Midwest Studies in Philosophy, XXXI (2007)

Philosophical Thought Experiments, Intuitions, and Cognitive Equilibrium

TAMAR SZABÓ GENDLER

I. INTRODUCTION

It is a commonplace that contemplation of an imaginary particular may have cognitive and motivational effects that differ from those evoked by an abstract description of an otherwise similar state of affairs. In his *Treatise on Human Nature*, Hume ([1739] 1978) writes forcefully of this:

> There is a noted passage in the history of Greece, which may serve for our present purpose. Themistocles told the Athenians, that he had form'd a design, which wou'd be highly useful to the public, but which 'twas impossible for him to communicate to them without ruining the execution, since its success depended entirely on the secrecy with which it shou'd be conducted. The Athenians, instead of granting him full power to act as he thought fitting, order'd him to communicate his design to Aristides, in whose prudence they had an entire confidence, and whose opinion they were resolv'd blindly to submit to. The design of Themistocles was secretly to set fire to the fleet of all the Grecian commonwealths, which was assembled in a neighbouring port, and which being once destroy'd wou'd give the Athenians the empire of the sea without any rival. Aristides return'd to the assembly, and told them, that nothing cou'd be more advantageous than the design of Themistocles but at the same time that nothing cou'd be more unjust: Upon which the people unanimously rejected the project. (*Treatise* II.iii.6.3)

This anecdote, Hume reports, was shocking to his contemporary, the widely-read French historian Charles Rollin, who found it astounding that the Athenians would reject—merely on grounds of injustice—a strategy so "advantageous" that it would give them "the empire of the sea without any rival." Indeed, Rollin suggests, the episode is "one of the most singular that is any where to be met with," revealing a truly astonishing sense of justice among the Athenian people.

Hume's own interpretation is rather more mundane:

> For my part I see nothing so extraordinary in this proceeding of the Athenians. . . . [T]ho' in the present case the advantage was immediate to the Athenians, yet as it was known only under the general notion of advantage, without being conceiv'd by any particular idea, it must have had a less considerable influence on their imaginations, and have been a less violent temptation, than if they had been acquainted with all its circumstances: Otherwise 'tis difficult to conceive, that a whole people, unjust and violent as men commonly are, shou'd so unanimously have adher'd to justice, and rejected any considerable advantage. (*Treatise* II.iii.6.4)

Hume's diagnosis has a straightforward corollary. When two options are presented abstractly, the choice made between them may go one way; presented under some "particular idea" that "influence[s]" the "imagination," the choice made between them may go the other. Engagement of the cognitive mechanisms associated with vivid imagining may lead a subject to reverse a prior commitment, selecting as preferable the option previously rejected, and shunning the option previously embraced.

Philosophical thought experiments, I will suggest, exploit exactly the discrepancy that led to Rollin's perplexity and Hume's insight. In the remainder of this article, I will explore three corollaries of this central suggestion. First, that by presenting content in a suitably concrete or abstract way, thought experiments recruit representational schemas that were otherwise inactive, thereby evoking responses that may run counter to those evoked by alternative presentations of relevantly similar content. Second, that exactly because they recruit heretofore uninvolved processing mechanisms, thought experiments can be expected to produce responses to the target material that remain in disequilibrium with responses to the same material under alternative presentations, so that a true sense of cognitive equilibrium will, in many cases, prove elusive. And finally, that when thought experiments succeed as devices of persuasion, it is because the evoked response becomes dominant, so that the subject comes (either reflectively or unreflectively) to represent relevant non-thought experimental content in light of the thought experimental conclusion. In each case, I will present some recent results from psychology and related disciplines that support the interpretation I am advancing.

II. COGNITIVE UNDERPINNINGS

Nearly a century of empirical investigation has confirmed the extent to which tasks with the same formal structure but different contents may prompt different rates of success, presumably because the alternate framings activate different processing mechanisms. In this section, I will review some of the literature that has been taken by psychologists to establish this claim decisively. These cases provide a useful foil to the philosophical examples to be discussed in the remainder of the article, since it is straightforward to isolate their formal from their content properties, and straightforward to ascertain what a correct response amounts to. The survey is intended to be suggestive, not comprehensive, and for those even moderately familiar with the literature, there is unlikely to be anything of novelty. Its main purpose is to make vivid to those unfamiliar with this research program some of the striking ways that content effects can enable or inhibit reasoning skill.

Though tacit recognition of such effects goes back millennia (see section IV for a 3,000-year-old example) and explicit recognition goes back at least centuries (cf. Hume above), modern study of the phenomenon can be dated to the work by E. L. Thorndike and his students in the third decade of the last century. In 1922, Thorndike published an article entitled "Effect of Changed Data on Reasoning," in which he described a series of studies that involved presenting students with familiar algebra problems. Across subjects, the structures of the problems were held constant; the only differences were in the symbols embedded within them. So, for instance, one group confronted equations whose variables were indicated by x and y, while those in the second group faced structurally identical equations whose variables were indicated by b_1 and b_2. The results of these small changes were dramatic: Error rates for the first group were six percent; error rates for the second were twenty-eight percent. Similar results were obtained by changing x^2 to 4^2, or a and x to r_1 and r_{11}. (Thorndike 1922, 36). Thorndike's conclusion was sweeping. He maintained that "any disturbance whatsoever in the concrete particulars reasoned with will interfere somewhat with the reasoning, making it less correct or slower or both" (Thorndike 1922, 33).[1]

Six years later, his student Minna Cheves Wilkins undertook a dissertation-length study of the issue, concluding that the "ability to do formal syllogistic reasoning is very much affected by a change in the material reasoned about." She presented subjects with a range of syllogistic tasks, asking them to judge whether certain conclusions followed from certain pairs of premises. Some involved terms that were "familiar and concrete" ("Some of the girls in the chorus wear their hair braided; all the girls in the chorus wear their hair bobbed; therefore . . ."); others involved symbols ("All x's are z's; all x's are y's; therefore . . ."). Yet others involved complicated nonsense terms ("No juritobians are cantabilians; no cantixianti are

1. Thorndike explained these results in strictly associationist terms: He held that "the mind is ruled by habit throughout" with reasoning being no more than "the organization and cooperation of habits" (Thorndike 1922, 33). Inheritors of his research program have tended to reject Thorndike's explanation of the mechanisms involved, but have continued to observe similar results.

cantabilians; therefore . . .") or terms with which the subjects had antecedent views about the relations among the terms ("If New York is to the right of Detroit; and Chicago is to the left of New York; then . . ."). Across subjects, results were quite consistent[2]: "Most items increase in difficulty as the material is changed from familiar to symbolic, etc., but a few items representing very common fallacies are much less difficult in symbolic material than in familiar" (Wilkins 1928, 52–77).

In the eight decades following Wilkins's and Thorndike's pioneering work, much light has been shed on which sorts of embeddings facilitate and which impede reasoning. Though the nuances are manifold, Wilkins's fundamental observation—that subjects' tendency to reason validly is typically improved when materials are presented with familiar content, though there are also cases where familiar content may interfere with their ability to identify valid structures—has been borne out. In cases where subjects are asked to attend to formal properties alone, the presence of certain sorts of content seems to enhance or inhibit their ability to draw appropriate conclusions on the basis of structural features.

Much of the research demonstrating these sorts of interference effects has made great use of two well-known paradigms: syllogism tasks (described in this paragraph) and Wason selection tasks (described below). In the first, subjects are presented with a set of premises, and asked to determine whether a particular conclusion follows logically from them. Stimuli vary along two dimensions: Some of the reasoning patterns are valid whereas others are invalid; and some of the conclusions are independently plausible whereas others are independently implausible. Presented with such stimuli, subjects consistently exhibit *belief-bias*: Structurally identical valid inferences are far less likely to be judged valid when their conclusions are implausible ("some vitamin tablets are not nutritional") than when their conclusions are plausible ("some highly trained dogs are not police dogs"); structurally identical invalid inferences are far less likely to be judged invalid when their conclusions are plausible than when their conclusions are implausible.[3] (cf. Evans et al. 1983, reviewed in Evans 2003.)

In the second, the Wason selection task (Wason 1966), subjects are presented with four cards and told that each card has an A-type feature (say, a number) on one side and a B-type feature (say, a letter) on the other. The subject is then presented with a (material) conditional statement that takes the following form: "If a card's F-feature is x, then its G-feature is y" and asked which cards she would

2. Wilkins was careful to note that there were individual differences among her subjects; some provided correct answers in (nearly) all cases. In recent years, these differences have been explored in detail, most notably by Keith Stanovich and Richard West (see, e.g., Stanovich and West 2000; cf. also Epstein et al. 1996).

3. Recent suggestive fMRI data may provide clues about the associated functional neuroanatomy. Studies by Vinod Goel and colleagues suggest "the engagement of a left temporal lobe system during belief-based reasoning and a bilateral parietal lobe system during belief-neutral reasoning." Their data suggest that "activation of right lateral prefrontal cortex was evident when subjects inhibited a prepotent response associated with belief-bias and correctly completed a logical task, a finding consistent with its putative role in cognitive monitoring. By contrast, when logical reasoning was overcome by belief-bias, there was engagement of the ventral medial prefrontal cortex, a region implicated in affective processing" (Goel and Dolan 2003, B11; cf. also Goel et al. 2000).

need to turn over to verify the statement's truth. The first card shows an instance of an F-feature that is *x* (F/*x*); the second shows an instance of an A-feature that is not *x* (F/not-*x*); the third shows an instance of a G-feature that is *y* (G/*y*); the fourth shows an instance of a G-feature that is not *y* (G/not-*y*). The appropriate response to such a question is to turn over exactly two cards: the first (F/*x*) card and the fourth (G/not-*y*) card.

Presented with certain abstract versions of the task, subjects tend to perform poorly. If, for example, subjects are asked to verify the (material conditional) statement "if there is an A on one side, there is a 3 on the other" for the set of cards pictured below, fewer than ten percent correctly turn over exactly the "A" and the "7"; instead, they typically turn over the "A" and the "3," or the "A" only.

A D 3 7

If the task is altered slightly, however, so that subjects are presented with the same set of four cards, but with the instruction "if there is an A on one side, there is not a 7 on the other," subjects nearly universally turn over the correct pair of cards. This tendency to match response to cue (note that in the first case the consequent mentioned "3," whereas in the second case, it mentioned "7") goes by the name *matching bias*.[4]

Interestingly, subjects are far less prone to matching bias in certain cases that embed the selection task within a practical realm and make appeal to some sort of deontic rule.[5] So, for example, success rates are extremely high when the sentences to be verified resemble this one: "if a person is drinking beer, then the person must be at least 21 years of age." In such cases, a vast majority of subjects (correctly) turn over the "beer" and the "16" (and not, in parallel to the previous case, the "beer" and the "21") (cf. Griggs and Cox, 1982).[6]

Beer Coke 21 years 16 years

4. For an overview of the enormous body of research conducted using this paradigm, see Evans (1998) and relevant articles mentioned in its bibliography. For a fascinating discussion of a process of training subjects to inhibit matching bias, along with intriguing data about its possible neural underpinnings, see Houdé et al. (2000).

5. At least five features seem consistently to produce increased speed and accuracy in Wason-style tasks: the use of concrete and meaningful terms in articulating the rule and describing the cards; presenting the task as one of determining a rule violation rather than the truth or falsity of a statement; embedding the task within the context of a scenario where the subject is given a particular role to play; providing the subject with a rationale or justification for the rule; and relating the two rule components in a meaningful way (Dominowski 1995, 45).

A number of hypotheses have been offered to explain the patterns of response, among them that certain embedded tasks trigger a pragmatic reasoning schema (cf., e.g., Cheng and Holyoak, 1985; Cheng et al. 1986), that they trigger a modular social exchange algorithm (cf., e.g., Cosmides 1989; Gigerenzer and Hug 1992), and that different mental models are activated by different presentations of conditional content (cf., e.g., Johnson-Laird and Byrne 2002); others have argued on Bayesian grounds that typical reasoning patterns on Wason-style tasks are actually rational (cf., e.g., Oaksford and Chater 1994, 1996). None of these accounts has been universally accepted, and it seems likely that the full story will turn out to be quite complicated.

6. Interestingly, the effect is reduced in cases where the pairing is judged as unlikely: Fewer subjects turn over the final card if it reads "12 years" and fewer still if it reads "4 years" (Kirby 1994).

Related content-based effects can be found in tasks involving a wide array of different sorts of forced choices. In a 1994 study, for example, Veronika Denes-Raj and Seymour Epstein presented subjects with pairs of platters containing varying numbers of red and white jelly beans. Subjects were told that they would win $1 for each trial in which they drew a red jelly bean, and then given a choice about which of the two platters they would prefer to draw from blindly. The first platter always contained one red jelly bean and nine white beans, while the other contained 100 beans total, with the proportion of red to white ranging from 9 : 100 (9 red and 91 white) to 5 : 100 (5 red and 95 white.) Each platter was labeled with an index card clearly indicating the percentage of red jelly beans that it contained (ten percent, nine percent, eight percent, etc.)

Despite the presence of the monetary incentive and the explicit information about relative likelihood of success, well over half the subjects chose the 9 : 100 and 8 : 100 platters over the 1 : 10 platter, and—astonishingly—more than a quarter chose the 5 : 100 platter over the 1 : 10. Overall, more than eighty percent of subjects made at least one nonoptimal choice in the five trials each faced. When asked about their selections, "subjects reported that although they *knew* the probabilities were against them, they *felt* they had a better chance when there were more red beans . . . They made statements such as, 'I picked the ones with more red jelly beans because it looked like there were more ways to get a winner, even though I knew there were also more whites, and that the percents were against me' " (Denes-Raj and Epstein 1994, 819, 823.)

The literature on heuristics and biases is replete with such examples.[7] Readers are presumably familiar with many of Daniel Kahneman and Amos Tversky's famous cases. In the Linda-the-bank-teller case, for example, subjects are presented with a description of an imaginary character, Linda, that reads as follows:

> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. (Tversky and Kahneman 1983, 297)

Subjects are then presented with a set of eight statements about Linda, and are asked to rank them in order of likelihood. Among the statements are the following:

> Linda is a bank Teller. (T)
> Linda is a bank Teller and is active in the Feminist movement. (T + F)

Even when subjects are highly educated, even when they are graduate students in a decision science program, even when they are asked to bet money on their choice, even when they are explicitly reminded that "bank teller" does not mean

7. Three classic collections are Kahneman et al (1982), Kahneman and Tversky (2000), and Gilovich et al (2002).

"mere bank teller," even when the logical relations between the two statements are made transparent—even in all these cases—there is a striking tendency for subjects to choose T + F as more probable than T (cf. Tversky and Kahneman 1983; Crandall and Greenfield 1986; Epstein et al 1999). As Stephen Jay Gould remarks in his own reminiscence about encountering the case: "I know that [T + F] is least probable, yet a little homunculus in my head continues to jump up and down, shouting at me—'but she can't just be a bank teller; read the description'" (Gould 1991, 469).

The same goes for each of the cases discussed above. Even subjects who regularly provide correct answers to abstract match-violating Wason tasks are consistently faster (and consistently more accurate under conditions of cognitive load) at solving suitably matched or embedded tasks. Similar results obtain in the case of belief-matched (as opposed to belief-mismatched) syllogisms, and in tasks like Denes-Raj and Epstein's number/proportion task. Everyone—even those who are ultimately able to override (or endorse for the right reasons) the inclination that leads to error (or success) in the cases under consideration—feels the pull of the competing response.

One promising framework for explaining these patterns of response is the family of theories that go by the name *dual systems* accounts. According to such accounts, there are at least two clusters of subsystems involved in mental processing: one associative and instinctive, operating rapidly and automatically; the other rule-based and regulated, operating in a relatively slow and controlled fashion. Numerous formulations of this distinction have been proposed—diverging in important details that matter a great deal for a number of important debates. But for our purposes, their commonalities are more important than differences. Two examples will suffice to give a flavor of such accounts.[8]

According to Paul Sloman's *Two Systems* model, human reasoning makes use of both an *Associative* and a *Rule-Based System*. The Associative System operates on principles of similarity and contiguity; takes personal experience as its source of knowledge; operates on concrete and generic concepts, images, stereotypes, and feature sets; makes use of relations of association that serve as soft constraints; exhibits processing that is reproductive but capable of similarity-based generalization; uses overall feature computation and constraint satisfaction; is automatic; and is exemplified by functions such as intuition, fantasy, creativity, imagination, visual recognition, and associative memory. By contrast, the Rule-Based System operates on principles of symbol manipulation; takes language, culture, and formal systems as its sources of knowledge; operates on concrete, generic, and abstract concepts, abstracted features, and compositional symbols; makes use of causal, logical, and hierarchical relations that serve as hard constraints; exhibits processing that is productive and systematic; uses abstractions of relevant features; is strategic; and is exemplified by functions such as deliberation,

8. For additional representative discussions, see the articles collected in Chaiken and Trope (1999), Evans (2003, 2008), Evans and Over (1996), Gigerenzer and Regier (1996), Hinton (1990), Smolensky (1988), Stanovich (1999), and Stanovich and West (2000). For intriguing early discussions, see James (1890), Piaget (1929), and Neisser (1963).

explanation, formal analysis, verification, ascription of purpose, and strategic memory (Sloman 1996, 7).

According to Seymour Epstein's cognitive–experiential self-theory, or CEST (cf. Epstein 1990),

> [I]ndividuals apprehend reality by two interactive, parallel processing systems. The *rational system*, a relative newcomer on the evolutionary scene, is a deliberative, verbally mediated, primarily conscious analytical system that functions by a person's understanding of conventionally established rules of logic and evidence. The *experiential system*, which is considered to be shared by all higher order organisms (although more complex in humans), has a much longer evolutionary history, operates in a holistic, associationist manner, is intimately associated with the experience of affect, represents events in the form of concrete exemplars and schemas inductively derived from emotionally significant past experiences, and is able to generalize and construct relatively complex models for organizing experience and directing behavior by the use of prototypes, metaphors, scripts, and narratives. (Denes-Raj and Epstein 1994, 819)

As Daniel Gilbert points out, however, there is nothing sacred about the "dual" in dual processing. He writes,

> [T]he neuroscientist who says that a particular phenomenon is the result of two processes usually means to say something unambiguous [about] . . . the activities of two different brain regions . . . [but] dry psychologists who champion dual-process models are not usually stuck on two. Few would come undone if their models were recast in terms of three processes, or four, or even five . . . claims about dual processes in dry psychology are not so much claims about how many processes there *are*, but claims about how many processes there *aren't*. And the claim is this: There aren't one. (Gilbert 1999, 3–4)

For our purposes, the moral is simply this. Decades of research in cognitive psychology have demonstrated that when content is presented in a suitably concrete or abstract way, this may result in the activation or fortification of a representational schema that was otherwise inactive or subordinate; the result of this may be to evoke responses that run counter to those evoked by alternative presentations of relevantly similar content. So far from being an anomalous or idiosyncratic feature of arcane or unusual cases, the discrepancy described in our opening story is—in fact—a central feature of our mental lives.

III. THOUGHT EXPERIMENTS AND ELUSIVE EQUILIBRIUM

So far, we have been considering cases where it is clear what the right answer is, and where (at least in some cases) we have a fairly systematic understanding of the sorts of factors that lead subjects astray. When subjects turn over the A and the 3

in the A-D-3-7 task described above, they make a mistake; when they turn over the A and the 7, they do not. When the sentence to be confirmed is: "If there is an A on one side, there is a 3 on the other," even subjects who ultimately respond correctly are somewhat drawn to the card with the 3; when the sentence to be confirmed is: "If there is an A on one side, there is not a 7 on the other," even subjects who face persistent difficulties with the A-3 formulation are easily able to turn over the correct cards. Likewise with the syllogism tasks: When subjects conclude that an invalid inference with a true conclusion is valid, they err—and when they err, it tends to be because an independent judgment about the truth or falsity of the conclusion interferes with their judgment concerning the inference's validity. Like optical illusions, these cognitive illusions seem to be artifacts of deep features of our cognitive architecture: The "little homunculus in [our] head" will continue to "jump up and down," whether or not we can train ourselves to discount its cries when non-homuncular reasoning is called for. Just as we cannot simply talk ourselves out of seeing Müller-Lyer lines as different in length, we cannot simply talk ourselves out of feeling drawn toward turning over the 3.[9]

What implications does this have for philosophical methodology? It seems to me that the implications are both liberating and disturbing—and that these implications are two sides of the same coin. For if something akin to dual processing lies at the root of most human reasoning, then a philosophical theory may be correct even if we consistently and resiliently react to specific cases in ways that run counter to the theory's predictions. This introduction of an additional degree of freedom into the enterprise of philosophical explanation may introduce a feeling of vertigo.[10]

Recent neuroimaging work on moral reasoning has brought this challenge to the fore in the context of the "trolley problem." Though most readers are presumably familiar with this widely discussed example, here is Judith Jarvis Thomson's 1985 presentation of the scenario:

> Some years ago, Philippa Foot drew attention to an extraordinarily interesting problem (Foot 1978). Suppose you are the driver of a trolley. The trolley rounds a bend, and there come into view ahead five track workmen, who have been repairing the track. The track goes through a bit of valley at that point, and the sides are steep, so you must stop the trolley if you are to avoid

9. Habits of attention may mitigate the effects somewhat; one can learn to approach questions of validity by automatically mentally substituting content-neutral expressions for content-distracting ones. I discuss this in more detail in the context of philosophical thought experiments in section IV below.

10. For a related discussion of these matters that comes to somewhat similar conclusions, see Sunstein (2005). He writes, "In short, I believe that some philosophical and philosophical analysis, based on exotic moral dilemmas, is inadvertently and even comically replicating the early work of Kahneman and Tversky: uncovering situations in which intuitions, normally quite sensible, turn out to misfire. The irony is that while Kahneman and Tversky meant to devise cases that would demonstrate the misfiring, some philosophers develop exotic cases with the thought that the intuitions are likely reliable and should form the building blocks for sound moral judgments. An understanding of the operation of heuristics offers reason to doubt the reliability of those intuitions, even when they are very firm" (Sunstein 2005). I suspect I am a bit more sanguine than Sunstein about the possibility of intuition-driven moral theorizing—but only a bit.

> running the five men down. You step on the brakes, but alas they don't work. Now you suddenly see a spur of track leading off to the right. You can turn the trolley onto it, and thus save the five men on the straight track ahead. Unfortunately, Mrs. Foot has arranged that there is one track workman on that spur of track. He can no more get off the track in time than the five can, so you will kill him if you turn the trolley onto him. Is it morally permissible for you to turn the trolley?
> Everyone to whom I have put this hypothetical case says, Yes, it is. (Thomson 1985, 1395)

In the remainder of the article, Thomson runs through a number of simple and complex cases that evoke intuitions of various kinds, attempting to identify systematic principles that underlie those intuitions; among the cases she considers is this one:

> Consider a case—which I shall call *Fat Man*—in which you are standing on a footbridge over the trolley track. You can see a trolley hurtling down the track, out of control. You turn around to see where the trolley is headed, and there are five workmen on the track where it exits from under the footbridge. What to do? Being an expert on trolleys, you know of one certain way to stop an out-of-control trolley: Drop a really heavy weight in its path. But where to find one? It just so happens that standing next to you on the footbridge is a fat man, a really fat man. He is leaning over the railing, watching the trolley; all you have to do is give him a little shove, and over the railing he will go, onto the track in the path of the trolley. Would it be permissible for you to do this? Everyone to whom I have put this case says it would not be. (Thomson 1985, 1409)

As readers of the popular press are no doubt aware, recent neuroimaging and lesion work has suggested one explanation for this difference in response: It appears that whereas the original trolley case produces increased neural activity in "higher cognitive" regions of the brain, cases such as fat man (where the imagined action is "up close and personal") produce increased neural activity in "emotional/social" regions (cf. Greene et al. 2001). Intriguing confirmation of this suggestion can be found in recent work by Antonio Damasio suggesting that subjects with ventromedial prefrontal cortex damage (damage associated, among other things, with impaired emotional processing) are more than twice as likely as controls to consider it morally acceptable to push the fat man (or to suffocate a crying baby in order to save a group of people who are hiding) (Koenigs et al. 2007.)

All of this is fully compatible with there being a genuine deep moral difference between the two acts—deep enough to render the one morally mandatory and the other morally prohibited. Nothing that I have said here or elsewhere should be taken to deny the possibility that—as Mill writes at the beginning of *Utilitarianism*—"whatever steadiness and consistency our moral beliefs have attained has been mainly due to the tacit influence of a standard not yet recognized" (Mill [1861] 2001, 3).

That said, it is worth taking seriously other work that suggests that intuitions about such cases may vary along dimensions that are (presumably) completely morally irrelevant. Psychologist David Pizarro presented subjects with "fat man" trolley cases that differed only in the nature of the sacrifice involved: In the one case, a man named Chip Ellsworth III could be thrown off a bridge to stop a trolley hurtling toward 100 members of the Harlem Jazz Orchestra; in the other, a man named Tyrone Peyton could be thrown off to save 100 members of the New York Philharmonic.[11] Subjects were significantly more likely to consider it morally acceptable to sacrifice Chip to save the Harlem Jazz Orchestra than to sacrifice Tyrone to save the New York Philharmonic (presumably an overcorrection of an initial instinctively racist response) (Pizarro et al., manuscript).

Whether or not there is a moral difference between the original trolley case and the fat man case, it seems clear that there is no moral difference between sacrificing Tyrone and sacrificing Chip.[12] But if our only basis for thinking that there is a moral difference between fat man and original trolley is that subjects tend to respond differently to them, we should be disturbed to discover that parallel differences can be evoked by what seem clearly to be morally irrelevant differences.[13]

Even more disturbingly, additional work by Pizarro suggests that subjects' responses to moral dilemmas can be made to vary through techniques of unconscious priming. Presented with otherwise identical scenarios in which American (or Iraqi) troops cause anticipated but unintentional collateral damage to Iraqi (or American) civilians, politically conservative subjects are significantly more likely to judge the American-on-Iraqi damage to be morally acceptable than the other way around, whereas politically liberal subjects make precisely the opposite judgment. But Pizarro discovered that these effects can be induced simply by prompting subjects to unscramble sentences containing terms associated either with patriotism or with multiculturalism[14]: Subjects primed with patriotism terms tend to assess the America–Iraq case in ways akin to conservatives, whereas subjects primed with multiculturalism terms respond much like liberals (Pizarro et al., manuscript).[15]

11. Non-American readers may be helped by learning that the name "Chip Ellsworth III" evokes images of a wealthy white man, whereas "Tyrone Peyton" evokes images of a man of African descent; likewise, the New York Philharmonic is an elite largely white and Asian orchestra, whereas the Harlem Jazz Orchestra is associated with the African-American community.

12. Of course, this judgment is itself grounded in some sort of intuitive judgment. For discussion of the unavoidability of appeal to intuition in philosophical reasoning, see Bealer (1998), Goldman (2007), Pust (2000), Sosa (2007a, 2007b), and Williamson (2005).

13. Admittedly, the differences that Pizarro observes are decidedly less extreme than those evoked by the original trolley/fat man contrast. (But there are good naturalistic reasons to expect this.)

14. "Scrambled sentence" tasks—in which subjects are presented with a series of word clusters that they are asked to form into sentences (e.g., "flies high the olives flag" or "ribbons very dogs are loyal")—are a standard technique in social psychology for "priming" unconscious associations.

15. Pizarro's work is representative of a large research program in contemporary psychology exploring the status and source of moral intuition. See, for example, deWaal (1996), Haidt (2001), Haidt and Joseph (2004), Hauser (2006), and sources cited therein.

Nor is there anything special about moral intuitions in this regard. Take the case of Keith Lehrer's Mr. Truetemp:

> Suppose a person, whom we shall name Mr. Truetemp, undergoes brain surgery by an experimental surgeon who invents a small device which is both a very accurate thermometer and a computational device capable of generating thoughts. The device, call it a tempucomp, is implanted in Truetemp's head so that the very tip of the device, no larger than the head of a pin, sits unnoticed on his scalp and acts as a sensor to transmit information about the temperature to the computational system of his brain. This device, in turn, sends a message to his brain causing him to think of the temperature recorded by the external sensor. Assume that the tempucomp is very reliable, and so his thoughts are correct temperature thoughts. All told, this is a reliable belief-forming process. Now imagine, finally, that he has no idea that the tempucomp has been inserted in his brain, is only slightly puzzled about why he thinks so obsessively about the temperature, but never checks a thermometer to determine whether these thoughts about the temperature are correct. He accepts them unreflectively, another effect of the tempucomp. Thus, he thinks and accepts that the temperature is 104 degrees. It is. Does he know that it is? (Lehrer 1990, 163–64)

Jonathan Weinberg and colleagues have discovered that "(1) willingness to attribute knowledge in the Truetemp Case increases after being presented with a clear case of non-knowledge, and (2) willingness to attribute knowledge in the Truetemp Case decreases after being presented with a clear case of knowledge" (Swain et al., manuscript, 1). John Hawthorne and I demonstrate related sorts of shiftiness in fake barn cases (Gendler and Hawthorne 2005[16]). And Joshua Knobe and Shaun Nichols have found presentation-dependent differences in judgments of free will and moral responsibility: "When subjects are asked the abstract question whether agents in [a deterministic universe] are fully morally responsible, 86% say that they are not: no agent can be fully morally responsible for doing what he is fully determined to do. However, when a dastardly deed is attributed with a wealth of detail to a particular agent in [that world], and those same subjects are asked whether that agent is then fully morally responsible, 72% report that in their view he is!" (Sosa, 2007a, 104, discussing Nichols and Knobe, forthcoming).

Though specific stories can be told about each of the cases discussed, overall, the accumulated implications can seem dizzying.[17] If intuitions cannot serve as a

16. For an overview of the issue of intuitions and epistemology, see Alexander and Weinberg (2007).

17. In addition to concerns about intrasubjective variation, there are also grounds for unease about intersubjective variation. Widely touted work by Jonathan Weinberg, Stephen Stich, and Shaun Nichols seems to suggest that there are important cultural differences in how subjects respond to some of the central examples in the epistemological literature (Weinberg et al. 2001). Similar worries are raised in an intracultural context by Robert Cummins, who notes, "It is commonplace for researchers in the current Theory of Content to proceed as if [Twin Earth] intuitions were undisputed . . . Nor is the reason for this practice far to seek. The Putnamian

fixed point for philosophical theorizing, then much that has been widely taken as philosophical orthodoxy may be up for grabs. On the basis of related considerations, for example, Brian Weatherson writes,

> Intuitively, Gettier cases are instances of justified true beliefs that are not cases of knowledge. Should we therefore conclude that knowledge is not justified true belief? Only if we have reason to trust intuition here. But intuitions are unreliable in a wide range of cases. And it can be argued that Gettier intuitions have a greater resemblance to unreliable intuitions than to reliable intuitions. What's distinctive about the faulty intuitions, I argue, is that respecting them would mean abandoning a simple, systematic and largely successful theory in favour of a complicated, disjunctive and idiosyncratic theory. So maybe respecting the Gettier intuitions was the wrong reaction, we should instead have been explaining why we are so easily misled by these kinds of cases. (Weatherson 2003, 1)

Though careful work regarding particular cases may allow the reclaiming of some aspects of traditional intuition-based methodology,[18] the accumulated evidence reviewed in sections 1 and 3 suggests that the utility of philosophical thought experiments may lie in another direction. It is to this issue that I turn in the final section.

IV. THOUGHT EXPERIMENTS AS DEVICES OF PERSUASION

A common insight lies at the heart of both Kantian and utilitiarian moral theorizing: To reason in accord with the dictates of morality is to view oneself as unexceptional. Immanuel Kant's Categorical Imperative requires that "I should never act except in such a way that I can also will that my maxim should become a universal law" (Kant [1785] 1981, 402). That is, morality requires that personal desires be filtered through a universalizing lens: My own desires may serve as bases for willed action only if I can at the same time coherently will that others in similar circumstances would act in the way that I am choosing to act. Despite important differences between the views, a similar core insight lies at the heart of Jeremy Bentham's famous utilitarian formulation that "everybody [is] to count for one, nobody for more than one" (Bentham, cited by Mill [1861] 2001). Here, too, one's own interests may legitimately enter into decision-making only insofar as they are weighed equally alongside the interests of others: First-person exceptionalism is morally prohibited.

. . . take on these cases is widely enough shared to allow for a range of thriving intramural sports among believers. Those who do not share the intuitions are simply not invited to the games . . . [I]t is all too easy for insiders to suppose that dissenters just do not understand the case. If we are honest with ourselves, I think we will have to confront the fact that selection effects . . . are likely to be pretty widespread in contemporary philosophy" (Cummins 1998).

18. For some reflections on this issue, see Weinberg et al (manuscript) and sources cited therein.

For the purposes of discussion in this section, let's call a *moral stance* one that prohibits first-person exceptionalism. How might one make this stance cognitively available to the subject at moments of moral decision-making?

In answering this question, it is worth reminding ourselves that among the most resilient of our cognitive tendencies is exactly the tendency to hold ourselves to different standards than we hold others. So, for example, study after study has shown that "people overestimate the extent to which they personally are influenced by 'objective' concerns and/or overestimate the extent to which others are influenced by 'self-serving' concerns" (Pronin et al. 2004). As Emily Pronin notes in a review article, summarizing a wide range of recent work, "they assume that people who work hard at their jobs are motivated by external incentives such as money, whereas they claim that they personally are motivated by internal incentives" (Pronin 2006, 37–38); they consistently overestimate the likelihood that they will act generously or selflessly, while accurately predicting the ungenerosity and selfishness of others (whom they most likely turn out to resemble). Repeated studies have shown that "people on average tend to think they are more charitable, cooperative, considerate, fair, kind, loyal, and sincere than the typical person but less belligerent, deceitful, gullible, lazy, impolite, mean, and unethical." The same holds for specific predictions of behavior: "[P]eople generally think they are more likely than their peers to rebel in the Milgram obedience studies, cooperate in a prisoner's dilemma game, distribute collective funds equitably, and give up their seat on a crowded bus to a pregnant woman ( ) . . . [they] tend to believe they will resolve moral dilemmas by selecting the saintly course of action but that others will behave more selfishly" (Epley and Dunning 2000.) And they "tend to see their futures as overly rosy, to see their traits as overly positive, to take too much credit for successful outcomes and to disregard evidence that threatens their self esteem" (Pronin 2006, 37). It is no exaggeration to say that the tendency toward first-person exceptionalism is among the most widespread and pervasive of our tendencies toward bias.

This tendency finds powerful voice in the Biblical story of David and Bathsheba (2 Sam. 11–12).[19] David, who is King of Israel, is walking along the roof of his palace when he catches sight of an attractive woman—Bathsheba—washing herself nearby. Taken by her beauty, he has her brought to the palace, where he proceeds to lie with her, though she is married to another man. She becomes pregnant, and David arranges to have her husband Uriah sent to fight "in the forefront of the hottest battle . . . that he may be smitten and die." Uriah is killed, and David proceeds to take Bathsheba as his wife.

God is (understandably enough) rather displeased by David's behavior, and seeks to help him see the ways in which it is problematic. But God recognizes the deep human tendency toward first-person exceptionalism, and seeks a way to speak to David that will circumvent this tendency. So "the Lord sent Nathan unto David," and Nathan proceeds to tell David the following story:

19. Thanks to Tim Crane for pointing out to me the philosophical potential of this story in a related context.

> There were two men in one city; the one rich, and the other poor. The rich man had exceeding many flocks and herds: But the poor man had nothing, save one little ewe lamb, which . . . grew up together with him, and . . . was unto him as a daughter. And there came a traveler unto the rich man, and he spared to take of his own flock . . . but took the poor man's lamb, and dressed it for the man that was come to him.

When David hears this story, his "anger [is] greatly kindled against the man." He holds the man to be deserving of disapprobation and punishment, and says to Nathan: "As the Lord liveth, the man that hath done this thing shall surely die. And he shall restore the lamb fourfold, because he did this thing, and because he had no pity."

At this point, the circumstances have been set for the delivery of the punch line. Nathan says famously to David,

> Thou art the man . . . thou hast killed Uriah the Hittite with the sword, and hast taken his wife to be thy wife, and hast slain him with the sword of the children of Ammon . . .

With a shock of recognition, David reframes his understanding of the circumstances in which he has placed himself, and says to Nathan, "I have sinned against the Lord."

By framing the story so that David is not in a position to exhibit first-person bias with respect to what turns out to be his own actions, Nathan has enabled David to acknowledge a moral commitment that he holds in principle, but has failed to apply in this particular case. There is no ambiguity here about which commitment, on reflection, David endorses: The story he has been told is fully effective; it reshapes his cognitive frame, and brings him to view his own previous actions in its light.

Despite being relatively schematic, the story is a vivid one, engaging the reader's imagination as she hears about David's and Nathan's actions, and David's imagination as he hears of the behavior of the imaginary rich man who slays the poor man's sheep. Within the domain of philosophy, broadly construed, there is a tradition that emphasizes the capacity of the literary form to appropriately represent moral complexity, contrasting this with the tradition of austere philosophical theorizing. Martha Nussbaum maintains that "there may be some views of the world and how one should live in it . . . that cannot be fully and adequately stated in the language of conventional philosophical prose . . . but only in a language and in forms themselves more complex, more allusive, more attentive to particulars" (Nussbaum 1990, 3). Noting that there has been a "predominant tendency in contemporary Anglo-American philosophy . . . to ignore the relation between form and content . . . or . . . [to] treat[] style as largely decorative—as irrelevant to the stating of content," she emphasizes instead the "importance of taking style seriously in its expressive and statement-making functions" (Nussbaum 1990, 8).

While Nussbaum is surely right that the dominant tendency in Western philosophical theorizing has been one that holds form and content to be isolable in

these ways, there is also an important strand—even among the most austere of philosophical writing—that explicitly or tacitly acknowledges the force that presentational features can play. Even Kant, who held that "worse service cannot be rendered morality than that an attempt be made to derive it from examples" (Kant [1785] 1981, 408) gives some weight to this perspective. In the course of the *Grounding for the Metaphysics of Morals*, he famously formulates the Categorical Imperative in a number of different ways. Though he maintains that these "ways of representing the principle of morality are at bottom only so many formulas of the very same law," he remarks that "nevertheless there is a difference in them which is subjectively rather than objectively practical, viz., it is intended to bring an idea of reason closer to intuition (in accordance with a certain analogy) and thereby closer to feeling" (Kant [1785] 1981, 436).

Viewed in this light, moral and political philosophy have a secondary task that runs alongside the task of ascertaining what morality demands, namely, that of providing the reader with resources that enable him or her to make the perspective shift that the moral stance requires at the moment of moral decision-making. In this regard, one of the tasks of such philosophical inquiry is to identify images that can play the role that Nathan's story did with respect to David: images that will bring the readers to reframe their experience of some morally valenced situation, in such a way that their apprehension of the morally relevant features of it are re-experienced in light of the scenario presented. It is this role, I want to suggest, that is played by some of the most famous thought experiments in moral and political theorizing.

Take, for example, one of the most widely discussed aspects of John Rawls's enormously influential *A Theory of Justice*—his "device of representation" for thinking about the principles that would govern the basic structure of a just society. In the first chapter of the book, he introduces the famous example of the "original position"—a "purely hypothetical" situation where "no one knows his place in society, his class position or social status, nor does any one know his fortune in the distribution of natural assets and abilities, his intelligence, strength and the like." From behind this "veil of ignorance," principles are chosen that will regulate "the kinds of social cooperation that can be entered into and the forms of government that can be established." Just principles will be those that "free and rational persons concerned to further their own interests would accept in [such] an initial position of equality as defining the terms of their association" (Rawls [1971] 1999, section 3). In the remainder of the section, Rawls identifies two fundamental principles that he maintains would be adopted by subjects in such a circumstance: that each person has equal rights to certain basic liberties, and that inequalities exist only if their presence can reasonably be expected to benefit all, including those who are least well-off.

Or take a Rawls-inspired example from a recent blog entry by the philosopher Elizabeth Anderson (Anderson 2006). Anderson writes,

> Let's conduct a thought experiment. You have to play a mountain-climbing game. The higher you climb, the better off you are. Rarely, players climb solo. Most of the time, they climb in teams. The members of each team are

> connected by pulleys and gears in such a way that, if everyone climbs in a cooperative fashion, everyone in the team goes higher than if each just climbed the team rope in an uncoordinated way. The job of the team leaders—those highest on the rope—is to figure out how to get everyone to coordinate their climbing so as to get the maximum total lifting force for the whole team. However, depending on the gear setup, the lifting force of each member's step may accrue unequally to each team member. (In most setups, those at the top get lifted higher by any team member's step than anyone below.) The mountain face is swept by gales, although the winds tend to be milder at higher altitudes than at lower ones. Sometimes the gales blow you or even your whole team off your rope. Other times, the team leaders—those at the top of the team rope–eject you from the team and toss you off the team rope. If you are lucky, your mountain-climbing skills may be attractive enough to another team that they extend you a part of their team rope before you hit the ground. Or you will have family or friends who will toss you a safety rope to catch you on your way down. But you may not find a team with an open place on their rope that they will offer you, and you may not have family or friends willing to offer a rope, or the rope they are able to offer may be too frail to stop your fall.
>
> You don't know your initial place on your rope, nor which rope it is, nor your mountain-climbing skills, nor how well-off, benevolent, and numerous your family and friends are. In this state of ignorance, you get to choose some of the rules of the game you must play. Which rules would you prefer to play by? Here are your choices . . .

Anderson goes on to enumerate a number of options for how the rules of this "game" might be arranged—*Free Fall*, *Safety Net*, *Long Bungee Cord*, *Short Bungee Cord*, *Maximin*, *Strict Equality*, and *No Rules Dictatorship*—which differ in the extent to which the fates of players are yoked together. In each case, the cost of having some sort of protective mechanism in place is some reduction in the maximum altitude reached by the most successful climbers, with costs to the higher climbers proportionate to benefits to the lower climbers. (So, for example, on the rules of *Safety Net* "there is a safety-net placed somewhere between the ground and the lowest-altitude player that will catch you before you hit the ground . . . everyone will climb at a slightly slower pace than if the net were not there," whereas on the rules of *Long Bungee Cord* "in addition to a safety-net for those who never get going on a rope, you have a bungee cord anchored to a point on the mountain equal to your highest achieved altitude [that] . . . prevents you from falling more than 60% of the way down the mountain . . . everyone will climb at a modestly slower pace than if they were not supplied with the cord.") Anderson suggests that among these options, only *Maximin*, *Short Bungee Cord*, and *Long Bungee Cord* represent "credible" options, and reports a personal preference for playing "a game somewhere between *Long Bungee Cord* and *Short Bungee Cord* (Anderson 2006).

In both cases, I want to set aside the conclusions that the author draws on the basis of reasoning within the context of the scenario. That is, I want to set aside Rawls's suggestion that subjects in the original position behind the veil of igno-

rance would endorse the two principles of justice alluded to above, and Anderson's suggestion that only certain of the rule-sets enumerated are "credible options." Though these also raise interesting methodological issues, they are subsequent to the issues that I want to address here. What I want to think about is what we can learn about philosophical methodology as the result of thinking carefully about the cognitive effects of framing the question in the way that each does.

In Rawls's case, the scenario presented serves as a device of representation: It exhibits in vivid fashion the notion that Rawls takes to lie at the core of the concept of justice. In Anderson's case, the scenario presented serves as a tool for clarification: It encourages the reader to think through questions of risk and reward that Anderson takes to be common between her climbing game and the social structures governing resource distribution. In Anderson's cases, the example can play its intended role only if a certain isomorphism holds: If the trade-offs that are assumed to operate at the level of the game-rules do not hold in the context of the target subject matter, then judgments made in the context of the thought-experimental scenario will be irrelevant to the question they are intended to illuminate.

It is far from clear that this latter condition is met: It may well be that the target subject matter is structured quite differently from the subject matter of Anderson's scenario. Moreover, this may be a case where content matters: Perhaps risk adversity of the sort Anderson endorses is appropriate (in whatever sense) in the context of the game, but not in the larger-scale situation which Anderson intends the game to illuminate. Granting all of this, it is nonetheless worthwhile to think about the cognitive effects of presenting the scenario as it has been presented. My suggestion is that, to the extent that it is effective, it plays the role that Nathan's story of the rich man and the sheep played: It provides the subject with a powerful frame through which the target material—decisions about the appropriate structure for resource distribution in society—can be reconceptualized. It seeks to make the moral stance cognitively available at moments of moral decision-making.

A similar status—with corresponding caveats about whether the posited isomorphism holds—can be claimed for Judith Jarvis Thomson's famous violinist example (Thomson 1971), intended to provide the subject with a vivid reframing of how he thinks about the relation between the fetus and the mother:

> You wake up in the morning and find yourself back to back in bed with an unconscious violinist. A famous unconscious violinist. He has been found to have a fatal kidney ailment, and the Society of Music Lovers has canvassed all the available medical records and found that you alone have the right blood type to help. They have therefore kidnapped you, and last night the violinist's circulatory system was plugged into yours, so that your kidneys can be used to extract poisons from his blood as well as your own. The director of the hospital now tells you, "Look, we're sorry the Society of Music Lovers did this to you—we would never have permitted it if we had known. But still, they did it, and the violinist is now plugged into you. To unplug you would be to kill him. But never mind, it's only for nine months. By then he will have

> recovered from his ailment, and can safely be unplugged from you." Is it morally incumbent on you to accede to this situation? . . . What if the director of the hospital says, "Tough luck. I agree. But now you've got to stay in bed, with the violinist plugged into you . . . Because remember this. All persons have a right to life, and violinists are persons. Granted you have a right to decide what happens in and to your body, but a person's right to life outweighs your right to decide what happens in and to your body . . ."

Thomson's thought experiment "works" if it brings about a reframing of the subject's attitudes in the domain it is intended to illuminate—if he comes, either reflectively or unreflectively, to represent the question of the fetus–mother relationship in ways akin to those that he represents the violinist–patient relationship. Anderson's "works" if, when faced with decisions about whether to institute certain sorts of social policies, the subject sees her decision as being relevantly similar to that faced in Anderson's scenario. Rawls's "works" if, when considering questions of whether a particular social arrangement is just, the subject makes use of principles that would be endorsed in the original position.

In all of these cases, we see the force of Hume's observation above: When two options are presented abstractly, the choice between them may go one way; presented under some "particular idea" that "influence[s]" the "imagination," the choice may go the other. To say this is to say nothing about which option is the correct one; my concern here is only with the cognitive underpinnings of a certain philosophical methodology.

We return then to the themes of the opening section. I have suggested that by presenting content in a suitably concrete or abstract way, thought experiments may recruit representational schemas that were previously inactive. As a result, they can be expected to evoke responses that run counter to those evoked by alternative presentations of relevantly similar content. But exactly because of this, the responses they evoke may well remain in disequilibrium with responses evoked in alternative ways. When thought experiments succeed as devices of persuasion, it is because the evoked response becomes dominant, so that the subject comes (either reflectively or unreflectively) to represent relevant non-thought experimental content in light of the thought experimental conclusion.[20]

REFERENCES

Alexander, Joshua, and Weinberg, Jonathan M. 2007. "Analytic Epistemology and Experimental Philosophy." *Philosophy Compass* 2 (1): 56–80.

20. For comments on talks that served as distant predecessors to this article, I am grateful to audiences at the Conference on Intuitions, Fribourg, Switzerland; Bergen Community College; Cornell University; the University of North Carolina at Chapel Hill; Union College; and the CUNY Graduate Center. For comments on a more recent incarnation, I am grateful to an audience at the University of Toronto Workshop on Thought Experiments, organized by James Robert Brown. For comments on previous written versions of this article, I thank Carolyn Caine and Zoltán Gendler Szabó.

Anderson, Elizabeth. 2006. "Which Game Would You Rather Play?" Blog post on Left2Right, February 8, 2006. Retrieved January 4, 2007, from http://left2right.typepad.com/main/2006/02/what_game_would.html.

Bealer, George. 1998. "Intuition and the Autonomy of Philosophy." In Michael DePaul and William Ramsey, eds., *Rethinking Intuition*. Lanham, MD: Rowman and Littlefield, 1998, 201–40.

Chaiken, Shelly, and Trope, Yaacov, eds. 1999. *Dual-Process Theories in Social Psychology*. New York: Guilford Press.

Cheng, P. W., and Holyoak, K. J. 1985. "Pragmatic Reasoning Schemas." *Cognitive Psychology* 17: 391–416.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., and Oliver, L. M. 1986. "Pragmatic versus Syntactic Approaches to Training Deductive Reasoning." *Cognitive Psychology* 18: 293–328.

Cosmides, Leda. 1989. "The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task." *Cognition* 31: 187–276.

Crandall, C. S., and Greenfield, B. 1986. "Understanding the Conjunction Fallacy: A Conjunction of Effects?" Social Cognition 4: 408–19.

Cummins, Robert. 1998. "Reflections on Reflective Equilibrium." In Michael DePaul and William Ramsey, eds., *Rethinking Intuition*. Lanham, MD: Rowman and Littlefield, 1998, 113–28.

Denes-Raj, Veronika, and Epstein, Seymour. 1994. "Conflict between Intuitive and Rational Processing: When People Behave against Their Better Judgment." *Journal of Personality and Social Psychology* 66(5): 819–29.

de Waal, Fritz. 1996. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge, MA: Harvard University Press.

Dominowski, R. L. 1995. "Content Effects in Wason's Selection Task." In *Perspectives on Thinking and Reasoning*, ed. S. E. Newstead and Jonathan St. B. T. Evans, 41–65. Hove, UK: Psychology Press.

Epley, N., and Dunning, David. 2000. "Feeling 'Holier Than Thou': Are Selfserving Assessments Produced by Errors in Self or Social psychology?" *Journal of Personality and Social Psychology* 79: 861–75.

Epstein, Seymour. 1990. "Cognitive-Experiential Self-Theory." In *Handbook of Personality Theory and Research*, ed. L. Pervin, 165–92. New York: Guilford Publications.

Epstein, Seymour, Donovan, S., and Denes-Raj, V. 1999. "The Missing Link in the Paradox of the Linda Conjunction Problem: Beyond Knowing and Thinking of the Conjunction Rule, the Intrinsic Appeal of Heuristic Processing." *Personality and Social Psychology Bulletin* 25: 204–14.

Epstein, Seymour, Pacini, R., Denes-Raj, V., and Heier, H. 1996. "Individual Differences in Intuitive-Experiential and Analytical-Rational Thinking Styles." *Journal of Personality and Social Psychology* 71: 390–405.

Evans, Jonathan St. B.T. 1998. "Matching Bias in Conditional Reasoning: Do We Understand It after 25 Years?" *Thinking and Reasoning* 4: 45–82.

———. 2003. "In Two Minds: Dual Processing Accounts of Reasoning." *Trends in Cognitive Sciences* 7(10): 454–59.

———. Forthcoming 2008. "Dual-Processing Accounts of Reasoning, Judgment and Social Cognition." *Annual Review of Psychology*.

Evans, Jonathan St. B. T., Barston, J. L. and Pollard, P. 1983. "On the Conflict between Logic and Belief in Syllogistic Reasoning." *Memory and Cognition* 11: 295–306.

Evans, Jonathan St. B. T., and Over, D. E. 1996. "Rationality in the Selection Task: Epistemic Utility versus Uncertainty Reduction." *Psychological Review* 103: 356–63.

Foot, Philippa. 1978. "The Problem of Abortion and the Doctrine of Double Effect." In *her Virtues and Vices and Other Essays in Moral Philosophy*, 19–35. Berkeley, CA: University of California Press; Oxford: Blackwell.

Gendler, Tamar Szabó, and Hawthorne, John. 2005. "The Real Guide to Fake Barns: A Catalogue of Gifts for Your Epistemic Enemies." *Philosophical Studies* 124: 331–52.

Gigerenzer, Gerd, and Hug, K. 1992. "Domain-Specific Reasoning: Social Contracts, Cheating, and Perspective Change." *Cognition* 43: 127–71.

Gigerenzer, Gerd, and Regier, T. P. 1996. "How Do We Tell an Association from a Rule?" *Psychological Bulletin* 119(1): 23–26.
Gilbert, Daniel T. 1999. "What the Mind's Not." In *Dual-Process Theories in Social Psychology*, ed. Shelly Chaiken and Yaacov Trope, 3–11. New York: Guilford Press.
Gilovich, Tom, Griffin, D., and Kahneman, Daniel, eds. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.
Goel, Vinod, Buchel, C., Frith, C., and Dolan, R. J. 2000. "Dissociation of Mechanisms Underlying Syllogistic Reasoning." *Neuroimage* 12: 504–14.
Goel, Vinod, and Dolan, R. J. 2003. "Explaining Modulation of Reasoning by Belief." *Cognition* 87: B11–B22.
Goldman, Alvin. 2007. "Philosophical Intuitions: Their Target, Their Source, and Their Epistemic Status," *Grazer Philosophische Studien* 74: 1–25.
Gould, Stephen Jay 1991. *Bully for Brontosaurus: Reflections on Natural History*. New York: Norton.
Greene, Joshua, Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. 2001. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293: 2105–08.
Griggs, R. A., and Cox, J. R. 1982. "The Elusive Thematic Materials Effect in the Wason Selection Task." *British Journal of Psychology* 73: 407–20.
Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108: 814–34.
Haidt, Jonathan, and Joseph, C. 2004. "Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues." *Daedalus* 133(4): 55–66.
Hauser, Marc D. 2006. *Moral Minds: How Nature Designed Our Sense of Right and Wrong*. New York: Harper Collins.
Hinton, G. E. 1990. "Mapping Part-Whole Hierarchies into Connectionist Networks." *Artificial Intelligence* 46(1): 47–76.
Houdé, O., Zago, L., Mellet, E., Moutier, S., Pineau, A., Mazoyer, B., and Tzourio-Mazoyer, N. 2000. "Shifting from the Perceptual Brain to the Logical Brain: The Neural Impact of Cognitive Inhibition Training." *Journal of Cognitive Neuroscience* 12: 721–28.
Hume, David. [1739] 1978. *A Treatise on Human Nature*, ed. L. A. Selby-Bigge. Oxford: Clarendon.
James, William. [1890] 1950. *The Principles of Psychology*. New York: Dover Publications.
Johnson-Laird, P N., and Byrne, Ruth. 2002. "Conditionals: A Theory of Meaning, Pragmatics, and Inference." *Psychological Review* 109(4): 646–78.
Kahneman, D., Slovic, P., and Tversky, A., eds. 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.
Kahneman, Daniel, and Tversky, Amos, eds. 2000. *Choices, Values and Frames*. Cambridge: Cambridge University Press.
Kant, Immanuel. [1785] 1981. *Grounding for the Metaphysics of Morals*, trans. James Wesley Ellington. Indianapolis: Hackett.
Kirby, Kris N. 1994. "Probabilities and Utilities of Fictional Outcomes in Wason's Four-Card Selection Task." *Cognition* 51: 1–28.
Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., and Damasio, A. 2007. "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgment." *Nature* 446: 908–11.
Lehrer, Keith. 1990. *Theory of Knowledge*. Boulder, CO: Westview Press.
Mill, John Stuart. [1861] 2001. *Utilitarianism*. Indianapolis: Hackett.
Neisser, Ulrich. 1963. "The Multiplicity of Thought." *British Journal of Psychology* 54: 1–14.
Nichols, Shaun, and Knobe, Joshua. Forthcoming. "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions." *Nous*.
Nussbaum, Martha. 1990. *Love's Knowledge: Essays on Philosophy and Literature*. Oxford: Oxford University Press.
Oaksford, Mike, and Chater, Nick. 1994. "A Rational Analysis of the Selection Task as Optimal Data Selection." *Psychological Review* 101(4): 608–31.
———. 1996. "Rational Explanation of the Selection Task." *Psychological Review* 103(2): 381–91.
Piaget, Jean. 1929. *The Child's Conception of the World*. London: Routledge and Kegan Paul.

Pizarro, D. A., Uhlman, E. L., Tannenbaum, D., and Ditto, P. H. Manuscript. "The Motivated Use of Moral Principles."

Pronin, Emily. 2006. "Perception and Misperception of Bias in Human Judgment." *Trends in Cognitive Sciences* 11(1): 37–43.

Pronin, Emily, Gilovic, T. D., and Ross, L. 2004. "Objectivity in the Eye of the Beholder: Divergent Perceptions of Bias in Self versus Others." *Psychological Review* 111: 781–99.

Pust, Joel. 2000. *Intuitions as Evidence*. New York: Garland Press.

Rawls, John. [1971] 1999 (Revised). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Sloman, S. A. 1996. "The Empirical Case for Two Systems of Reasoning." *Psychological Bulletin* 119: 3–22.

Smolensky, Paul. 1988. "On the Proper Treatment of Connectionism." *Behavioral and Brain Sciences* 11: 1–23.

Sosa, Ernest. 2007a. "Experimental Philosophy and Philosophical Intuition." *Philosophical Studies* 132: 99–107.

———. 2007b. "Intuitions: Their Nature and Epistemic Efficacy," *Grazer Philosophische Studien* Forthcoming.

Stanovich, Keith E. 1999. "Who Is Rational?" *Studies of Individual Differences in Reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stanovich, Keith E., and West, Richard F. 2000. "Individual Differences in Reasoning: Implications for the Rationality Debate" (with discussion and replies). *Behavioral and Brain Sciences* 23: 645–726.

Sunstein, Cass. 2005. "Moral Heuristics" (with discussion and replies). *Behavioral and Brain Sciences* 28(4): 531–73.

Swain, S., Alexander, J., and Weinberg, J. M. Manuscript. "The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp." Draft of 1/30/2006. Available at http://www.indiana.edu/~eel/.

Thomson, Judith Jarvis. 1971. "A Defense of Abortion." *Philosophy and Public Affairs*. 1/1(Fall): 47–66.

———. 1985. "The Trolley Problem." *Yale Law Journal* 94: 1395–1415.

Thorndike, E. L. 1922. "The Effect of Changed Data on Reasoning." *Journal of Experimental Psychology* 5(1): 33–38.

Tversky, Amos, and Kahneman, Daniel. 1983. "Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement." *Psychological Review* 90: 293–315.

Wason, Peter C. 1966. "Reasoning." In *New Horizons in Psychology*, ed. B. Foss, 135–51. Harmondsworth: Penguin Books.

Weatherson, Brian. 2003. "What Good Are Counterexamples?" *Philosophical Studies* 115: 1–31.

Weinberg, Jonathan M., Crowley, Steve, Gonnerman, Chad, Swain, Stacey and Vandewalker, Ian. Manuscript. "Intuition and Calibration." Version of 9/18/05. Available at http://www.indiana.edu/~eel/.

Weinberg, Jonathan M., Nichols, Shaun, and Stich, Stephen. 2001. "Normativity and Epistemic Intuitions." *Philosophical Topics* 29(1/2): 429–60.

Wilkins, Minna C. 1928. "The Effect of Changed Material on Ability to Do Formal Syllogistic Reasoning." *Archives of Psychology* 102: 1–84.

Williamson, Timothy. 2005. "Armchair Philosophy, Metaphysical Modality and Counterfactual Thinking." *Proceedings of the Aristotelian Society* 105: 1–23.