Law and Philosophy (2005) 24: 393–434
DOI 10.1007/s10982-004-8197-3

J.D. TROUT

PATERNALISM AND COGNITIVE BIAS

I. INTRODUCTION: COSTLY BIASES

As clever and enterprising as humans sometimes are, we are also hobbled by a host of systematic and psychologically stubborn biases of reason and emotion. And these biases persist even when there are no limits on the evidence made available to the decision-maker.[1] We severely underestimate our health risks, from HIV to heart disease and cancer, and so don't take adequate precautions (the "optimistic bias"). We discount the future value of resources, and so radically undersave for a variety of important and foreseeable prospects, ranging from the costs of college education and health care to retirement (the "discounting bias"). These biases of reason and emotion are in no way exotic; they afflict normal people under normal stresses. Their effects are both routine and expensive. Because they are allowed to go uncorrected, people unnecessarily suffer disease and poverty. The cost to each individual is clear. Equally clear is that, had they known that their reasoning were so unreliable, they would have chosen strategies to counteract this sorry decision-making performance. But these decisions are not just personally costly; they are socially costly as well. The optimistic bias and irrational discounting cost our health care and social welfare systems billions of dollars annually in uninsured treatment and indigent services.

I will argue that institutional assistance can improve judgment, and that these improvements can promote our auton-

[1] See M. Bishop and J. D. Trout, *Epistemology and the Psychology of Human Judgment* (New York: Oxford University Press, 2005) for a systematic response to the empirical literature on judgment and decision-making, and a positive theory for evaluating the quality of reasoning.

omy, avoiding costly outcomes when the biases go unchecked.[2] The last 50 years of research on human judgment supplies the knowledge to design and implement such institutional prosthetics. I will examine three such examples of institutional prosthetics in section V. The proposal to institutionally assist human judgment in defined settings of health and financial risk has precedent, and can be found in FDA regulations that prohibit false and misleading drug advertisements, and the SEC's constraints on the forward-looking statements of brokerages and mutual fund companies. In these cases, much of the crucial information is theoretically arcane. Is it possible for a normal adult to identify misleading ads, and correct their attraction to the product? Is it possible for a normal adult to ferret out false drug claims or recognize false or irresponsible brokerage guarantees of future returns? Perhaps, with enough time and training, initiates could do so. But most such policies, designed to "protect the public" in a market economy, do not count on mere possibilities. A premium on autonomy should not turn all citizens into vigilant researchers, ever on the lookout for false claims that require technical sophistication to identify.

This paper makes the case for the legitimacy of governmental regulation on behalf of a person's good for selected classes of cognitive bias. Regulation can be permissible even when it runs counter to that person's spontaneous wishes, particularly when the regulation advances the agent's considered judgments or implicit long-term goals. The biases undermine people's welfare in well-defined and understood contexts. When our best psychological science has identified the conditions under which an individual's spontaneous judgment is systematically and seriously compromised, and individuals do not want to be so compromised, they can arrange (or endorse the arrangement of) institutional structures

[2] J. Rachlinski 'The Uncertain Psychological Case for Paternalism', *Northwestern University Law Review* 97 (2003), pp. 1165–1225, correctly notes that much of the literature on both theory and policy is quick to infer the need for institutional correctives from the existence of cognitive biases. The present paper evaluates the quality of some of the reasons for this inference.

required to assist in achieving that person's or a group's ultimate ends. But there are better and worse ways of interfering. When a person's spontaneous decisions are inconsistent with her long-term goals, institutional assistance should be as unimposing as possible. This is the advantage of bias-harnessing methods, methods that actually *use* a bias to advance that person's chosen ends. I will offer two examples of such methods, designed to reduce the risk of financial hardship in retirement, and to reduce health risk through screening. Institutional assistance for decision-making imposes no such restrictions on our actions or knowledge in other walks of life. And, because individuals are unable to counteract the ill effects of our heuristics in reasoning, and their effects are systematic, the biases have exactly the properties necessary for efficient institutional treatment.

II. THE BIASES: HOW BAD ARE THEY?

The Enlightenment philosopher Condorcet judged that, "no bounds have been fixed to the improvement of the human faculties; that the perfectibility of man is absolutely indefinite…."[3] Contemporary cognitive science treats this Enlightenment opinion as a potentially damaging, if quaint, expression of hope. After more than 50 years of systematic experimental research on the nature of rational judgment, researchers summarize the consensus that the empirical results of the heuristics and biases program have "bleak implications" for human rationality[4] and that "individuals are generally affected by systematic deviations from rationality."[5] The causes of these biases are broad and deep, and so their correction would re-

[3] J. de Condorcet, *Sketch for a Historical Picture of the Progress of the Human Mind* (London: Weidenfeld and Nicolson, 1792/1955), Chapter 24.

[4] R. E. Nisbett and E. Borgida, 'Attribution and the Psychology of Prediction', *Journal of Personality and Social Psychology* 32 (1975), p. 935.

[5] M. H. Bazerman and M. A. Neale, 'Heuristics in Negotiation: Limitations to Effective Dispute Resolution', in H. R. Arkes and R. R. Hammond (eds.), *Judgment and Decision Making: An Interdisciplinary Reader* (Cambridge: Cambridge University Press, 1986), p. 317.

quire measures that are, potentially, sweeping and significant. While the brief review that follows cannot do justice to the expansive scope and everyday regularity of these decision-making traps, it will make clear that they produce far more than minor annoyances or containable predicaments; their effects are potentially very costly, in terms of both economic currency and human prospects.

In the 1970s, a bias of availability swept the nation. The media coverage of alar, a chemical preservative sprayed on apples, produced a scare that made this risk more psychologically salient to an anxious public than far more serious ones.[6] As it turns out, the alar scare was without basis, but still cost millions in regulation and its effects. A litany of other examples should demonstrate just how pervasive the cognitive biases are. People are chronically overconfident in judging the probability of their correctness on factual questions. They are powerfully prone to preferring one option over another when both have equal expected value, favoring options framed as gains rather than losses. People anchor on a mentioned value, even when they know it to be irrelevant to the desired calculation. People are self-serving in their interpretation of their own behavior, so much so that it is difficult to find anyone who reports being below average along any desirable dimension. To make matters worse, we suffer from a kind of "status quo bias" best represented by "the endowment effect": people shoulder larger risks to preserve the status quo than they would to obtain the item or achieve the goal in the first place.

A thorough picture of the biases, and so a complete basis for the institutional assistance of decision-making, requires a fuller presentation of their dispositional origin. We will examine two contexts in which institutional assistance in decision-making significantly improves people's welfare, but presenting isolated cases may convey the impression that contexts alone cause the biases.

Our review of the biases will show that they are virtually as stable, durable, and universal as reflexes. And like reflexes,

[6] T. Kuran and C. Sunstein, 'Availability Cascades and Risk Regulation', *Stanford Law Review* 51 (1999), pp. 704–768.

their effects can be anticipated and often counteracted. Our best psychological theories provide a detailed understanding of the mechanisms that produce these biases and the environments that trigger them. Forty years of empirical research converging on a single moral – that the Enlightenment vision is profoundly mistaken – is bound to have important practical consequences. The most important practical consequence of the biases is that they frustrate our efforts to achieve our goals.

Poor planning catches up with people in all endeavors. Academics are all-too familiar with the experience of agreeing, on several independent occasions, to write articles for solicited projects. On each occasion of agreement, the task seems easily achievable, so we say "yes" to the request, and then find ourselves surprised to be swamped with obligations. The impulse is certainly admirable, and the disposition to take on such responsibilities speaks of a robust sense of effectiveness. But missing deadlines is normally both socially and personally unacceptable – and in business and government settings, extremely costly – and the best hope of preventing it begins with an understanding of the steps that lead to this embarrassment.

In 'Intuitive prediction: Biases and corrective procedures', Kahneman and Tversky[7] describe the planning fallacy:

> The planning fallacy is a consequence of the tendency to neglect distributional data and to adopt what may be termed an internal approach to prediction, in which one focuses on the constituents of the specific problem rather than on the distribution of outcomes in similar cases. The internal approach to the evaluation of plans is likely to produce underestimation. A building can only be completed on time, for example, if there are no delays in the delivery of materials, no strikes, no unusual weather conditions, and so on. Although each of these disturbances is unlikely, the probability that at least one of them will occur may be substantial... This combinatorial consideration, however, is not adequately represented in people's intuitions (Bar-Hillel 1973). Attempts to combat this error by adding a slippage factor

[7] D. Kahneman and A. Tversky, 'Intuitive Prediction: Biases and Corrective Procedures', in D. Kahneman, P. Slovic and A. Tversky (eds.), *Judgment Under Uncertainty: Heuristics and Biases* (New York: Cambridge University Press, 1979), pp. 414–421.

are rarely adequate, since the adjusted value tends to remain too close to the initial value that acts as an anchor.[8]

So, if the goal is to not overextend yourself, the trick is to assemble distributional information about your current and past projects. Kahneman and Tversky suggest a 5-step procedure,[9] which I will apply to a recognizable academic example: (1) look at your CV, (2) select the reference class (if the solicitation is for an article, then articles are your reference class), (3) count the number of articles you were able to complete over a representative period of time (perhaps 5 years or so), (4) ask whether there is anything unusual about the article solicited that would make it unrepresentative of those on your CV. Is it longer or shorter? Would it be more or less closely related to the topics you have been working on? (5) predict the likelihood of completing the project by the deadline. Self-serving biases are seductive here.[10] We often tell ourselves that there are special

[8] Kahneman and Tversky (1979, p. 415).

[9] In the balance of Kahneman and Tversky (1979), they offer a five steps corrective procedure that recognizes that, "[t]he prevalent tendency to underweigh or ignore distributional information is perhaps the major error of intuitive prediction" (p. 416). This five-step exercise is as follows:

(1) Selection of a reference class. Identify the class to which this case belongs, a class for which there is known distributional information.
(2) Assessment of the distribution of the reference class. Determine the relative frequency of these kinds of cases for this class.
(3) Intuitive estimation. Based on specific information about the particular case of interest, ask how this case differs from other members of this class. This exercise should allow you to assess how well this case can serve as a basis for accurate prediction of outcomes.
(4) Assessment of predictability. Without the product-moment correlation between predictions and outcomes, one must rely on subjective estimates of prediction-outcome correlation.
(5) Correction of the intuitive estimate. Use frequency information to correct any incorrect subjective estimates that might misrepresent base-rate information. The biggest threats here are the hindsight bias and overconfidence bias. The goal is to reduce the distance between the intuitive estimate and the average.

[10] T. Gilovich, *How We Know What Isn't So* (New York: The Free Press, 1991).

reasons to think that we could meet the deadline *in this particular case*, even though the number of projects currently on our plate is twice the 5-year average (e.g., in the earlier period, we had a young child, a heavier teaching load, an ugly divorce, etc.). We either discount, or fail to consider, the *efficiencies* we enjoyed during that time as well, and along with them, the fact that those efficiencies may no longer obtain. All of the other biases have in common with the planning fallacy their unreliable intuitive basis and their ability to be corrected by a procedure that focuses on distributional information. But in many cases providing the relevant distributional information is a job of institutional proportion.

A. *The Availability Bias*

The availability heuristic is an implicit cognitive rule that inclines us to infer the representativeness of an event from the ease with which it can be recalled or visualized. A particular kind of event may be rendered unavailable (or less available) if it is difficult to generate instances of this kind of event. For example, Tversky and Kahneman[11] asked participants if, in a representative body of English language text, there more words that begin with 'k' or have 'k' in the third position. About 69% said that words beginning with 'k' are more common. As a matter of fact, words with 'k' in third position are about twice as probable. Because it is easier to generate instances with 'k' in first rather than third position, people tend to overestimate the frequency of 'k'-initial words.

Another source of availability bias is the jarring or unusual character of an event. News and gossip items not only produce vivid images, but keep them in public consciousness. School shootings are big news items but, for example, more people died from lightning strikes in 1998 – a representative year for lightning fatalities (the mean is about 90 fatalities per year in the U.S. between 1959 and 1994) – than from school shootings

[11] D. Kahneman and A. Tversky, 'On the Psychology of Prediction', *Psychological Review* 80 (1973), pp. 237–251, reprinted in D. Kahneman, P. Slovic and A. Tversky (eds.), *Judgement Under Uncertainty: Heuristics and Biases* (Cambridge: Cambridge University Press, 1982), pp. 48–68.

in any year in the U.S. (the mean is about 2.5 fatalities per year from 1979 to 2002, and the worst year for fatal school shootings was 1999, with 16).[12] The real threat here, of course, is that the information that we receive and that we attend to is not an inaccurate representation of the actual risk. In particular, an impression of risk that prompts hysteria or other undue concern could place demands on resources that then 'crowd out' other worthy concerns.

B. *The Overconfidence Bias*

The overconfidence bias is one of the most robust findings in contemporary psychology. Fischhoff, Slovic and Lichtenstein[13] asked subjects to indicate the most frequent cause of death in the U.S., and to estimate their confidence that their choice was correct (in terms of 'odds'). When subjects set the odds of their answer's correctness at 100:1, they were correct only 73% of the time. Remarkably, even when they were so certain as to set the odds between 10,000:1 and 1,000,000:1, they were correct only between 85% and 90% of the time. It is important to note that, like the other biases, the overconfidence effect is systematic (it is highly replicable and survives changes in task and setting) and directional (the effect is overwhelmingly in the direction of over rather than under-confidence). These judgments are utterly

[12] www.ribbonofpromise.org/stats and www.nssl.noaa.gov/papers/techmemos/NWS-SR-193/techmemo-sr193.html; last accessed on January 22, 2004. The social importance of an event is not mitigated by its low frequency, of course. So there are other reasons that school shootings may be psychologically prominent or available despite their low relative frequency (when compared to, say, fatal lightning strikes). For example, we assign a higher cost to the death of a young person than an older person, and school shooting fatalities are typically young, while lightning strikes do not select for age. In addition, lightning strikes seem uncontrollable, so the resulting fatalities appear practically unavoidable, whereas school shootings seem avoidable when appropriate precautions are taken. But availability can certainly be one reason an event is seen as higher in frequency than it actually is.

[13] B. Fischhoff, P. Slovic, and S. Lichtenstein, 'Knowing with Certainty: The Appropriateness of Extreme Confidence', *Journal of Experimental Psychology: Human Perception & Performance* 3(4) (November 1977), pp. 552–564.

representative of those made in medical care, financial services, and a host of other settings of 'expert' decision-making.[14]

C. *Hindsight Bias*

People are notably unaware of the influence that outcome information has on them. This is precisely the retrodictive epistemic position of the explainer. The explainer says, after the fact, how the causes brought about an effect. The traditional manner of establishing the hindsight bias begins by asking subjects to estimate the likelihood of various outcomes of an upcoming event, and then retesting them after the event, asking them to recall how likely they had found each of the possible outcomes the first time around. Fischhoff and Beyth[15] did just that in an early study of the hindsight bias. Prior to President Nixon's trip to China and the Soviet Union in 1972, subjects were asked how likely they found a variety of possible outcomes (e. g., whether Nixon would meet Mao, that the Soviet Union and U.S. would establish a joint space program, etc.). Two weeks to six months after the trip the subjects were asked to fill out the same questionnaire. They were asked to recall the probabilities they assigned initially to the same events and, if they couldn't recall, to assign the probability they would have assigned immediately before Nixon's trip. They were also asked if each of the listed outcomes had, in fact, occurred.

The results were a striking demonstration of the distorting influence of hindsight. For those outcomes that subjects thought had occurred, they remembered their estimates as more

[14] What about scientists? Surely scientists' training and experience delivers them from the overconfidence bias in their areas of expertise. Alas, no–or at least, not always, Physicists, economists, and demographers have all been observed to suffer from the overconfidence bias, even when reasoning about the content of their special discipline. See M. Henrion and B. Fischoff, 'Assessing Uncertainty in Physical Constants', *American Journal of Physics* 54 (1986), pp. 791–798. It would appear that scientists, too, place more faith in the subjective trappings of judgment than is warranted. Philosophers have supported this bad habit.

[15] B. Fischhoff and R. Beyth, "'I Knew it Would Happen'': Remembered Probabilities of Once-future Things', *Organizational Behavior & Human Decision Processes* 31(1) (February 1975), pp. 1–16.

accurate than they in fact were. For those outcomes thought not to have occurred, subjects recalled their estimates as having been lower than they in fact were. The effect seems to strengthen with the passage of time. After three to six months, 84% of the subjects displayed hindsight biases. Therefore, after learning the results of Nixon's trip, subjects believed the outcomes were more predictable that they actually were.

We conceptualize the event as inevitable, and thus people tend to say that the event was fairly predictable all along. Thus, the hindsight bias is also known as the "I-knew-it-all-along effect". In particular, people tend to overestimate how probable they thought the event was before it occurred. Without proper acknowledgment of our error, the hindsight bias removes the incentive to respond to the data with humility and deference, and so to approach evidence with critical scrutiny.

Scientific research on the frailties of human judgment have yet had virtually no influence on those aspects of the judicial and legislative branches of the U.S. government's policies designed to improve human welfare. These include shaping attitudes toward poverty, the importance of lower and middle-income people to save, the threat of racism, etc. Despite little attention to this judgment research in public policy, another government agency, the CIA, was quick to see the importance of this research, in this case, for the shaping of public opinion and deflection of blame for poor intelligence. In recently published internal papers, a 1978 document analyzes the first studies on the hindsight bias (which received funding from the Defense Department):

> These results indicate that overseers conducting postmortem evaluations of what analysts should have been able to foresee, given the available information, will tend to perceive the outcome of that situation as having been more predictable than was, in fact, the case. Because they are unable to reconstruct a state of mind that views the situation only with foresight, not hindsight, overseers will tend to be more critical of intelligence performance than is warranted.[16]

[16] CIA report, 1999, p. 7; http://www.cia.gov/csi/books/19104/art16.html; last accessed on January 22, 2004.

The CIA may have been interested in the hindsight bias, but the U.S. court system has been far less interested and vigilant. More than twenty years after this data was published there was still no systematic response by the courts, even after a number of hindsight studies were done in legal contexts in which liability is assessed and damages are awarded.[17]

D. *Self-serving Biases and Overconfidence in the Reliability of Our Subjective Reasoning Faculties*

Humans are naturally disposed to exaggerate the powers of their subjective or intuitive assessments of evidence. A very prominent example of this is the interview effect. When gatekeepers (e.g., hiring and admissions officers, parole boards, etc.) are allowed personal access to applicants in the form of unstructured interviews, they are still outperformed by simple prediction rules (based on demographic information) that take no account of the interviews. In fact, unstructured interviews actually *degrade* the reliability of human prediction.[18] That is, gatekeepers degrade the reliability of their predictions by availing themselves of unstructured interviews.

Highly educated people ignore the obvious practical implications of the interview effect. Even when they know that the interview information they are considering is not diagnostic, they cannot look away. They tell themselves that *this time* their spontaneous subjective estimate is more reliable than the proven strategy. The common arrogance of each individual's conviction that they are exceptional leads people to defect from the better strategy, piece by piece, and surrender their accuracy to, as George Eliot once put it, ''that pleasureless yielding to the small solicitations of circumstance, which is a commoner history of perdition than any single momentous bargain.''[19]

[17] For an excellent study of the hindsight bias in a legal context, see J. Rachlinski, 'A Positive Psychological Theory of Judging in Hindsight', *University of Chicago Law Review* 65 (1998), pp. 571–625. For a thorough post-mortem of this era of neglect, see C. Sunstein, R. Hastie, R. Payne, D. Schkade, and W. Viscusi, *Punitive Damages: How Juries Decide* (Chicago: University of Chicago Press, 2002), especially chapter 6.

[18] For citations, see Robyn Dawes, *House of Cards* (Free Press, 1994).

[19] George Eliot, *Middlemarch*, chapter 79.

The interview and related effects occur because we have unwarranted confidence in our subjective ability to 'read' people. We suppose that our insight into human nature is so powerful that we can plumb the depths of a human being in a 45 minute interview – unlike the lesser lights who were hoodwinked by disciplined attachment to the prediction rules. Our (over)confidence survives because we typically don't get complete feedback about the quality of our judgments (e.g., we can't compare the long-term outcomes of our actual decisions against the decisions we would have made if we hadn't interviewed the candidates).

E. *Framing Bias*

One of the most powerful influences prompting an availability bias is the phenomenon of 'framing'. Framing is the process whereby a problem is presented to an audience, preparing them to see a certain range of possible options, solutions, evidential bearing, and so on. The audience's intellectual habits and explanatory expectations allow carefully framed narrative descriptions to yield defective inductions. Framing typically gets the reader or listener to ignore important quantitative, sampling information (recall, this was the same downfall caused by the planning fallacy). A number of studies have shown that whether subjects find an option acceptable or not depends on how the alternatives are presented rather than on quantitative information that, on the typical paradigm of these studies, ensures equally probable alternatives.

The following passage is representative of a wide range of instances of framing: respondents in a telephone interview evaluated the fairness of an action described in the following vignette, which was presented in two versions that differed only in the bracketed clauses:

> A company is making a small profit. It is located in a community experiencing a recession with substantial unemployment [but no inflation/and inflation of 12 percent]. The company decides to [decrease wages and salaries 7 percent/increase salaries only 5 percent] this year.[20]

[20] D. Kahneman, J. Knetsch, and R. Thaler, 'Fairness as a Constraint on Profit Seeking: Entitlements in the Market', *American Economic Review* 76(4) (September 1986), 728–741, at p. 731.

Although the loss of real income is very similar in the two versions, the proportion of respondents who judged the action of the company "unfair" or "very unfair" was 62% for a nominal reduction but only 22% for a nominal increase.

The framing effect is an effect of formulating problems in different ways. Suppose you are asked to choose between two alternatives: (A) A sure gain of $240, or (B) A 25% chance to gain $1000, and a 75% chance to gain nothing. Would you choose A or B? As it happens, when gains are at stake people tend to be risk averse. In fact, 84% of participants chose A over B.

Now consider which of the following you would prefer: (C) A sure loss of $750, or (D) A 75% chance to lose $1000, and a 25% chance to lose nothing. In this case, would you choose C or D? Well, people tend to take risks when only losses are at stake, and 87% of the participants selected D.

These patterns are so robust that, if we consider the two choices as a pair, 73% chose A and D, and only 3% chose B and C. The interesting fact is, however, that B and C is a better pair than A and D. We can arrive at this conclusion by simply calculating the expected value of each pair.[21] As we will see in section V, the framing of options is crucial in allowing people to make choices that represent both their spontaneous choices and their implicit rationales, the tacit long-term objectives from which we so easily defect.

F. *Status Quo Bias*

The status quo bias causes people to assign a premium to existing courses of action over alternatives that they would otherwise agree have greater value. In the classic experiment of the "endowment effect" on which the status quo bias is based, mugs are given randomly to some people in a group. Those who now have them are asked to state a price to sell their mug; those without one are asked to name a price at which they will purchase one. Usually, the average sales price is substantially higher than the average offer price. Put in policy terms, people don't appreciate when an existing option is more costly than a

[21] A. Tversky and D. Kahneman, 'The Framing of Decisions and the Psychology of Choice', *Science* 211 (January 1981), pp. 453–458.

new option, because people rate existing or 'default' options more positively than alternatives; they favor the status quo. However, if we look only at costs and benefits, there is little reason to favor existing options without a calculation of the opportunity costs of maintaining the current arrangement. Further, the premium for the status quo is seldom explained by the aversion to either risk of the unknown or start-up costs.[22]

G. *Anchoring and Adjustment*

Background information plays an important role in generating accurate predictions or estimates. Anchoring is a process in which irrelevant background information fixes points of reference. In an experiment by Tversky and Kahneman,[23] a wheel is spun and, when the arrow stops on the number 65, participants are asked if the percentage of African countries in the United Nations is greater than or less than 65%. For another group of participants, the wheel is spun and stops at 10, at which time they are asked if the percentage of African countries in the United Nations is greater or less than 10%. Now, no one would have supposed that people's ultimate estimates would be sensitive to such irrelevant information. After all, the process of spinning a wheel could not have anything to do with the percentage of African countries in the United Nations. But this irrelevant information has a shockingly potent effect. In the 65% condition, the median estimate of African countries in the United Nations was 45%. In the 10% condition the median estimate was 25%.

The anchoring bias was examined in real-world settings as well. Gregory Northcraft and Margaret Neale[24] showed that

[22] D. Kahneman, J. Knetsch, and R. Thaler, 'Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias', *Journal of Economic Perspectives* 5(1) (1991), pp. 193–206. Also see W. Samuelson and R. Zeckhauser, 'Status Quo Bias in Decision-making', *Journal of Risk and Uncertainty* 1 (1988), pp. 7–59.

[23] A. Tversky and D. Kahneman, 'Judgment Under Uncertainty: Heuristics and Biases', *Science* 185 (September 1974), pp. 1124–1131.

[24] G. Northcraft and M. Neale, 'Experts, Amateurs, and Real Estate: An Anchoring-and-Adjustment Perspective on Property Pricing Decisions', *Organizational Behavior & Human Decision Processes* 39(1) (February 1987), pp. 84–97.

anchoring points go so far as to influence home-pricing decisions of realtors. A group of real estate agents estimated the value of a house. The agents had all of the usual information available to them for pricing: property characteristics, prices of neighboring properties, and a chance to see and inspect the house. Even under the circumstances of free information, listing price acted as a powerful anchor on their estimate.

Similar anchoring effects are found in legal contexts. In a series of experiments in the field, Birte Englich and Thomas Mussweiler,[25] a nonexpert advised an expert on the proper solution to a problem. Despite the fact that the expert recognized that the source had relatively low credibility (a layperson making a judgment about an arcane matter), and so their "sentencing demand" to be irrelevant, the mere suggestion was enough to anchor the experts' judgments. In particular, judges who have an average of more than 15 years of experience in deciding criminal sentences are influenced by what they themselves identify as an irrelevant sentencing demand. Anchoring effects are often invoked to explain why such different sentences are given for nearly identical crimes. If jurors make the decisions as to guilt, judges typically determine the sentence. In order to do so, judges rely on a recommended or required sentence. As it turns out, judges' decisions are anchored by the sentence demanded by the prosecutor. This happens even when the judge avows that the sentence demand is irrelevant, and it happens even for judges of different levels of experience.

These biases have a number of common features. Virtually everyone displays them, they operate powerfully in one direction, and they stubbornly resist efforts to control them by spontaneous acts of will. The news, however, is not altogether bad. Although we are prone to very basic errors in nearly every area of human reason, we may be able to turn problems into solutions. We sometimes use rules of thumb, or heuristics, to make decisions about complicated matters, such as picking stocks, or purchasing a home. And sometimes a good

[25] B. Englich and T. Mussweiler, 'Sentencing Under Uncertainty: Anchoring Effects in the Courtroom', *Journal of Applied Social Psychology* 31(7) (July 2001), pp. 1535–1551.

rationale can be found to continue using these heuristics. In a few cases, in practice the heuristics may be difficult to improve upon.

While the news is not terrible for untutored, unaided judgment about problems of routine complexity, this is only if we compare it to the performance of, say, a dog – ok, maybe a chimp – on similar tasks. Of course, our standards are higher, for we traffic in more cognitively complex problems. Perhaps this verdict is controversial, but what is uncontroversial is: (1) deciding the matter requires careful scientific investigation, (2) demonstrated frailties resist easy repair, (3) these frailties are freely prompted by familiar tasks, and (4) we do not know how many successful heuristics in human judgment to expect, given the total number we implement – a crucial factor in deciding when a good heuristic is expected by chance.

I will argue that these biases are, in general, extremely difficult for individuals to correct, and so are, for practical purposes, psychologically incorrigible. The twin demands of efficiency and welfare, then, suggest institutional remedies. These remedies will, to some extent, limit the range of an individual's spontaneous choices, and so will be subject to the charge of paternalism. But I will argue that these remedies are not paternalistic in any substantial sense of the term.

III. AUTONOMY, THE STRUCTURE OF PATERNALISM, AND THE PRICE OF CONCEIT

Implementing institutional assistance might seem to raise a threat to our autonomy, but in fact it is these very biases that damage autonomy; they compromise our ability to effectuate our considered, long-term plans. Why, then, might some people see institutional assistance of biased judgment as paternalistic? Here, a general definition of paternalism might be useful. Paternalism is widely regarded as the interference with a person's actions or knowledge, against that person's will, for the purpose of promoting that person's good. But there are many ways that restrictions are imposed. Such interference is thought to be inconsistent with respect for autonomy because the intervention involves a judgment that the person is not able to

decide for herself how best to pursue her own good. Autonomy is a condition for self-determined conduct, and so paternalism is thought to entail a lack of respect for autonomy.

When the government intervenes, there are at least four kinds of autonomy thought to be affronted:

> When applied to individuals the word 'autonomy' has four closely related meanings. It can refer either to the *capacity* to govern oneself, which of course is a matter of degree; or to the *actual condition* of self-government and its associated virtues; or to an *ideal of character* derived from that conception; or (on the analogy to a political state) to the *sovereign authority* to govern oneself, which is absolute within one's own moral boundaries (one's 'territory', 'realm', 'sphere', or 'domain').[26]

Clearly, the notion of self-government is central to our conception of autonomy.

Due to this centrality, there should be a strong presumption against interfering with individuals' considered judgments about their best interests, the long-term goals they formulate by reflection. When a regulation assists the person in effectuating their long-term goals or plans, there is no relevant interference, and so that regulation is not paternalistic. Were we to treat all instances of government regulation of self-regarding behavior necessarily as instances of paternalism, we risk trivializing the notion of paternalism altogether, stripping it of any meaning independent of regulation and intervention. But once properly separated, we can first ask whether a particular intervention is paternalistic, and then whether the intervention is justified.

'Paternalism' has been defined in a number of ways, many of them controversial. To many, the term must capture the experience that prompts opposition to the regulation of conduct. On this view, paternalism requires that the institutional choice be imposed against the will of the person affected. We can see this in one prominent account:

---

[26] J. Feinberg, 1986, *Harm to Self* (New York: Oxford University Press), p. 28. Feinberg's book contains perhaps the most influential modern discussion of paternalism. A very useful discussion of paternalism, and its legal implications, can be found in Lawrence O. Gostin, *Public Health Law: Power, Duty, Restraint* (Berkeley: University of California Press, 2000).

Paternalism is the interference of a state or an individual with another person, against their will, and justified by a claim that the person interfered with will be better off or protected from harm.[27]

Our government requires us to contribute to Social Security, a form of pension system. The government prohibits the sale of assorted drugs considered ineffective. It both forbids the sale of various drugs believed to be harmful, and regulates the names of prescription drugs on the market. The government forbids consent to certain forms of assault to be a defense against prosecution. Some of our state governments legislate that motorcyclists must wear helmets. The list of government-regulated actions could be continued at length.

In all of these cases, the regulation is justified by appeal to the agent's best interests. But, no matter how common this conception of paternalism, it contains a necessary condition – that the interference be against the person's will – that vastly overestimates the insult to liberty produced by insensitivity to current, episodic, will. After all, one's spontaneous desires are often ill-considered, and one's actual, long-term ends are often obscure to even that agent. Indeed, in order to promote their autonomy, normal people often bind themselves to act in ways that might be contrary to their current will, and so control for the caprice of current desires. True, people can resent the reach of government agencies for all sorts of reasons, but not all of those reasons track the promotion of our autonomy, or the securing of our long-term goals.

If we are to address the questions whether, and when, cognitive error warrants government intervention, it is worth knowing under what conditions government intervention, is *ever* warranted. The most obvious case is the one in which the conditions for Mill's principle of liberty are not satisfied, and (the consequences of) an individual's decisions affect many others. The decision to wear a motorcycle helmet or a seatbelt can affect not just the pertinent actor, but also family members

[27] G. Dworkin, 'Paternalism', in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2002 Edition). http://plato.stanford.edu/archives/win2002/entries/paternalism/

who must care for the driver when he or she suffers a traumatic head injury or death. This consequence will also be a cost in the group insurance pool. If you think a government requirement to wear seatbelts is impermissibly paternalistic, you are very likely to think that institutional solutions to cognitive error are as well.

As Hume observed, reason alone does not move man to action. So knowing what is right (to radically reduce traffic fatalities at minimal cost) does not thereby prompt compliance. The obstacles are many, but arrogance is surely one. In particular, people don't like being told what to do. Some don't even like the very idea that they *can* be told what to do. In this case, it may be tempting to derogate all institutional regulations with the typically pejorative "paternalism". Given a sufficiently austere conception of negative liberty, outcries of paternalism can arise from little more than the childish desire to not be told what to do. Every institutional constraint on behavior is bemoaned as an unacceptable intrusion on one's freedom, a case of undue paternalism.

But as we have seen, genuine cases of paternalism are more difficult to find than one might imagine. Paternalism is a very specific relation. The existence of "implicit rationales" or actual (but often unrealized) motives accounts for cases in which one interferes with the behavior of another, but is justified in doing so not because one knows what is best for the other, but because the behavior is inconsistent with the implicit rationales of the actor. Interference with a recreational drug-induced suicide attempt is a typical case of a liberty-limiting action that is justified. But is it paternalistic?

Consider the two standard types of paternalism, described by Feinberg:

> Hard paternalism will accept as a reason for criminal legislation that it is necessary to protect competent adults, against their will, from the harmful consequences even of their fully voluntary choices and undertakings." (Feinberg 1986, p.12)

> Soft paternalism holds that the state has the right to prevent self-regarding harmful conduct (so far as it *looks* paternalistic) *when but only when* that

conduct is substantially nonvoluntary, or when temporary intervention is necessary to establish whether it is voluntary or not. (Feinberg 1986, p.12)

What kinds of conditions make choice substantially nonvoluntary? Among them: "[I]gnorance, coercion, derangement, or other voluntariness-vitiating factors" (p.12). If these factors predominate, then the conduct "does not come from his own will, and might be as alien to him as the choices of someone else." (p.12)

Many legal scholars advocate intervention not just because it would be best for the decision-makers, but because it will be best for others as well. It is worth noting that, according to the definition of paternalism above, intervention that is based on third-party effects is not paternalistic. If the presumption of noninterference can be overridden by the dramatic increase in safety for the interested party – as happens in the case of a compulsory seat belt law – why shouldn't it be overridden in the case of large benefits to other parties? For example, pooled insurance resources can get sapped by injuries and deaths from failing to wear seatbelts. So insurance rates would go up without the law.[28]

The existence of the biases raises serious doubts about the ability of normal individuals to make decisions that are, *by their own lights*, as effective or otherwise as good as tested institutional decisions. The common good calls for the participation of individuals even if it is not in their individual interests to participate, but rather in the interests of other people, and in the public interest, to participate. It is less

[28] A paternalistic attitude toward children in the legal system is warranted by their inability to effectively participate in their own case ('adjudicative competence'). Systematic cognitive error in adults cannot be accurately analogized by the incompetence of children, because many of the biases could be corrected with sustained attention and education, but no amount of attention and education will make a child competent. A hint of how to treat cognitive error resides in the sensitivity of this account to issues of maturity. One reason that children are appropriate objects of paternalism is that they are not mature judges of facts; their immaturity limits them in this respect. However, adults are similarly limited in important respects, and it is worth noting that the evidence for these biases reflects the adult inability, in important cases, to control their own judgment.

expensive and more effective to place fluoride in the local water supply than to distribute fluoride pills to individuals or households.[29] Because your participation is for their sake and not solely for yours, the institutional intervention is not paternalistic. Instead, it is of a piece with ordinary social policies. The legislated compliance results not from the claim that the government knows what is best for you and so can secure it for you without your consent. Rather, it results from the fact that it is a more expensive and less effective way to achieve a social priority to permit defection from the policy, even if a small number of people want to defect and would be personally financially better off if they did so.

So in order to legitimately charge paternalism, it must be shown that the legislation of this behavior is justified solely in terms of the best interests of the affected individual. In one sense, then, institutional assistance in judgment need not be an issue about paternalism at all. If the effect of the institutional strategy is one that the agent would support, it cannot be said that the institutional measure is paternalistic. If people are brought to recognize the costs of these biases, they can adopt strategies that bind themselves at one remove, with effective voluntary control over doing it or not. In that case, it is not paternalism – for it is the citizen rather than a government body that is making a decision about how best to handle the otherwise treacherous matters. In these frequent cases, the institutional assistance is minimally compatible with the agent's "implicit rationales".[30]

IV. THE FULLY INFORMED AGENT, DEBIASING, AND PATERNALISM

If the agents are competent and otherwise rational, what is the warrant for intervention in cases of cognitive bias? This basis of this warrant can be expressed as a test: The intervention is warranted if, against a background of a fully informed deci-

[29] On efficiency considerations, see J. Feinberg, *Harm to Self* (New York: Oxford University Press, 1986), p. 377.

[30] See R. Arneson, 'Mill vs. Paternalism', *Ethics* 90(4) (1980), pp. 471–472.

sion-maker and an unbiased standard, the decision-maker would not have, say, driven dangerously fast. In the present case, if the decision-maker knows that they can't resist taking that risk, they would have consented to any number of low-cost corrective measures.

The relation between liberty, choice, and welfare is a complicated one. But the balance of legal doctrine and political philosophy deny that unfettered liberty and unlimited choice reign absolute. Neither liberty nor choice is an unconditional value. I have argued that cognitive biases are common, spontaneous, and costly. Because of this pervasiveness and intractability, the damage they cause is bound to be great, and costly (at the start) to reverse. And like cognitive practices with costly and potentially self-destructive effects, the costs may be great enough to warrant regulating an individual's choice about a largely self-regarding issue.

Debiasing promotes rather than undermines autonomy. After all, the biases threaten our ability to meet our considered and long-term interests. By recognizing the threat of the biases, we can increase the probability that we can meet our long term interests. The approach I favor (the "fully informed agent" approach described above) achieves this not by casting institutional intervention as soft paternalism; nor does it suppose that individuals have diminished capacity to make decisions that contribute to achieving long-term interests (as soft paternalism does when we intervene in the decisions of children). At the same time, the approach I am suggesting does not impose a substantial conception of the good on individuals. Rather, it enhances their chances of effectuating those goals. This approach leaves room for many cases of genuine paternalism – for example, in measures that dictate that fully competent individuals act on behalf of a substantial conception of the good that they would not have chosen if fully informed.

There is little question that, with a minimum of assistance, we could achieve greater accuracy in calculating what our priorities require. However, in light of the superior performance of institutional prosthetics in decision-making, what explains the resistance to them? After all, the rationale for permitting them

does not require telling people what they should want, but rather how to ensure that they have the most accurate available solution to securing what they want. At the bottom of resistance to institutional assistance in decision-making is the conviction that individuals should be permitted to simply choose a course of action, no matter how inferior that course of action is for that agent. While the capacity for informed choice leads libertarians to reject any independent institutional dictates on competent agents, libertarians are not alone in the focus they give to individual choice. The case for the capability of autonomous judgment is compromised just to the extent that the capacity for individual choice is limited. It is no accident, then, that Kant likens intellectual irresponsibility to the diminished decision-making capacity of the child:

> Enlightenment is man's emergence from his self-imposed immaturity. Immaturity *is* the inability to use one's understanding without guidance from another. This immaturity is *self-imposed* when its cause lies not in lack of understanding, but in lack of resolve and courage to use it without guidance from another. *Sapere Aude*! 'Have courage to use your own understanding!' – that is the motto of the enlightenment.[31]

However, Kant makes intellectually responsible judgment appear both simpler and safer than it really is. Fifty years of experimental psychology demonstrates that humans have an ample supply of ''courage'' to use their own judgment, and that results of this confidence are less than impressive.[32]

There are many conditions that paternalism does not require. Paternalism does not entail coercion or interference with the affected person's liberty of action. In medical settings, a doctor may lie to a person on their deathbed when the information could only make what was left of their life worse.

---

[31] I. Kant, 'An Answer to the Question: What is Enlightenment?', in *Perpetual Peace and Other Essays*, trans. Ted Humphrey (Indianapolis and Cambridge: Hackett Publishing Company, 1983), p. 41.

[32] Obviously, my purpose here is not to take a gratuitous swipe at Kant for ignoring psychological research; the research in question would not even be available for another 200 years after Kant's essay on enlightenment. Rather, the point is that we have inherited an Enlightenment tradition that promoted false beliefs about the accuracy of adult judgment, and about the risk of overconfidence.

Such decisions are based on justifying the treatment given by appeal to their own good, but do not involve coercion or interference with liberty of action. In order for a constraint to be seen as paternalistic, it must attempt to protect the individual from acts arising from that person's will. As we have seen, the biases themselves arise independently of the will; they are a factor external to it. So there is no issue of protecting someone from themselves or interfering with their liberty. The biases present dangers, but the person is not acting in regard to this danger in any respect; typically, they are unaware of the bias, and so unmindful of its harmful consequences. Therefore the protection sought from an institutional constraint focuses on something that is not the actor, a consequence that is not of an intended action. If individuals routinely make crucial and predictable errors in judgments about their own welfare, and are unable to control doing so without turning life into an existence of contemplative paralysis or one of distorted value otherwise disavowed, then we should ask for an argument *against* introducing institutional prosthetics. That is, if people want to retire with minimal security but normal humans irrationally discount the future, we need to either bind humans to defer gratification and save for retirement, as the U.S. government does by making Social Security provisions compulsory, or make the incentives to do so more attractive, as the U.S. government does by allowing a limited voluntary contribution to retirement savings to be pretax.[33]

If one were to confuse the numerical increase in choices with an increase in free choice, one might suppose that it is preferable to choose less wisely as long as you are doing so freely. But freedom of this sort is ironically constraining. In fact, there is reason to suppose that unlimited spontaneous choice is not even desirable. Sheena Iyengar has done ingenious work on the

[33] I don't propose to address those charges of paternalism prompted by extreme sensitivity to interference. Sensitivity to institutional regulation varies greatly across individuals. For some, the placement of fluoride in the water, or mandatory seatbelt use, is a scourge on liberty. For others, steep taxation for state pension pools is treated as a legitimate and even cherished policy.

rationale for limiting choice. In the Godiva chocolate study, people in one condition chose from 6, and in the other from 30, different kinds of chocolates. Those in the '6' condition were more likely to choose, and to make a choice that was satisfying. It appears that you feel more responsibility for your choice the more options you have. This may be because we can imagine how much better other options might have been. The same goes for 401(k) plan options. As the options increase, people are less likely to enroll. Choice overload causes choice-noise, making us unable to decide with any fidelity among the many options. Paralysis is the natural outcome of 'choice-noise'. The reason for this paralysis is simple. With so many options, a poor choice is a real defeat. This leads to a heightened sense of responsibility for the choice made, and an unwillingness to commit to selection should it turn out unfavorably. Effective liberty requires only a menu of options (which is what people select from anyway), and decisional paralysis occurs when there are too many options.[34]

Before cries of pessimism and bondage erupt, let me be clear about lessons of the empirical work on judgment and choice. The psychological findings do not show that people CAN'T make good choices – in fact, there has been far less research on correcting biases than establishing their existence, so we really don't know whether, as a routine matter, they CAN'T. Rather, the psychological findings show that people, even devoting enormous cognitive resource to correction, DON'T make good choices. Accordingly, we engage in behavior we would not if we could properly bind ourselves in some second-order way. Moreover, institutional constraints on defection are violations of the desire for spontaneous choice, not free choice *per se*. Indeed, if people were made aware of the sub-optimal nature of, say, their retirement investing behavior, they might freely consent to a bureaucratic arrangement that binds them.

[34] S. Iyengar, W. Jang, and G. Huberman, 'How Much Choice is Too Much?: Determinants of Individual Contributions in 401(K) Retirement Plans', in O. S. Mitchell and S. P. Utkus (eds.), *Developments in Decision-Making Under Uncertainty: Implications for Retirement Plan Design and Plan Sponsors* (Oxford: Oxford University Press, 2004).

Now, it might be that we *can* make good choices if each individual is willing to tolerate the exorbitant costs required to counteract the natural pull of the biases. But the cost of that attention has its limits. If we shouldn't spend every dime of revenues to ensure protection of our property, we should not devote unlimited resources to costly individual reasoning strategies just to escape administrative regulation.[35] But where there are comparably effective debiasing methods, we should favor the one that imposes the least restrictive interventions.

V. LIBERTY-SENSITIVE AND WELFARE-SENSITIVE DEBIASING

How, then, can we efficiently assist, and thereby correct, these errant judgment processes. Under these circumstances, how can we improve judgment and action? Two strategies are suggested. An *inside strategy* is a voluntary reasoning process designed to improve the accuracy of judgment by creating a fertile corrective environment *in the mind*. Psychologists have developed a number of inside strategies for correcting biases – consider-the-opposite, perspective-taking, theory-based adjustment – but even the most effective inside strategies are only a qualified success.

The most prominent inside strategy, applied to correct overconfidence and hindsight biases, is called the "consider the opposite strategy". According to one of the groundbreaking studies on debiasing, people "have a blind spot for opposite possibilities" when making social and policy judgments.[36] And so, this 'inside' strategy urges people to consider alternative hypotheses for the occurrence of the very event that they believe they understand. While it is perhaps too much to ask that

[35] We cannot conclude that we make good decisions most of the time, simply because things don't go horribly wrong all of the time. At best, we can conclude that people's reasoning systems satisfice most of the time. In this case, bureaucratic intervention is designed not to paternalistically protect individuals from themselves, but to ensure an allocation of resources that is as socially responsible as possible.

[36] C. Lord, M. Lepper, and E. Preston, 'Considering the Opposite: A Corrective Strategy for Social Judgment', *Journal of Personality & Social Psychology* 47(6) (1984), pp. 1231–1243.

people shoulder technical burdens in lay life here, "consider-the-opposite" is a portable inside strategy that is marginally effective. For any belief that we can hold with undue certainty (e.g., "New York State is the largest state on the Eastern seaboard", "Los Angeles is west of Reno" or, more tragically, "the defendant is guilty beyond a reasonable doubt"), we can follow a simple rule: "Stop to consider why your judgment might be wrong."[37] For example, ask yourself whether, respectively, you have considered South Atlantic states that get less press, the orientation of the U.S., and your confusion over the DNA evidence. When asked to generate pros and cons for a judgment made, Koriat, Lichtenstein and Fischhoff[38] demonstrated that overconfidence bias was reduced. Indeed, they found that it was the generation of *opposing* reasons that did all of the bias-reducing work. What did not work to decrease bias, was the mere instruction to think harder about the problem, to concentrate, or to give the problem greater attention.

For certain kinds of inside strategies of debiasing to be effective, then, a fairly demanding set of conditions must exist. Decision-makers have to be (a) motivated to give an accurate judgment, (b) aware of the potentially distorting influence, and (c) aware of the direction and magnitude of this influence. Understandably, these are difficult conditions to meet. The decision-maker must also invest effort in generating specific alternative outcomes, and in order to do so they must have the cognitive capacity, attentional focus, and undistracting environment to carry it out. These conditions are seldom jointly available. And so, an inside strategy allows the conditions to persist that tempt defection. In addition to these conditions that are difficult to realize in practice, the process of correction often recruits the same cognitive mechanisms responsible for the trouble in the first place. If distinct incentives have little effect, it is not surprising that bland instructions to concentrate or attend to the evidence are ineffective as well. Such instructions

[37] S. Plous, *The Psychology of Judgment and Decision-Making* (New York: McGraw-Hill, 1993), p. 228.

[38] A. Koriat, S. Lichtenstein, and B. Fischhoff, 'Reasons for Confidence', *Journal of Experimental Psychology: Human Learning and Memory* 6 (1980), pp. 107–118.

simply invoke the already defective cognitive routines: "[B]iases in social judgment can be corrected only by a change in strategy, not just by investing greater effort in a strategy has led to biased judgments in the first place."[39]

Inside strategies, such as consider-the-opposite, are liberty-sensitive. Like most inside strategies, consider-the-opposite won't engender much resistance, because you can decide to make the alternative considerations or not. Participation is voluntary, not compulsory. At the same time, the consider-the-opposite approach, like most inside strategies, has predictable limitations. On the one hand, we normally defect from such voluntary prescriptions; if left unsupervised and untrained, we would not acknowledge the need to debias. So reasoners need to be effectively educated (which may require being effectively motivated). On the other hand, educational remedies would require that the government unleash a veritable army of teachers into our schools and businesses, in the hopes of reaping even the modest benefits that inside debiasing strategies offer. Ironically, then, this "liberty-sensitive" remedy promises to be costly, inefficient, bureaucratically cumbersome, intrusive, and in the end, ineffective to boot.

Now that we know the treachery of subjective judgment and the limited effectiveness of inside strategies, it would be dishonest to ignore it or to persist in the sanguine conceit that common sense counteracts these systematic and costly inaccuracies. We now know from the above studies that general admonitions to concentrate or attend to the evidence does not improve people's performance.

More effective debiasing strategies also tend to be more restrictive, and these are predominantly outside strategies. An *outside strategy* identifies features *of the environment* whose presence can be manipulated to produce the most accurate or desirable available outcome. Outside strategies place a priority on welfare, inside strategies on liberty.[40] Of course,

[39] C. Lord, M. Lepper, and E. Preston (1984), pp. 1236–1237.

[40] S. Hurley, 'Imitation, Media Violence, and Freedom of Speech', *Philosophical Studies* 117(1/2), (2004), pp. 165–218, gives a name – bypass effects – to the effects that bypass autonomous deliberative processes from those that do not.

these are just the end points on a continuum of strategies that range from governmental limits on our options, to a planned information format submitted to a decision-maker, to mental exercises that a decision-maker is simply asked to use. When appropriate, the outside strategy can ''trick'' the chooser, enforcing a different choice set or rule of conduct. A behavioral policy based on an outside strategy recommends, for example, that the alcoholic avoid the bar in the first place, thus eliminating the conditions that tempt defection. This outside, 'policy' approach improves decision-making by changing the dimensions of the choice-set. A good example of an outside strategy is the prevention of 'independent' auditors from working with a bank or brokerage firm for more than, say, five consecutive years.[41] Rather than simply advising auditors to be impartial, or expecting them to be professional and direct in delivering bad news to the company responsible for their employer's financial growth, the outside strategy removes the threat to integrity by eliminating its source. In so doing, an outside strategy might require that you select a solution that is not intuitively satisfying, but is objectively correct.[42]

Because the biases are systematic and psychologically incorrigible, the model of bias-correction I recommend treats biases as though they are addictions. This 'addiction' model addresses judgment errors by adopting tested strategies and then erecting barriers to defection from these strategies. The barriers can take many forms. Defection can be extremely costly, painful, enormously inconvenient, or embarrassing. These barriers – disincentives of various sorts – bind us to a reliable strategy, and in so doing, prevent defection. Self-binding strategies often work because they reduce the number

[41] F. Reeves, 'The Psychology of Accounting Fraud', *Pittsburgh Post-Gazette* (citing research by Don Moore, Max Baxerman and George Loewenstein), Wednesday, December 11, 2002; http://www.post-gazette.com/businessnews/20021211moorep2.asp

[42] D. Kahneman and D. Lovallo, 'Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking', *Management Science* 39 (1993), pp. 17–31. The designations of 'inside' and 'outside' strategies can be found there.

and kind of spontaneous decisions that the actor must make, and because they form successful strategy that is easy to implement.

We tend to blame the failure of spontaneous individual decisions on generic limitations like weakness of will. But, like addictions, the biases have a stable biological source, reinforced by habit, and are very difficult, in practice, to counteract. And, like addictions, they are best treated by not tempting defection: never permitting a forbidden taste "just this once", and reducing exposure to environments that trigger the bias. Unlike addictions, however, cognitive biases are not the burden of an unfortunate subpopulation of diminished capacity – the traditional objects of paternalistic intervention; rather, they are the province of humanity, part of our natural condition. So institutional solutions, rather than spontaneous individual efforts, offer the best hope for bringing our damaging tendencies into line with our objective interests and the common good. The route to transforming persistent and self-serving oversights into disciplined and painless commitment lies with decisions of small but incremental impact, structured in the right way. Institutions provide the structure, and 50 years of scientific research on the psychology of human judgment supplies the material for the improvement of happiness and human welfare.

One way to avoid defection is to adjust the structure of a reasoning problem or learn to represent the problem in a more transparent way. To the extent that these particular strategies work, their desirability is based on particular features of the problem: their generality (the scope of the problems they address), their frequency (how frequently the types of problems they address actually occur), their significance (how important the problems are to human welfare), and the costs of implementation (how simply and cheaply the problems can be addressed by these methods). Let's briefly consider one such effort.

A. *Adjusting the Structure of the Problem*

One proposal to improve individual reasoning is to engage in a bit of problem-engineering: analyze the structure of the

problem and adjust the way the problem is represented so that human cognitive capacities are able to appreciate the features key to its solution. Frequency formats are a good example here. People tend to interpret risk in terms of probabilities: my risk of heart attack is determined by a number of factors each of which, as predictors, has a different accuracy: gender, age, cholesterol level, amount/regularity of exercise, etc. So people tend, spontaneously, to equate their risk of heart attack with the probability that people with all of their factors will have a coronary event.

Gigerenzer[43] shows how to significantly improve people's reasoning on diagnosis problems without a lot of complicated statistical training. It turns out that people do much better on these sorts of problems when they are framed in terms of frequencies rather than probabilities. Let us begin with a problem adapted from Gigerenzer.[44] Imagine that, as a woman between 40 and 50 years of age, you are advised to participate in routine breast cancer screening. Mammography is the standard screening test. For women in this age range that are symptom free, the information below is made to available to them so that they can interpret the results of their mammogram. There are two mathematically equivalent formulations of a diagnosis problem. They are presented in two different formats.

B. *Probability Format*

For women at age 40 who participate in routine screening, the probability of breast cancer is 1%. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 10% that she will get a positive mammography. Suppose a woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? ___%.

[43] G. Gigerenzer, *Adaptive Thinking: Rationality in the Real World* (New York: Oxford University Press, 2000).

[44] Gigerenzer (2000).

C. *Frequency Format*

About 10 out of every 1000 women in this group (age 40–50 who participate in routine screening) have breast cancer. About 8 of every 10 women with breast cancer will get a positive mammography. Out of the 990 women *without* breast cancer, 99 will also get a positive mammography. Now you are introduced to a new representative sample of women at ages 40–50 with no symptoms who got a positive mammography in routine screening. How many of these women actually have breast cancer? ___ out of ___.

People with no training in statistics tend to do much better on problems presented in the frequency format. Gigerenzer reports that 16% of subjects faced with probability formats got the correct answer (arrived at by Bayes's formula), while 46% of subjects faced with frequency formats got the correct answer.[45]

There is no mystery why subjects have an easier time with the frequency format than the probability format. First, the frequency format makes the base rate information transparent, the frequency with which the event occurs in the relevant population (the population of women between 40 and 50 years of age). Second, the frequency format requires performing a much easier calculation. These results suggest an obvious reasoning strategy: when faced with a diagnosis problem, people should learn to represent and solve the problem in a frequency format.

This advice to use frequency formats is more easily given than implemented. Our heuristics apply to reasoning problems as they spontaneously arise, and many, if not most, of these are cast in terms of probabilities. Frequency formats have undeniable advantages, but what resources must be spent to transform spontaneous probabilities into frequencies? Adjusting informational structure can be costly. Depending upon how sweeping the intended reform, doctors would have to be trained in the understanding and presentation of risk

[45] G. Gigerenzer, 'The Psychology of Good Judgment: Frequency Formats and Simple Algorithms', *Medical Decision Making* 16 (1996), pp. 273–280.

information, news agencies would have to assess the impact of untutored presentation of statistical information in newspapers, radio, and television news, etc. In the case of frequency formats, the informational structure does most of the work for the cognizer. But in creating the appropriate informational structure, the start up costs may exceed, by a large margin, the opportunity costs of relying on untutored judgment in unstructured settings.

These adjustments may themselves require institutional policies or constraints as rich as those that prompted the anti-paternalistic objections. The compulsory transformation of informational structure will involve bureaucratic intrusions that are unacceptable to those who value ownership of their own spontaneous error, even when they know the bureaucratic route to avoid it. For those with a conservative reading of liberty-limiting regulations, the most obvious alternative to governmental intrusion is individual education. The hope is that people can *learn* to correct each of the biases that tempts them. But here again, correction by individuals would require an educational program, a bureaucracy that trains and certifies teachers and administers their employment and their participation in schools and businesses. In short, the correction of these spontaneous errors will require institutional interventions, at either the point of problem presentation or in the process of learning. And these interventions are self-defeating features of any debiasing strategy intended to be liberty-sensitive.

VI. PROMOTING AUTOMOMY AND ENHANCING WELFARE BY HARNESSING BIASES

For problems that can be so handled, we should use a bias-harnessing procedure – neither imposed from without nor guided from within. In order to illustrate the structure of bias-harnessing, we will begin with a simple perceptual case. After that, we will briefly consider two cognitive cases of bias-harnessing – message framing to encourage preventive medical testing, and the creative inertia of the (SMT) pension plan –

that make our biases work for us.[46] Both make use of the psychological research on judgment to formulate the least intrusive means of effecting solutions that maximize the welfare of the affected individuals. Of the institutional strategies, bias-harnessing methods such as message framing and creative inertia may be the least intrusive of the options. It preserves, and so honors, the individual's ability to choose to participate in the course of action.

Perceptual cases of bias are well-known and understood, because perception is a rigid and fast process that pays a price in stupidity for its haste and inflexibility. Yet, this stupidity can be used for good. The Department of Transportation has run experiments using optical illusions to get speeders to slow down. Chevron markings, distance cues that make the road appear to be narrowing, "convince drivers that they are traveling faster than they really are" according to a research study by the American Automobile Association. And this design has been implemented with palpable effect. In Japan, chevrons reduced by 40% the crashes across six locations. Although there is some evidence of adaptation to chevrons for repeat drivers, this design has clear applications. Department of Transportation Commissioner Lynn announced that "this is a proven, simple and inexpensive way to slow down drivers who are approaching dangerous intersections or residential neighborhoods at high speeds."[47]

[46] C. Sunstein and R. Thaler, 'Libertarian Paternalism Is Not an Oxymoron', *University of Chicago Law Review* 70 (2003), pp.1159–1202. They, too, discuss the Save More Tomorrow plan in connection with paternalism, but to emphasize a different issue. Sunstein and Thaler use it as a nice example of 'libertarian paternalism', in which there is an effective choice of actions – and so there is freedom of choice – and the opportunity to defect or 'opt-out' without imposing a high cost on that choice. Also see Colin Camerer, Samuel Issacharoff, George Loewenstein, Ted O'Donoghue, and Matthew Rabin, 'Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism', *University of Pennsylvania Law Review* 151 (2003), pp. 1211–1254. Interesting cautions about the 'libertarian' label can be found in Barbara Fried, 'Left Libertarianism: A Review Essay', *Philosophy & Public Affairs* 32(1) (2004), pp. 66–92.

[47] http://www.transalt.org/press/magazine/972MarApr/04-5reclaiming.html

In the Netherlands, researchers addressed the problem of slowing drivers on 80 km/h roads, particularly near village entry points. Researchers decided on a kind of rippled shoulder, which, in effect, made the lane narrower. An effectively narrower lane would make it more difficult to drive, and so slow the drivers. The lanes did not actually change in width, but the shoulder rippling, researchers feared, might cause drivers to edge toward oncoming traffic. In fact, it did have this effect. This tendency was counteracted by widening the center line and narrowing the rippling on the shoulder. A simulation of this new configuration demonstrated a significant effect in speed reductions.[48] When used on a real road in a before-after study, this configuration allowed a 20% decrease in accidents after 2-years.

At least some of these traffic patterns don't involve much interaction among drivers. So it is the driver's own behavior that places him or her at risk, a potential harm not to others but to themselves. In cases where the government is attempting to prevent harm to self, is it paternalistic for a governmental agency, such as the Department of Transportation, to use an optical illusion to secure compliance? Because the driver is not aware that they are reacting to the illusion that they are going faster than they in fact are, they are not choosing to slow down on the basis of the responsible use of information. Instead, they are, in effect, being tricked.

Here is the question: is it paternalistic to use this 'trick' – the harnessing of a bias by using chevrons in road design? The automatic perceptual response it evokes bypasses the process of deliberate evaluation of options. But, it does secure compliance, and not just idly; it saves lives. What is the price for this trick? Well, in general, people don't like to be deceived, especially by an arm of the government. Now, suppose they could make a choice, being told that they are objectively safer in contexts in which chevrons and other engineered precautions will extract compliance. Is this a kind

[48] R. Van der Horst, and W. Hoekstra, 'Testing Speed Reduction Designs for 80 Kilometre per Hour Roads With Simulator', *Transportation Research Record* 1464 (1994), pp. 63–68.

of circumvention of deliberate choice that a reasonable person would reject? If so, why would a reasonable person choose a riskier option when there is no cost associated with a safer one? And if not, what is the basis of the more general concern about institutional prosthetics for human judgment – particularly when the success of the outcome is not only measurable but striking?

Intervention is warranted here because, at negligible cost, the outside strategy serves the driver's goal of living a longer life. This type of goal is not especially controversial, and we should assign presumptive favor to any measure that promotes our autonomous pursuit of this goal. If given the choice, few people would select the course of action that they believed would lead to the shorter life. But the threats to our long-term goals can be complicated enough that we can benefit from some cognitive prosthetics.

A particularly dramatic and recent success comes from bias-harnessing research in behavioral finance. At the end of 2003, employees of many U.S. companies were able to use a plan called Save More Tomorrow when they make contributions to their retirement plan. Developed by business professors Richard Thaler at the University of Chicago Graduate School of Business and Schlomo Benartzi of the Anderson School of Business at UCLA, the SMT Plan allows employees to direct a portion of their future salary increases toward retirement savings. This plan is unique because it uses psychological research in behavioral economics to design a plan that employees will join and then not defect from. SMT uses our own inertia and procrastination to our advantage. We say we want to save more but don't take the necessary steps. People procrastinate. So the plan application asks prospective participants if they would like to start three months from now, and commits them to doing so at the time of enrollment. This allows them to experience the deferral of commitment to an unfamiliar or effortful change in a course of action, and so to experience whatever they find attractive about procrastination. But once in the plan, inertia takes over (abetted by the status quo bias), and people tend not to opt out. As well, judgment

research shows that people are loss averse, weighing losses far more heavily than gains, and this prevents them from enrolling in a program in which they can witness the decrease in their paycheck. So Thaler and Benartzi built loss aversion into the plan, taking the increased contribution out of the pay raise, so that the participant does not experience it as a loss or reduction. The result? By predicting and anticipating the pitfalls of procrastination and defection, saving rates more than tripled, from 3.5% to 11.6%, over 28 months.

With the number of individuals with self-directed pension plans in the millions, this bias-harnessing procedure, preventing discounting, has the potential to create a demographic of financially comfortable, rather than desperate, retirees, with all of the benefits that their happiness and welfare can bring. This is an admittedly important financial risk averted, but not all risks are financial. Consider 'message framing' in the area of health care. Messages carrying the same statistical information can be cast in different ways, to different effect. You can report to the public that people within a certain age range and illness stage who get tested for HIV or breast cancer and test positive have a 70% chance of living beyond 7 years, or a 30% chance of dying in under 7 years. These messages have different impacts. In order to test the relative effectiveness of health messages, researchers at Yale and the University of Minnesota created videotapes that were either gain or loss framed.[49] The gain-framed videos explained the positive effects of healthy behavior and regular breast exams, and the loss-framed video attempted to focus attention by frightening the viewer with the bad things that could happen if they don't see a doctor. The subjects in the gain-framed message condition were significantly more likely to arrange mammograms.

[49] T. R. Schneider, P. Salovey, A. M. Apanovitch, J. Pizarro, D. McCarthy, J. Zullo, and A. J. Rothman, 'The Effects of Message Framing and Ethnic Targeting on Mammography Use Among Low-income Women', *Health Psychology* 20 (2001), pp. 256–266.

Message framing has been equally successful in motivating HIV testing.[50]

Would a reasonable person be indifferent to the presentation format of a problem if they knew that, were they carrying the illness, the positive frame would make it more likely they would make a decision that would save their life? Once again, the researchers had harnessed a bias – the framing bias – in order to cultivate behavior that enhances welfare. And, once again, the warrant for intervention in individual decision-making depends on a principle that virtually anyone can abide: if there were a debiasing option, or if the decision-maker were fully informed, they would not have made the decision they did.

There is no evidence that the same corrective success could be effectively or routinely achieved by inside strategies, strategies of individual motivation that attempt to acquire a more accurate representation, or to consider alternative possibilities. Outside, welfare-enhancing strategies may require institutional arrangements, in the form of government-sponsored regulations, or advertisements to make salient the options that would be selected if the agent were fully informed and unbiased. But, like governmental restrictions by the SEC or the FDA on misleading statements, institutional support for higher-fidelity information enhances welfare at low costs. And when the support is for bias-harnessing strategies, such as message framing or the SMT plan, there is no insult to effective choice. These are not cases of justified paternalism; they are not instances of paternalism at all.

Earlier, we considered people who want to ''own'' their mistakes, even when they recognize it may make their long-term goals more difficult or impossible to attain. But these individuals are in

[50] P. Salovey and P. Williams-Piehota, 'Field Experiments in Social Psychology: Message Framing and the Promotion of Health Protective Behaviors', *American Behavioral Scientist,* 47 (2004), pp. 488–505; A. Apanovitch, D. McCarthy, and P. Salovey, 'Using Message Framing to Motivate HIV Testing Among Low-income, Ethnic Minority Women', *Health Psychology* 22 (2003), pp. 60–67. See more generally, A. J. Rothman, and P. Salovey, 'Shaping Perceptions to Motivate Healthy Behavior: The Role of Message Framing', *Psychological Bulletin* 121 (1997), pp. 3–19; and R. Cialdini, 'Crafting Normative Messages to Protect the Environment', *Current Directions in Psychological Science* 12(4) (2003), pp. 105–109.

the minority, and law and policy is forged to handle the conditions of the majority. For example, there is a strong presumption that 18 year olds can consent. Indeed, unless there is positive evidence of cognitive disability, reaching 18 years of age is a proxy for adult competence. Perhaps it would be more accurate to test each individual for competence. But this level of attention would soon become unmanageable, defeating the goals that competence was thought to serve. The practical demands of law and policy often require simple observable standards, especially where greater accuracy could be purchased only at exorbitant costs. And this is so even if it constrains the liberty of a minority of individuals. The same goes for majority policies like the 40 hours work week. Some people may want to make more money by working 50 hours per week, and so oppose the Fair Labor Standards Act. These people may feel that this Act constrains their liberty. But for the great majority, this Act protects individuals from exploitation.

If outside strategies, such as government regulations or other institutional measures, are more efficient and effective and are not paternalistic, then those are the ones that should be implemented. At the same time, the government's interest can be balanced against the individual's interest in enjoying the most unconstrained liberty possible that is consistent with making decisions that effectuate our autonomous, considered judgments.

The attractions of outside strategies have been implicitly acknowledged in some of the policy proposals in legal scholarship. Consider the research on the tort liability of a defendant-injurer. Because the hindsight bias will cause us to overestimate the liability of defendants, we should seek alternatives to tort liability as a means of encouraging precaution; in particular, we must establish ex ante safety regulations. This proposed policy is clearly an outside strategy, and has been explored by a number of legal scholars.[51]

[51] K. Kamin and J. Rachlinski, 'Ex Post $\neq$ Ex Ante: Determining Liability in Hindsight', *Law and Human Behavior* 19(1) (1995), pp. 89–104; C. Jolls, C. Sunstein, and R. Thaler, 'A Behavioral Approach to Law and Economics', *Stanford Law Review* 50(5) (1998), pp. 1471–1550; R. Korobkin, and T. Ulen, 'Law and Behavioral Science: Removing the Rationality Assumption from Law and Economics', *California Law Review* 88 (2000), pp. 1051–1144.

Most of the anticipated cases of institutional prosthetics entail neither hard nor soft paternalism; in fact, they are not paternalistic at all. By the standards set out in this paper, the cases of bias-harnessing are not paternalistic. In the instance of retirement investing, the person chooses to enroll, and the participant can opt out of the program at any time, without penalty. Because it involves precommitting to a course of action that might not be continued if revisited, it is a case of self-binding.

What should the limits of outside, institutional strategies be? The government should not restrict liberties in a way that would produce preferences that are mere artifacts of the restricted choice set. In such a case, people are rendered unable to examine the quality of information that warrants the restriction, or that suggests the means for opting out. As a result, such restrictions on choice set may not promote autonomy. And, this is an effective threat. For example, even when people know that using a prediction rule is more accurate than subjective judgment, they defect from the rule when offered the opportunity. The lesson is clear: if you want people to make accurate assessments, then, once they identify the most reliable strategy, it is best that the option to defect or ''opt out'' of the strategy not be made salient. And, when possible, the institutional arrangement should allow the agent to opt out, without significant penalty. Even this measure need not restrict liberty. The agent could freely sign on to just such a ''no defection'' option. But it should be noted that *ex ante* restrictions on information and choice, when necessary, should be imposed with great care. Ideally, information relevant to the warrant for the agent's action should be effectively accessible. These are complicated issues,[52] and implementing

[52] Richard Arneson has anticipated a number of complications that involve discipline, defection, and self-binding. Arneson notes that quite different restrictions on liberty arise from 'preventing someone from doing what he already wants to do, and preventing someone from ever entertaining the option of doing a thing and forming a desire for it by prior restriction of the choice set'. (R. Arneson, 'Paternalism, Utility, and Fairness', in G. Dworkin (ed). *Mill's On Liberty: Critical Essays* (Lanham, MD: Rowman & Littlefield, 1997), pp. 83–114; quoted passage from pp. 107–108.)

liberty-respecting remedies may be challenging, but those measures already implemented provide a basis for optimism.

VII. CONCLUSION

Poor individual judgment is an extremely costly consequence of life in a democratic society. If this is indeed the necessary price of freedom, nearly everyone believes that this price is well worth paying. But opt-in, opt-out strategies exact no such cost. And in any case, it should also be admitted that no one has really done an accounting. The important message of a half-century of psychological research is that we need not simply accept the effects of poor individual judgment. We can improve, and we can do so in a way that satisfies our implicit rationales, our aims and priorities, even when those preferences are ill-formed. We make personal decisions of finance and health, and social decisions about political participation and trust. But accurate decisions about these matters typically demand reliable policies. The costly, sporadic and already harried attention of individuals is simply not up to the task. Anyone who assumes the adequate efficiency of debiasing through individual training is either ignoring the magnitude of institutional intervention required for such educational programs, or ignoring the cognitive costs to the individual of correcting such biases. Either way, these strategies of individual training do not take seriously the best available science.

As an instrument, decision-debiasing promotes an agent's autonomy by enhancing the accuracy of the agent's judgment, without imposing a substantial conception of the good. It promotes the agent's autonomy by intervening when the agent's decision is not one that, if fully informed and cognitively unbiased, the agent would have made. A government may arrange a system of debiasing instruments that doesn't micromanage the substantial goals of autonomous agents. In the end, the supplanting of individual judgment with institutional directives is not, by itself, paternalistic; ego-bruising maybe, but not paternalistic. If we are mature enough to accept with grace the narcissistic injury of being told to do something that we

would want to do anyway, we will not only reduce harm to self and others, but promote both autonomy and welfare.[53]

Philosophy Department and the Parmly Hearing Institute
Loyola University Chicago
6525 North Sheridan Road
Chicago, IL 60626
USA

[53] I would like to thank Paul Abela, Tom Carson, John Christman, Joseph Mendola, and Abe Schwab, for comments and conversations on this paper. I am especially indebted to Heidi Malm, who kindly provided detailed comments on two earlier drafts of this paper. I am also grateful to two anonymous referees for this journal, who supplied extremely specific and constructive comments.