “must” prefer X to Y if one prefers A to B. …[There is then a discussion of why the sure thing principle may fail here given the presence of an “integrating agent.”]

However one explains the hedonic reaction to uncertainty reduction, findings like these suggest, counter to the traditional literature on punishment certainty, that uncertain (but highly likely) punishments may generate more disutility (and therefore deterrence punch) than fully certain punishments. The existing death penalty system – in which a sub-set of Death Row inmates is nearly, but never truly certain, that they will be executed, might (however inadvertently) both be the most punitive system we could establish (if one wanted to increase the retributive punch of punishment) and, to the degree that people pre-cognized this sort of punishment, might most thoroughly deter the death-penalty eligible homicides. Auditing systems with long Statutes of Limitations might best deter certain forms of violation since they both make people feel pretty sure that they will be punished and anxious that they will, rather than allowing them to adapt to the reality that they will indeed be punished.

Fourth, and finally, H&B theorists posit that the duration of a bad (or good) event has far less impact on its *perceived* hedonic quality than one would expect. H&B-influenced researchers plainly treat duration neglect as an unambiguous reporting or measurement error, grounded in simple attribution substitution. It is difficult to recall and sum all the pains and pleasures one felt over any substantial period (whether as short as a colonoscopy or as long as a prison term), so people typically substitute the average of the peak pain and the end-point pain for a (cognitively unavailable) true summation. In the colonoscopy context, experimental subjects in Group A receive a regular half hour colonoscopy, reporting, say, a peak pain of 8 on a ten point scale and a final pain of four when the instrument is removed at the end of the procedure. For the next fifteen minutes, after the instrument is removed, the pain level is zero. They believe that the procedure as a whole produced a pain level of 6 (the average of the peak and end). Those in Group B receive the same regular procedure but receive 15 more minutes of moderate pain (the instrument is left in though it does not continue to probe): they report a peak pain level of 8 and a final pain level of 2 (the pain level when the instrument is left in.) Their peak/end report is that the procedure caused a pain level of only 5. Now, both people in groups A and those in group B experience identical experiences for the first half hour; those in Group B have a worse time of it for the next fifteen minutes (2 level, not 0 level pain). But those in group B report less pain *and* are more likely to show up for their next colonoscopy appointment.[26]

Kahneman is confident that this sort of duration neglect is a bias, a mere cognitive "reporting error". I have expressed my hesitations about that view in the past, and do not think it is especially important for these purposes to decide whether "peak/end" reporters are mistaken (as he believes) or expressing a more integrated view of life satisfaction and dissatisfaction, or engaged in construing their life-states by giving "meaning" to narrative events rather than thinking of themselves as dissolved mini-persons whose experience is the sum of the mini-people's experiences. But what is important for now is how it might impact punishment policy if we believed subjects typically neglected the duration of bad feelings when they considered the hedonic quality of punishments.

The standard H&B-influenced story would go something like this: People will not think a ten year prison term *was* worse than a six year term merely by virtue of its length. To the degree that specific deterrence works through memory of the hedonic quality of past punishment, then, a criminal who was previously punished for ten years will be no more likely to be specifically deterred than one punished for six. General deterrence effects will depend on the degree to which those who have been punished communicate

[26] See Daniel A. Redelmeier & Daniel Kahneman, "Patients' Memories of Painful Medical Treatments: Real Time and Retrospective Evaluations of Two Minimally Invasive Procedures," 66 *Pain* 3 (1996).

to would-be violators how they experienced their punishment. Worse still, from the vantage point of those who believe they can deter more crime by increasing prison terms, increasing terms may be counterproductive. *If* people indeed hedonically adapt to prison, those who serve longer terms will experience a more favorable "end" hedonic state. Assuming their peak state is no worse (though this may not be true if peak states worsen for those who enter prison anticipating a long and miserable term), their final evaluation of prison will be more favorable if they evaluate by averaging the peak and end state.[27]

c. *Systematic errors in providing appropriate deterrence signals*

I will be especially brief, and merely suggestive, in noting that policymakers influenced by H&B theory will tend to believe that the price signals that the legal system will often generate may reflect cognitive biases. [There is a discussion of the fact that we will over-deter reasonable behavior because hindsight bias may induce us to mischaracterize behavior that was reasonable *ex ante* as unreasonable *ex post*, and a brief discussion of the fact that legislators may overestimate the harms associated with salient risks. This second discussion harkens back to a discussion in an earlier chapter of the "moral heuristics"]

B. [I omit the section on retributivism and heuristics]

C. A Brief Note on Incapacitation

Obviously, some policy makers believe that the primary (or at least a substantial) reason we ought to punish (at least some subset of) offenders is that we can prevent them from offending (at least against the non-incarcerated population) while they are imprisoned. The question for policy makers interested in incapacitation is whether the criminal justice

[27] This picture of the counterproductive impact of longer terms is suggested in Robinson and Darley, supra note – at 189-93

system can distinguish people based on how dangerous they are (how likely it is that they will offend, and for what period they are likely to offend if left free to do so.)

F&F theorists are likely to believe that even if we are incapacitationists, we should use the conventional criminal law system to determine dangerousness levels. A single cue – did the defendant commit a certain form of offense that has been labeled in terms of its "severity" – should be sufficient to make adequate predictions of dangerousness. Any efforts by sentencing "experts" to predict dangerousness using multi-cue statistical methods are likely to be less accurate than decision rules grounded in one or a few cues. Thus, (pseudo)-sophisticated incapacitationists might be tempted to use regression equations or similar tools that attempt to predict how likely a person would be to recidivate if set free given a range of variables that include, but are not exhausted, by the crime he is convicted of but may include other factors (like job history, family background, scores on a variety of cognitive and affective psychological tests, an inventory of current relationships, drug or alcohol use/abuse etc.) But these tools are (presumptively?) less accurate than the use of fast and frugal heuristics.

Of course, F&F theory does not establish meaningful, non-tautological boundary conditions to specify when these sorts of heuristics are particularly valid. (Statements of the form, "the heuristic s work especially well when there is a high correlation between the value of the single cue and the outcome variable of interest" are tautological, not useful.) So we are left with a general "predisposition" (best thought of in terms of the "sociology of knowledge") among F&F theorist to believe in "less is more" effects.

Still, in studying bail granting practices of English magistrates (that closely resemble incapacitation-oriented sentencing decisions), F&F-influenced researchers indeed found that as a *descriptive* matter, the judges actually used fewer cues in deciding whether or not to grant conditional bail (involving curfews or some level of pre-trial detention) than they said they used. The two cues they used were simple: they essentially "passed the buck" and looked to see only if the prosecutors recommended conditional bail and whether conditions had been placed on the prisoner in past cases. But normatively, they did not assume (or find) that it was better to use the few cues they seemed to employ, and, quite to the contrary, seemed to worry that the exclusion of further cues might compromise legitimate Due Process goals. The case is complex, though, in terms of evaluating the utility of fast and frugal heuristics because to the degree they find "failure" it appears to be due more to a principal/agent problem than the intrinsic inadequacy of the heuristics. The magistrates do not share the principals' goal – to make "just" bail decisions – but seek a distinct selfish goal – to protect themselves from criticism. The "pass the buck" heuristics may indeed meet *that* goal quite efficaciously.[28]

Those who think about how decision-making may be biased are likely to worry that the attributes that make us think that a convicted offender (or, in preventive detention terms, a mere potential offender) are likely to be dangerous unless isolated are not enormously probative of true risk. Our perceptions of risk would be biased, largely in ways that would make us unduly punitive.

[28] For a fuller discussion, see Gerd Gigerenzer, "Heuristics," in *Heuristics and the Law* 17,28-30, 40 (Gerd Gigerenzer and Christoph Engel eds. 2006) One sees the "predisposition" point especially clearly in Ralph Hetwig, "Do Legal rules Rule Behavior?" in *Heuristics and the Law*391,407 (Gerd Gigerenzer and Chritoph Engel eds. 2006) (("Although [the magistrate's] approach deviates from the ideal of due process, it is impossible to find out how accurate the decision tree is. Judging from the good performance of other 'fast and frugal' decision-making heuristics, however, it may not result in less accurate judgments than due process."

Those influenced by H&B theory would expect, first, that decision makers would over-predict dangerousness: past cases of false negative predictions (situations in which a dangerous person had been released and committed a crime, especially a brutal crime) would be highly salient and available while false positives (in which people were needlessly isolated, though they would not have committed a crime had they been free) are not saliently observable by direct observation but only known through dry, statistical analysis. If it is right that we overestimate the objective probability of events that are readily recalled – in this case, the false negatives – we will tend to think we have to perpetually increase detention rates because we have always been underestimating the dangerousness of those we release.

Similarly, it is likely the case that the perpetually dangerous criminal is more "representative" of (at least) violent convicts: Once more, we will overestimate the probability that representative traits are (more) common traits than is actually the case. Of course, predictions about how decision makers will assess the degree to which one instantiation of a category is more prototypical of the category are not especially well-specified in H&B thought, but it appears at least plausible that an H&B-influenced criminal policy-maker would worry that dangerousness would be systematically overstated for this reason as well.

Anchoring effects would cut in the same direction, at least in thinking about incapacitating convicted criminals rather than in thinking about (purer) preventive detention. What we know about a convicted criminal is that she did offend during the period just prior to our sentencing decision: our baseline probability that she is someone who offends is 100%. Whether we adequately adjust in predicting what she will do in the future from the anchored starting point – she is a dead certain offender – is dubious if the general studies on anchoring are correct. Again, "anchoring theory" is not so well-specified that we can be sure that the anchor a decision maker will use is "the probability that the defendant offended recently", but it is a bit hard to imagine another anchor as plausible as that for convicted criminals. (I think specifying the likely anchors in purer preventive detention cases would be considerably more difficult: I doubt that people would anchor to the global rate of offending, "uncorrected" for predispositions about the offense rates of individuals with distinct demographic markers).

Finally, it may well be the case that decision makers engage in the sort of base rate neglect that H&B theorists frequently highlight when they try to figure out how likely a criminal is to commit a more serious crime than the one she was convicted of. Imagine (counterfactually, but it will only strengthen the case for incapacitationist punishment if this is true) that we know that every single person who will kill in the next year has committed a crime in the past (i.e. murder is *never* a first offense.) At the same time, we know that a number of offenders will not kill and that the base rate for murder is very low. Think of the parallel case I noted in chapter ---. all people who are truly HIV+ will test positive for the presence of HIV, just as all people who will kill will "test positive" for having committed a crime. But if the base rate for murder is, say, one in ten thousand -- just as the base rate for HIV infection in the low-risk population is that low – and 1 in 10,000 people who will not murder – like one in ten thousand people who are actually HIV negative -- will "test" positive (have committed crimes.) Just as those who neglect base rates will falsely believe that a person who has received a positive HIV test is likely to be positive since all people who are positive test positive, so will people systematically tend to believe that all who have committed crimes are more likely to kill than they are (since all who kill have indeed committed crimes.) Obviously, the false positive rate for killing among criminals is actually quite high, but if that rate is overlooked (in favor of a focus on the low false negative rate among killers), then the actual aggregate risk that criminals pose will be misunderstood….