Rationality and Philosophy

1. Less Wrong Rationality and Mainstream Philosophy
2. Philosophy: A Diseased Discipline
3. How You Make Judgments: The Elephant and its Rider
4. Your Evolved Intuitions
5. Intuition and Unconscious Learning
6. When Intuitions Are Useful
7. Concepts Don't Work That Way
8. Living Metaphorically
9. Intuitions Aren't Shared That Way
10. Philosophy Needs to Trust Your Rationality Even Though It Shouldn't
11. Train Philosophers with Pearl and Kahneman, not Plato and Kant

Less Wrong Rationality and Mainstream Philosophy

Part of the sequence: Rationality and Philosophy

Despite Yudkowsky's distaste for mainstream philosophy, Less Wrong is largely a philosophy blog. Major topics include epistemology, philosophy of language, free will, metaphysics, metaethics, normative ethics, machine ethics, axiology, philosophy of mind, and more.

Moreover, standard Less Wrong positions on philosophical matters have been standard positions in a movement *within* mainstream philosophy for half a century. That movement is sometimes called "Quinean naturalism" after Harvard's W.V. Quine, who articulated the Less Wrong approach to philosophy in the 1960s. Quine was one of the most influential philosophers of the last 200 years, so I'm not talking about an *obscure* movement in philosophy.

Let us survey the connections. Quine thought that philosophy was continuous with science - and where it wasn't, it was *bad* philosophy. He embraced empiricism and reductionism. He rejected the notion of libertarian free will. He regarded postmodernism as sophistry. Like Wittgenstein and Yudkowsky, Quine didn't try to straightforwardly *solve* traditional Big Questions as much as he either dissolved those questions or reframed them such that they *could* be solved. He dismissed endless semantic arguments about the meaning of vague terms like *knowledge*. He rejected *a priori* knowledge. He rejected the notion of privileged philosophical insight: knowledge comes from ordinary knowledge, as best refined by science. Eliezer once said that philosophy should be about cognitive science, and Quine would agree. Quine famously wrote:

> The stimulation of his sensory receptors is all the evidence anybody has had to go on, ultimately, in arriving at his picture of the world. Why not just see how this construction really proceeds? Why not settle for psychology?

But isn't this using science to justify science? Isn't that circular? Not quite, say Quine and Yudkowsky. It is merely "reflecting on your mind's degree of trustworthiness, using your current mind as opposed to something else." Luckily, the brain is the lens that sees its flaws. And thus, says Quine:

> Epistemology, or something like it, simply falls into place as a chapter of psychology and hence of natural science.

Yudkowsky once wrote, "If there's any centralized repository of reductionist-grade naturalistic cognitive philosophy, I've never heard mention of it."

When I read that I thought: *What? That's Quinean naturalism! That's Kornblith and Stich and Bickle and the Churchlands and Thagard and Metzinger and Northoff! There are* hundreds *of philosophers who do that!*

Non-Quinean philosophy

But I should also mention that LW philosophy / Quinean naturalism is not the *largest* strain of mainstream philosophy. Most philosophy is still done in relative ignorance (or ignoring) of cognitive science. Consider the preface to *Rethinking Intuition*:

> Perhaps more than any other intellectual discipline, philosophical inquiry is driven by intuitive judgments, that is, by what "we would say" or by what seems true to the inquirer. For most of philosophical theorizing and debate, intuitions serve as something like a source of evidence that can be used to defend or attack particular philosophical positions.
>
> One clear example of this is a traditional philosophical enterprise commonly known as *conceptual analysis*. Anyone familiar with Plato's dialogues knows how this type of inquiry is conducted. We see Socrates encounter someone who claims to have figured out the true essence of some abstract notion... the person puts forward a definition or analysis of the notion in the form of necessary and sufficient conditions that are thought to capture all and only instances of the concept in question. Socrates then refutes his interlocutor's definition of the concept by pointing out various counterexamples...
>
> For example, in Book I of the *Republic*, when Cephalus defines justice in a way that requires the returning of property and total honesty, Socrates responds by pointing out that it would be unjust to return weapons to a person who had gone mad or to tell the whole truth to such a person. What is the status of these claims that certain behaviors would be unjust in the circumstances described? Socrates does not argue for them in any way. They seem to be no more than spontaneous judgments representing "common sense" or "what we would say." So it would seem that the proposed analysis is rejected because it fails to capture our intuitive judgments about the nature of justice.
>
> After a proposed analysis or definition is overturned by an intuitive counterexample, the idea is to revise or replace the analysis with one that is not subject to the counterexample. Counterexamples to the new analysis are sought, the analysis revised if any counterexamples are found, and so on...
>
> Refutations by intuitive counterexamples figure as prominently in today's philosophical journals as they did in Plato's dialogues...
>
> ...philosophers have continued to rely heavily upon intuitive judgments in pretty much the way they always have. And they continue to use them in the absence of any well articulated, generally accepted account of intuitive judgment - in particular, an account that establishes their epistemic credentials.
>
> However, what appear to be serious new challenges to the way intuitions are employed have recently emerged from an unexpected quarter - empirical research in cognitive psychology.
>
> With respect to the tradition of seeking definitions or conceptual analyses that are immune to counterexample, the challenge is based on the work of psychologists studying the nature of concepts and categorization of judgments. (See, e.g., Rosch 1978; Rosch and Mervis 1975; Rips 1975; Smith and Medin 1981). Psychologists working in this area have been pushed to abandon the view that we represent

> concepts with simple sets of necessary and sufficient conditions. The data seem to show that, except for some mathematical and geometrical concepts, it is not possible to use simple sets of conditions to capture the intuitive judgments people make regarding what falls under a given concept...
>
> With regard to the use of intuitive judgments exemplified by reflective equilibrium, the challenge from cognitive psychology stems primarily from studies of inference strategies and belief revision. (See, e.g., Nisbett and Ross 1980; Kahneman, Slovic, and Tversky 1982.) Numerous studies of the patterns of inductive inference people use and judge to be intuitively plausible have revealed that people are prone to commit various fallacies. Moreover, they continue to find these fallacious patterns of reasoning to be intuitively acceptable upon reflection... Similarly, studies of the "intuitive" heuristics ordinary people accept reveal various gross departures from empirically correct principles...
>
> There is a growing consensus among philosophers that there is a serious and fundamental problem here that needs to be addressed. In fact, we do not think it is an overstatement to say that Western analytic philosophy is, in many respects, undergoing a crisis where there is considerable urgency and anxiety regarding the status of intuitive analysis.

Conclusion

So Less Wrong-style philosophy is part of a movement within mainstream philosophy to massively reform philosophy in light of recent cognitive science - a movement that has been active for at least two decades. Moreover, Less Wrong-style philosophy has its roots in Quinean naturalism from *fifty* years ago.

And I haven't even covered all the work in formal epistemology toward (1) mathematically formalizing concepts related to induction, belief, choice, and action, and (2) arguing about the foundations of probability, statistics, game theory, decision theory, and algorithmic learning theory.

So: Rationalists need not dismiss or avoid philosophy.

Update: To be clear, though, I *don't* recommend reading Quine. Most people should not spend their time reading even Quinean philosophy; learning statistics and AI and cognitive science will be far more useful. All I'm saying is that mainstream philosophy, especially Quinean philosophy, *does* make some useful contributions. **I've listed more than 20 of mainstream philosophy's useful contributions here, including several instances of classic LW dissolution-to-algorithm**.

But maybe it's a testament to the epistemic utility of Less Wrong-ian rationality training and *thinking like an AI researcher* that Less Wrong got so many things right *without* much interaction with Quinean naturalism. As Daniel Dennett (2006) said, "AI makes philosophy honest."

Next post: Philosophy: A Diseased Discipline

References

Dennett (2006). Computers as Prostheses for the Imagination. Talk presented at the International Computers and Philosophy Conference, Laval, France, May 3, 2006.

Kahneman, Slovic, & Tversky (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Nisbett and Ross (1980). *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall.

Rips (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Behavior, 12*: 1-20.

Rosch (1978). Principles of categorization. In Rosch & Lloyd (eds.), *Cognition and Categorization* (pp. 27-48). Lawrence Erlbaum Associates.

Rosch & Mervis (1975). Family resemblances: studies in the internal structure of categories. Cognitive Psychology, 8: 382-439.

Smith & Medin (1981). *Concepts and Categories*. MIT Press.

Philosophy: A Diseased Discipline

Part of the sequence: Rationality and Philosophy

Eliezer's anti-philosophy post Against Modal Logics was pretty controversial, while my recent pro-philosophy (by LW standards) post and my list of useful mainstream philosophy contributions were massively up-voted. This suggests a significant appreciation for mainstream philosophy on Less Wrong - not surprising, since Less Wrong covers so many philosophical topics.

If you followed the recent very long debate between Eliezer and I over the value of mainstream philosophy, you may have gotten the impression that Eliezer and I strongly diverge on the subject. But I suspect I agree more with Eliezer on the value of mainstream philosophy than I do with many Less Wrong readers - perhaps most.

That might sound odd coming from someone who writes a philosophy blog and spends most of his spare time doing philosophy, so let me explain myself. (Warning: broad generalizations ahead! There are exceptions.)

Failed methods

Large swaths of philosophy (e.g. continental and postmodern philosophy) often don't even *try* to be clear, rigorous, or scientifically respectable. This is philosophy of the "Uncle Joe's musings on the meaning of life" sort, except that it's dressed up in big words and long footnotes. You will occasionally stumble upon an *argument*, but it falls prey to magical categories and language confusions and non-natural hypotheses. You may also stumble upon science or math, but they are used to 'prove' things irrelevant to the actual scientific data or the equations used.

Analytic philosophy is clearer, more rigorous, and better with math and science, but only does a slightly better job of avoiding magical categories, language confusions, and non-natural hypotheses. Moreover, its central tool is intuition, and this displays a near-total ignorance of how brains work. As Michael Vassar observes, philosophers are "spectacularly bad" at understanding that their intuitions are generated by cognitive algorithms.

A diseased discipline

What about Quinean naturalists? Many of them at *least* understand the basics: that things are made of atoms, that many questions don't need to be answered but instead dissolved, that the brain is not an a priori truth factory, that intuitions come from cognitive algorithms, that humans are loaded with bias, that language is full of tricks, and that justification rests in the lens that can see its flaws. Some of them are even Bayesians.

Like I said, a few naturalistic philosophers are doing some useful work. But the signal-to-noise ratio is *much* lower even in naturalistic philosophy than it is in, say, behavioral economics or cognitive neuroscience or artificial intelligence or statistics. Why? Here are some hypotheses, based on my thousands of hours in the literature:

1. Many philosophers have been infected (often by later Wittgenstein) with the idea that philosophy is *supposed* to be useless. If it's useful, then it's science or math or something else, but not philosophy. Michael Bishop says a common complaint from his colleagues about his 2004 book is that it is *too useful*.
2. Most philosophers *don't* understand the basics, so naturalists spend much of their time coming up with new ways to argue that people are made of atoms and intuitions don't trump science. They fight beside the poor atheistic philosophers who keep coming up with new ways to argue that the universe was not created by someone's invisible magical friend.
3. Philosophy has grown into an abnormally backward-looking discipline. Scientists like to put their work in the context of what old dead guys said, too, but philosophers have a real *fetish* for it. Even naturalists spend a fair amount of time re-interpreting Hume and Dewey yet again.
4. Because they were trained in traditional philosophical ideas, arguments, and frames of mind, naturalists will anchor and adjust from traditional philosophy when they make progress, rather than scrapping the whole mess and starting from scratch with a correct understanding of language, physics, and cognitive science. Sometimes, philosophical work is useful to build from: Judea Pearl's triumphant work on causality built on earlier counterfactual accounts of causality from philosophy. Other times, it's best to ignore the past confusions. Eliezer made most of his philosophical progress on his own, in order to solve problems in AI, and only later looked around in philosophy to see which standard position his own theory was most similar to.
5. Many naturalists aren't trained in cognitive science or AI. Cognitive science is essential because the tool we use to philosophize is the brain, and if you don't know how your tool works then you'll use it poorly. AI is useful because it keeps you honest: you can't write confused concepts or non-natural hypotheses in a programming language.
6. Mainstream philosophy publishing favors the established positions and arguments. You're more likely to get published if you can write about how intuitions are useless in solving Gettier problems (which is a confused set of non-problems anyway) than if you write about how to make a superintelligent machine preserve its utility function across millions of self-modifications.
7. Even much of the *useful* work naturalistic philosophers do is not at the cutting-edge. Chalmers' update for I.J. Good's 'intelligence explosion' argument is the best one-stop summary available, but it doesn't get as far as the Hanson-Yudkowsky AI-Foom debate in 2008 did. Talbot (2009) and Bishop & Trout (2004) provide handy summaries of much of the heuristics and biases literature, just like Eliezer has so usefully done on Less Wrong, but of course this isn't cutting edge. You could always just read it in the primary literature by Kahneman and Tversky and others.

Of course, there *is* mainstream philosophy that is both good and cutting-edge: the work of Nick Bostrom and Daniel Dennett stands out. And of course there *is* a role for those who keep arguing for atheism and reductionism and so on. I was a fundamentalist Christian until I read some contemporary atheistic philosophy, so that kind of work definitely does some good.

But if you're looking to solve cutting-edge problems, mainstream philosophy is one of the *last* places you should look. Try to find the answer in the cognitive science or AI literature first, or try to solve the problem by applying rationalist thinking: like this.

Swimming the murky waters of mainstream philosophy is perhaps a job best left for those who already spent several years studying it - that is, people like *me*. I already

know what things are called and where to look, and I have an efficient filter for skipping past the 95% of philosophy that isn't useful to me. And hopefully my rationalist training will protect me from picking up bad habits of thought.

Philosophy: the way forward

Unfortunately, many important problems are fundamentally *philosophical* problems. Philosophy itself is unavoidable. How can we proceed?

First, we must remain vigilant with our rationality training. It is not easy to overcome millions of years of brain evolution, and as long as you are human there is no final victory. You will always wake up the next morning as *homo sapiens*.

Second, if you want to contribute to cutting-edge problems, even ones that seem philosophical, it's far more productive to study math and science than it is to study philosophy. You'll learn more in math and science, and your learning will be of a higher quality. Ask a fellow rationalist who is knowledgeable about philosophy what the standard positions and arguments in philosophy are on your topic. If any of them seem *really* useful, grab those particular works and read them. But again: you're probably better off trying to solve the problem by thinking like a cognitive scientist or an AI programmer than by ingesting mainstream philosophy.

However, I must say that I wish so much of Eliezer's cutting-edge work wasn't spread out across hundreds of Less Wrong blog posts and long SIAI articles written in with an idiosyncratic style and vocabulary. I would rather these ideas were written in standard academic form, even if they transcended the standard game of mainstream philosophy.

But it's one thing to complain; another to offer solutions. So let me tell you what I think cutting-edge philosophy should be. As you might expect, my vision is to combine what's good in LW-style philosophy with what's good in mainstream philosophy, and toss out the rest:

1. Write short articles. One or two major ideas or arguments per article, maximum. Try to keep each article under 20 pages. It's hard to follow a hundred-page argument.
2. Open each article by explaining the context and goals of the article (even if you cover mostly the same ground in the opening of 5 other articles). What topic are you discussing? Which problem do you want to solve? What have other people said about the problem? What will you accomplish in the paper? Introduce key terms, cite standard sources and positions on the problem you'll be discussing, even if you disagree with them.
3. If possible, use the standard terms in the field. If the standard terms are flawed, explain why they are flawed and then introduce your new terms in that context so everybody knows what you're talking about. This requires that you research your topic so you know what the standard terms and positions *are*. If you're talking about a problem in cognitive science, you'll need to read cognitive science literature. If you're talking about a problem in social science, you'll need to read social science literature. If you're talking about a problem in epistemology or morality, you'll need to read philosophy.
4. Write as clearly and simply as possible. Organize the paper with lots of heading and subheadings. Put in lots of 'hand-holding' sentences to help your reader along: explain the point of the previous section, then explain why the next

section is necessary, etc. Patiently guide your reader through every step of the argument, especially if it is long and complicated.
5. Always cite the relevant literature. If you can't find much work relevant to your topic, you almost certainly haven't looked hard enough. Citing the relevant literature not only lends weight to your argument, but also enables the reader to track down and examine the ideas or claims you are discussing. Being lazy with your citations is a sure way to frustrate precisely those readers who care enough to read your paper closely.
6. Think like a cognitive scientist and AI programmer. Watch out for biases. Avoid magical categories and language confusions and non-natural hypotheses. Look at your intuitions from the outside, as cognitive algorithms. Update your beliefs in response to evidence. [**This one is central. This is LW-style philosophy**.]
7. Use your rationality training, but avoid language that is unique to Less Wrong. Nearly all these terms and ideas have standard names outside of Less Wrong (though in many cases Less Wrong already uses the standard language).
8. Don't dwell too long on what old dead guys said, nor on semantic debates. Dissolve semantic problems and move on.
9. Conclude with a summary of your paper, and suggest directions for future research.
10. Ask fellow rationalists to read drafts of your article, then re-write. Then rewrite again, adding more citations and hand-holding sentences.
11. Format the article attractively. A well-chosen font makes for an easier read. Then publish (in a journal or elsewhere).

Note that this is *not* just my vision of how to get published in journals. It's my vision of *how to do philosophy*.

Meeting journals standards is *not* the most important reason to follow the suggestions above. Write short articles because they're easier to follow. Open with the context and goals of your article because that makes it easier to understand, and lets people decide right away whether your article fits their interests. Use standard terms so that people already familiar with the topic aren't annoyed at having to learn a whole new vocabulary just to read your paper. Cite the relevant positions and arguments so that people have a sense of the context of what you're doing, and can look up what other people have said on the topic. Write clearly and simply and with much organization so that your paper is not wearying to read. Write lots of hand-holding sentences because we always communicate less effectively then we *thought* we did. Cite the relevant literature as much as possible to assist your most careful readers in getting the information they want to know. Use your rationality training to remain sharp at all times. And so on.

That is what cutting-edge philosophy could look like, I think.

Next post: How You Make Judgments

Previous post: Less Wrong Rationality and Mainstream Philosophy

How You Make Judgments: The Elephant and its Rider

Part of the sequence: Rationality and Philosophy

Whether you're doing science or philosophy, flirting or playing music, the first and most important tool you are using is your *mind*. To use your tool well, it will help to know how it works. Today we explore how your mind makes judgments.

From Plato to Freud, many have remarked that humans seem to have more than one mind.[1] Today, detailed 'dual-process' models are being tested by psychologists and neuroscientists:

> Since the 1970s *dual-process* theories have been developed [to explain] various aspects of human psychology... Typically, one of the processes is characterized as fast, effortless, automatic, nonconscious, inflexible, heavily contextualized, and undemanding of working memory, and the other as slow, effortful, controlled, conscious, flexible, decontextualized, and demanding of working memory.[2]

Dual-process theories for reasoning,[3] learning and memory,[4] decision-making,[5] belief,[6] and social cognition[7] are now widely accepted to be correct to some degree,[8] with researchers currently working out the details.[9] Dual-process theories even seem to be appropriate for some nonhuman primates.[10]

Naturally, some have wondered if there might be a "grand unifying dual-process theory that can incorporate them all."[11] We might call such theories *dual-system* theories of mind,[12] and several have been proposed.[13] Such unified theories face problems, though. 'Type 1' (fast, nonconscious) processes probably involve many nonconscious architectures,[14] and brain imaging studies show a wide variety of brain systems at work at different times when subjects engage in 'type 2' (slow, conscious) processes.[15]

Still, perhaps there is a sense in which one 'mind' relies mostly on type 1 processes, and a second 'mind' relies mostly on type 2 processes. One suggestion is that Mind 1 is evolutionarily old and thus shared with other animals, while Mind 2 is recently evolved and particularly developed in humans. (But not *fully unique* to humans, because some animals do seem to exhibit a distinction between stimulus-controlled and higher-order controlled behavior.[16]) But this theory faces problems. A standard motivation for dual-process theories of reasoning is the conflict between cognitive biases (from type 1 processes) and logical reasoning (type 2 processes).[17] For example, logic and belief bias often conflict.[18] But both logic and belief bias can be located in the pre-frontal cortex, an evolutionarily *new* system.[19] So either Mind 1 is not entirely old, or Mind 2 is not entirely composed of type 2 processes.

We won't try to untangle these mysteries here. Instead, we'll focus on one of the most successful dual-process theories: Kahneman and Frederick's dual-process theory of judgment.[20]

Attribute substitution

Kahneman and Frederick propose an "attribute-substitution model of heuristic judgment" which claims that judgments result from both type 1 and type 2 processes.[21] The authors explain:

> The early research on judgment heuristics was guided by a simple and general hypothesis: When confronted with a difficult question, people may answer an easier one instead and are often unaware of the substitution. A person who is asked "What proportion of long-distance relationships break up within a year?" may answer as if she had been asked "Do instances of failed long-distance relationships come readily to mind?" This would be an application of the availability heuristic. A professor who has heard a candidate's job talk and now considers the question "How likely is it that this candidate could be tenured in our department?" may answer the much easier question: "How impressive was the talk?" This would be an example of one form of the representativeness heuristic.[22]

Next: what is attribute substitution?

> …whenever the aspect of the judgmental object that one intends to judge (the *target attribute*) is less readily assessed than a related property that yields a plausible answer (the *heuristic attribute*), individuals may unwittingly substitute the simpler assessment.[22]

For example, one study[23] asked subjects two questions among many others: "How happy are you with your life in general?" and "How many dates did you have last month?" In this order, the correlation between the two questions was negligible. If the dating question was asked first, however, the correlation was .66. The question about dating frequency seems to evoke "an evaluation of one's romantic satisfaction" that "lingers to become the heuristic attribute when the global happiness question is subsequently encountered."[22]

Or, consider a question in another study: "If a sphere were dropped into an open cube, such that it just fit (the diameter of the sphere is the same as the interior width of the cube), what proportion of the volume of the cube would the sphere occupy?"[24] The target attribute (the volumetric relation between cube and sphere) is difficult to assess intuitively, and it appears that subjects sought out an easier-to-assess heuristic attribute instead, substituting the question "If a *circle* were drawn inside a *square*, what proportion of the *area* of the square does the circle occupy?" The mean estimate for the 'sphere inside cube' problem was 74%, almost identical to the mean estimate of the 'circle inside square' problem (77%) but far larger than the correct answer for the 'sphere inside cube' problem (52%).

Attribution substitutions like this save on processing power but introduce systematic biases into our judgment.[25]

Some attributes are always candidates for the heuristic role in attribute substitution because they play roles in daily perception and cognition and are thus always accessible: cognitive fluency, causal propensity, surprisingness, mood, and affective valence.[26] Less prevalent attributes can become accessible for substitution if recently evoked or primed.[27]

Supervision of intuitive judgments

Intuitive judgments, say Kahneman and Frederick, arise from processes like attribute substitution, of which we are unaware. They "bubble up" from the unconscious, after which many of them are evaluated and either endorsed or rejected by type 2 processes.

You can feel the tension[28] between type 1 and type 2 processes in your own judgment when you try the Stroop task. Name the color of a list of colored words and you will find that you pause a bit when the word you see names a different color than the color it is written in, like this: green. Your unconscious, intuitive judgment uses an availability heuristic to suggest the word 'green' is shown in green, but your conscious type 2 processes quickly correct the unconscious judgment and conclude that it is written in red. You have no such momentary difficulty naming the color of this word: blue.

In many cases, type 2 processes have no trouble correcting the judgments of type 1 processes.[29] But because type 2 processes are slow, they can be interrupted by time pressure.[30] On the other hand, biased attribute substitution can sometimes be prevented if subjects are alerted to the possible evaluation contamination in advance.[31] (This finding justifies a great deal of material on Less Wrong, which alerts you to many cognitive biases - that is, possible sources of evaluation contamination.)

Often, type 2 processes fail to correct intuitive judgments, as demonstrated time and again in the heuristics and biases literature.[32] And even when type 2 processes correct intuitive judgments, the *feeling* that the intuitive judgments is correct may remain. Consider the famous Linda problem. Knowledge of probability theory does not extinguish the *feeling* (from type 1 processes) that Linda must be a feminist bank teller. As Stephen Jay Gould put it:

> I know [the right answer], 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!"[33]

Conclusion

Kahneman and Frederick's dual-process theory appears to be successful in explaining a wide range of otherwise puzzling phenomena in human judgment.[34] The big picture of all this is described well by Jonathan Haidt, who imagines his conscious mind as a rider upon an elephant:

> I'm holding the reins in my hands, and by pulling one way or the other I can tell the elephant to turn, to stop, or to go. I can direct things, but only when the elephant doesn't have desires of his own. When the elephant really wants to do something, I'm no match for him.
>
> ...The controlled system [can be] seen as an advisor. It's a rider placed on the elephant's back to help the elephant make better choices. The rider can see farther into the future, and the rider can learn valuable information by talking to other riders or by reading maps, but the rider cannot order the elephant around against its will...

> ...The elephant, in contrast, is everything else. The elephant includes gut feelings, visceral reactions, emotions, and intuitions that comprise much of the automatic system. The elephant and the rider each have their own intelligence, and when they work together well they enable the unique brilliance of human beings. But they don't always work together well.[35]

Next post: Your Evolved Intuitions

Previous post: Philosophy: A Diseased Discipline

Notes

[1] Plato divided the soul into three parts: reason, spirit, and appetite (Annas 1981, ch. 5). Descartes held that humans operate on unconscious mechanical processes we share with animals, but that humans' additional capacities for rational thought separate us from the animals (Cottingham 1992). Leibniz said that animals are guided by inductive reasoning, which also guides 'three-fourths' of human reasoning, but that humans can also partake in 'true reasoning' — logic and mathematics (Leibniz 1714/1989, p. 208; Leibniz 1702/1989, pp. 188-191). Many thinkers, most famously Freud, have drawn a division between conscious and unconscious thinking (Whyte 1978). For a more detailed survey, see Frankish & Evans (2009). Multiple-process theories of mind stand in contrast to monistic theories of mind, for example: Johnson-Laird (1983); Braine (1990); Rips (1994). Note that dual-process theories of mind need not conflict with massively modular view of human cognition like Barrett & Kurzban (2006) or Tooby & Cosmides (1992): see Mercier & Sperber (2009). Finally, note that dual-process theories sit comfortably with current research on situated cognition: Smith & Semin (2004).

[2] Frankish & Evans (2009).

[3] Evans (1989, 2006, 2007); Evans & Over (1996); Sloman (1996, 2002); Stanovich (1999, 2004, 2009); Smolensky (1988); Carruthers (2002, 2006, 2009); Lieberman (2003; 2009); Gilbert (1999).

[4] Sun et al. (2009); Eichenbaum & Cohen (2001); Carruthers (2006); Sherry & Schacter (1987); Berry & Dienes (1993); Reber (1993); Sun (2001).

[5] Kahneman & Frederick (2002, 2005).

[6] Dennett (1978, ch. 16; 1991); Cohen (1992); Frankish (2004); Verscheuren et al. (2005).

[7] Smith & Collins (2009); Bargh (2006); Strack & Deutsch (2004).

[8] Evans (2008); Evans & Frankish (2009). Or as Carruthers (2009) puts it, "Dual-system theories of human reasoning are now quite widely accepted, at least in outline."

[9] One such detail is: When and to what extent does System 2 intervene in System 1 processes? See: Evans (2006); Stanovich (1999); De Neys (2006); Evans & Curtis-Holmes (2005); Finucane et al. (2000); Newstead et al. (1992); Evans et al. (1994); Daniel & Klaczynski (2006); Vadenoncoeur & Markovits (1999); Thompson (2009). Other important open questions are explored in Fazio & Olson (2003); Nosek (2007); Saunders (2009). For an accessible overview of the field, see Evans (2010).

[10] Call & Tomasello (2005).

[11] Evans (2009).

[12] Dual-process and dual-system theories of the mind suggest multiple cognitive *architectures*, and should not be confused with theories of multiple *modes* of processing, or two kinds of *cognitive style*. One example of the latter is the supposed distinction between Eastern and Western thinking (Nisbett et al. 2001). Dual-process and dual-system theories of the mind should also be distinguished from theories that posit a *continuum* between one form of

thinking and another (e.g. Hammond 1996; Newstead 2000; Osman 2004), since this suggests there are not separate cognitive architectures at work.

[13] Evans (2003); Stanovich (1999, 2009); Evans & Over (1996); Smith & DeCoster (2000); Wilson (2002).

[14] Evans (2008, 2009); Stanovich (2004); Wilson (2002).

[15] Goel (2007).

[16] Toates (2004, 2006).

[17] Evans (1989); Evans & Over (1996); Kahneman & Frederick (2002); Klaczynski & Cottrell (2004); Sloman (1996); Stanovich (2004).

[18] Evans et al. (1983); Klauer et al. (2000).

[19] Evans (2009); Goel & Dolan (2003).

[20] Those who prefer video lectures to reading may enjoy a 2008 lecture on judgment and intuition, to which Kahneman contributed: 1, 2, 3, 4.

[21] They use the terms 'system 1' and 'system 2' instead of 'type 1' and 'type 2'. Their theory is outlined in Kahneman & Frederick (2002, 2005).

[22] Kahneman & Frederick (2005).

[23] Strack et al. (1988).

[24] Frederick & Nelson (2007).

[25] Cognitive biases particularly involved in attribute substitution include the availability heuristic (Lichtenstein et al. 1978; Schwarz et al. 1991; Schwarz & Vaughn 2002), the representativeness heuristic (Kahneman & Tversky 1973; Tversky & Kahneman 1982; Bar-Hillel & Neter 1993; Agnolia 1991), and the affect heuristic (Slovic et al. 2002; Finucane et al. 2000).

[26] Cognitive fluency: Jacoby & Dallas (1981); Schwarz & Vaughn (2002); Tversky & Kahneman (1973). Causal propensity: Michotte (1963); Kahneman & Varey (1990). Surprisingness: Kahneman & Miller (1986). Mood: Schwarz & Clore (1983). Affective valence: Bargh (1997); Cacioppo et al. (1993); Kahneman et al. (1999); Slovic et al. (2002); Zajonc (1980, 1997).

[27] Bargh et al. (1986); Higgins & Brendl (1995). Note also that attributes must be mapped across dimensions on a common scale, and we understand to some degree the mechanism that does this: Kahneman & Frederick (2005); Ganzach and Krantz (1990); Stevens (1975).

[28] Also see De Neys et al. (2010).

[29] Gilbert (1989).

[30] Finucane et al. (2000).

[31] Schwarz & Clore (1983); Schwarz (1996).

[32] Gilovich et al. (2002); Kahneman et al. (1982); Pohl (2005); Gilovich (1993); Hastie & Dawes (2009).

[33] Gould (1991), p. 469.

[34] See the overview in Kahneman & Frederick (2005).

[35] Haidt (2006), pp. 4, 17.

References

Agnolia (1991). Development of judgmental heuristics and logical reasoning: Training counteracts the representativeness heuristic. *Cognitive Development, 6*: 195–217.

Annas (1981). *An introduction to Plato's republic*. Oxford University Press.

Bar-Hillel & Neter (1993). How alike is it versus how likely is it: A disjunction fallacy in probability judgments. *Journal of Personality and Social Psychology, 41*: 671–680.

Bargh (1997). The automaticity of everyday life. *Advances in social cognition, 10*. Erlbaum.

Bargh, Bond, Lombardi, & Tota (1986). The additive nature of chronic and temporary sources of construct accessibility. *Journal of Personality and Social Psychology, 50(5)*: 869–878.

Bargh (2006). *Social psychology and the unconscious*. Psychology Press.

Barrett & Kurzban (2006). Modularity in cognition: Framing the debate. *Psychological Review, 113*: 628-647.

Berry & Dienes (1993). *Implicit learning*. Erlbaum.

Braine (1990). The 'natural logic' approach to reasoning. In Overton (ed.), *Reasoning, necessity and logic: Developmental perspectives*. Psychology Press.

Cacioppo, Priester, & Berntson (1993). Rudimentary determinants of attitudes: II. Arm flexion and extension have differential effects on attitudes. *Journal of Personality and Social Psychology, 65*: 5–17.

Call & Tomasello (2005). Reasoning and thinking in nonhuman primates. In Holyoak & Morrison (eds.), The Cambridge Handbook of Thinking and Reasoning (pp. 607-632). Cambridge University Press.

Carruthers (2002). The cognitive functions of language. *Behavioral and Brain Sciences, 25*: 657-719.

Carruthers (2006). *The architecture of the mind*. Oxford University Press.

Carruthers (2009). An architecture for dual reasoning. In Evans & Franklin (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 109-128). Oxford University Press.

Cohen (1992). *An essay on belief and acceptance*. Oxford University Press.

Cottingham (1992). Cartesian dualism: Theology, metaphysics, and science. In Cottingham (ed.), *The Cambridge companion to Descartes* (pp. 236-257). Cambridge University Press.

Daniel & Klaczynski (2006). Developmental and individual differences in conditional reasoning: Effects of logic instructions and alternative antecedents. *Child Development, 77*: 339-354.

De Neys, Moyens, & Vansteenwegen (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, affective, and behavioral neuroscience, 10(2)*: 208-216.

Dennett (1978). *Brainstorms: Philosophical essays on mind and psychology*. MIT Press.

Dennett (1991). Two contrasts: Folk craft versus folk science and belief versus opinion. In Greenwood (ed.), *The future of folk psychology: Intentionality and cognitive science* (pp. 135-148). Cambridge University Press.

De Neys (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science, 17*: 428-433.

Eichenbaum & Cohen (2001). *From conditioning to conscious reflection: Memory systems of the brain*. Oxford University Press.

Evans (1989). *Bias in human reasoning: Causes and consequences*. Erlbaum.

Evans (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*: 454-459.

Evans (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review, 13*: 378-395.

Evans (2007). *Hypothetical Thinking: Dual processes in reasoning and judgment*. Psychology Press.

Evans (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology, 59*: 255-278.

Evans (2009). How many dual-process theories do we need? One, two, or many? In Evans & Franklin (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 33-54). Oxford University Press.

Evans (2010). *Thinking Twice: Two minds in one brain*. Oxford University Press.

Evans & Over (1996). *Rationality and Reasoning*. Psychology Press.

Evans & Frankish, eds. (2009). *In Two Minds: Dual Processes and Beyond*. Oxford University Press.

Evans & Curtis-Holmes (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning, 11*: 382-389.

Evans, Barston, & Pollard (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11*: 295-306.

Evans, Newstead, Allen, & Pollard (1994). Debiasing by instruction: The case of belief bias. *European Journal of Cognitive Psychology, 6*: 263-285.

Fazio & Olson (2003). Implicit measures in social cognition research: Their meaning and uses. *Annual Review of Psychology, 54*: 297-327.

Finucane, Alhakami, Slovic, & Johnson (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making, 13*: 1-17.

Frankish (2004). *Mind and Supermind*. Cambridge University Press.

Frankish & Evans (2009). The duality of mind: An historical perspective. In Evans & Franklin (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 1-29). Oxford University Press.

Frederick & Nelson (2007). Attribute substitution in the estimation of volumetric relationships: Psychophysical phenomena underscore judgmental heuristics. Manuscript in preparation. Massachusetts Institute of Technology.

Ganzach and Krantz (1990). The psychology of moderate prediction: I. Experience with multiple determination. *Organizational Behavior and Human Decision Processes, 47*: 177–204.

Gilbert (1989). Thinking lightly about others: Automatic components of the social inference process. In Uleman & Bargh (eds.), *Unintended thought* (pp. 189–211). Guilford Press.

Gilbert (1999). What the mind's not. In Chaiken & Trope (eds.), *Dual-process theories in social psychology* (pp. 3–11). Guilford Press.

Gilovich (1993). *How we know what isn't so*. Free Press.

Gilovich, Griffin, & Kahneman, eds. (2002). *Heuristics and biases: the psychology of intuitive judgment*. Cambridge University Press.

Goel (2007). Cognitive neuroscience of deductive reasoning. In Holyoak & Morrison (eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 475-492). Cambridge University Press.

Goel & Dolan (2003). Explaining modulation of reasoning by belief. *Cognition, 87*: B11-B22.

Gould (1991). *Bully for brontosaurus. Reflections in natural history*. Norton.

Hammond (1996). *Human judgment and social policy*. Oxford University Press.

Haidt (2006). *The happiness hypothesis: Finding modern truth in ancient wisdom*. Basic Books.

Hastie & Dawes, eds. (2009). *Rational Choice in an Uncertain World, 2nd ed*. Sage.

Higgins & Brendl (1995). Accessibility and applicability: Some 'activation rules' influencing judgment. *Journal of Experimental Social Psychology, 31*: 218–243.

Jacoby & Dallas (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 3*: 306–340.

Johnson-Laird (1983). *Mental Models*. Cambridge University Press.

Kahneman & Tversky (1973). On the psychology of prediction. *Psychological Review, 80*: 237–251.

Kahneman et al. (1999). Economic preferences or attitude expressions? An analysis of dollar responses to public issues. *Journal of Risk and Uncertainty, 19*: 203–235.

Kahneman & Frederick (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In Gilovich, Griffin, & Kahneman (eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49-81). Cambridge University Press.

Kahneman & Frederick (2005). A model of heuristic judgment. In Holyoak & Morrison (eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 267-294). Cambridge University Press.

Kahneman & Miller (1986). Norm theory: Comparing reality with its alternatives. *Psychological Review, 93*: 136–153.

Kahneman & Varey (1990). Propensities and counterfactuals: The loser that almost won. *Journal of Personality and Social Psychology, 59(6)*: 1101–1110.

Klaczynski & Cottrell (2004). A dual-process approach to cognitive development: The case of children's understanding of sunk cost decisions. *Thinking and Reasoning, 10*: 147-174.

Kahneman, Slovic, & Tversky, eds. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Klauer, Musch, & Naumer (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107*: 852-884.

Leibniz (1702/1989). Letter to Queen Sophie Charlotte of Prussia, on what is independent of sense and matter. In Leibniz, *Philosophical essays* (pp. 186-192). Hackett.

Leibniz (1714/1989). Principles of nature and grace, based on reason. In Leibniz, *Philosophical essays* (pp. 206-213). Hackett.

Lichtenstein, Slovic, Fischhoff, Layman, & Combs (1978). Judged Frequency of Lethal Events. *Journal of Experimental Psychology: Human Learning and Memory, 4(6)*: 551-578.

Lieberman (2003). Reflective and reflexive judgment processes: A social cognitive neuroscience approach. In Forgas, Williams, & von Hippel (eds.), *Social judgments: Implicit and explicit processes* (pp. 44-67). Cambridge University Press.

Lieberman (2009). What zombies can't do: A social cognitive neuroscience approach to the irreducibility of reflective consciousness. In Evans & Franklin (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 293-316). Oxford University Press.

Mercier & Sperber (2009). Intuitive and reflective inferences. In Evans & Franklin (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 149-170). Oxford University Press.

Michotte (1963). *The perception of causality*. Basic Books.

Newstead (2000). Are there two different types of thinking? *Behavioral and Brain Sciences, 23*: 690-691.

Newstead, Pollard, & Evans (1992). The source of belief bias effects in syllogistic reasoning. *Cognition, 45*: 257-284.

Nisbett, Peng, Choi, & Norenzayan (2001). Culture and systems of thought: Holistic vs analytic cognition. *Psychological Review, 108*: 291-310.

Nosek (2007). Implicit-explicit relations. *Current Directions in Psychological Science, 16*: 65-69.

Osman (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin and Review, 11*: 988-1010.

Pohl, ed. (2005). *Cognitive illusions: a handbook on fallacies and biases in thinking, judgment and memory*. Psychology Press.

Reber (1993). *Implicit learning and tacit knowledge*. Oxford University Press.

Rips (1994). *The psychology of proof: Deductive reasoning in human thinking*. MIT Press.

Saunders (2009). Reason and intuition in the moral life: A dual-process account of moral justification. In Evans & Franklin (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 335-354). Oxford University Press.

Schwarz, Bless, Strack, Klumpp, Rittenauer-Schatka, & Simons (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology, 61*: 195–202.

Schwarz & Clore (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45(3)*: 513–523.

Schwarz (1996). Cognition and communication: Judgmental biases, research methods, and the logic of conversation. Erlbaum.

Schwarz & Vaughn (2002). The availability heuristic revisited: Ease of recall and content of recall as distinct sources of information. In Gilovich, Griffin, & Kahneman (eds.), *Heuristics & biases: The psychology of intuitive judgment* (pp. 103–119). Cambridge University Press.

Sherry & Schacter (1987). The evolution of multiple memory systems. *Psychological Review, 94*: 439-454.

Sloman (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*: 1-23.

Sloman (2002). Two systems of reasoning. In Gilovich, Griffin, & Kahneman (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press.

Slovic, Finucane, Peters, & MacGregor (2002). Rational Actors or Rational Fools: Implications of the Affect Heuristic for Behavioral Economics. *Journal of Socio-Economics, 31*: 329-342.

Smith & Collins (2009). Dual-process models: A social psychological perspective. In Evans & Franklin (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 197-216). Oxford University Press.

Smith & DeCoster (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*: 108-131.

Smith & Semin (2004). Socially situated cognition: Cognition in its social context. *Advances in experimental social psychology, 36*: 53-117.

Smolensky (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*: 1-23.

Stanovich (1999). *Who is rational? Studies of individual differences in reasoning*. Psychology Press.

Stanovich (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago University Press.

Stanovich (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In Evans & Franklin (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 55-88). Oxford University Press.

Strack & Deutsch (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*: 220-247.

Strack, Martin, & Schwarz (1988). Priming and communication: The social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology, 1*: 429–442.

Stevens (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. Wiley.

Sun (2001). *Duality of mind: A bottom-up approach towards cognition*. Psychology Press.

Sun, Lane, & Mathews (2009). The two systems of learning: An architectural perspective. In Evans & Franklin (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 239-262). Oxford University Press.

Toates (2004). 'In two minds' - consideration of evolutionary precursors permits a more integrative theory. *Trends in Cognitive Sciences, 8*: 57.

Toates (2006). A model of the heirarchy of behaviour, cognition, and consciousness. *Consciousness and Cognition, 15*: 75-118.

Thompson (2009). Dual-process theories: A metacognitive perspective. In Evans & Franklin (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 171-195). Oxford University Press.

Tooby & Cosmides (1992). The psychological foundations of culture. In Barkow, Cosmides, & Tooby (eds.), *The Adapted Mind* (pp. 19-136). Oxford University Press.

Tversky & Kahneman (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology, 5(2)*: 207–232.

Tversky & Kahneman (1982). Judgments of and by representativeness. In Kahneman, Slovic, & Tversky (eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84–98). Cambridge University Press.

Vadenoncoeur & Markovits (1999). The effect of instructions and information retrieval on accepting the premises in a conditional reasoning task. *Thinking & Reasoning, 5*: 97-113.

Verscheuren, Schaeken, & d'Ydewalle (2005). A dual-process specification of causal conditional reasoning. *Thinking & Reasoning, 11*: 239-278.

Whyte (1978). *The unconscious before Freud*. St. Martin's Press.

Wilson (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Belknap Press.

Zajonc (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35(2)*: 151–175.

Zajonc (1997). Emotions. In Gilbert, Fiske, & Lindzey (eds.), *Handbook of social psychology* (4th ed., pp. 591–632). Oxford University Press.

Your Evolved Intuitions

Part of the sequence: Rationality and Philosophy

We have already examined one source of our intuitions: attribute substitution heuristics. Today we examine a second source of our intuitions: *biological evolution*.

Evolutionary psychology

Evolutionary psychology[1] has been covered on Less Wrong many times before, but let's review anyway.

Lions walk on four legs and hunt for food. Skunks defend themselves with a spray. Spiders make webs. Each species is shaped by selection pressures, and is different from that of other species.

Certain evolved psychological mechanisms in humans are part of what makes us like each other and not like lions, skunks, and spiders.

These mechanisms evolved to solve specific adaptive problems. It is not an accident that people around the world prefer calorie-rich foods,[2] that women around the world prefer men with resources,[3] that men around the world prefer women with signs of fertility,[4] or that most of us inherently fear snakes and spiders but not cars and electrical outlets.[5]

An an example of evolutionary psychology at work, consider the 'hunter-gatherer hypothesis' that men evolved psychological mechanisms to aid in hunting, while women evolved psychological mechanisms to aid in gathering.[6] This hypothesis leads to a list of bold predictions. If the hypothesis is correct, then:

1. Men in modern tribal societies should spend a lot of time hunting, and women more time gathering.
2. Humans should show a greater tendency toward strong male coalitions than similar species in which males do not hunt much, because strong male coalitions are required to hunt big game.
3. Because meat from most game comes in quantities larger than a single hunter can consume, and because hunting success is highly variable (one week may be a success, but perhaps not the next week), humans should exhibit food sharing and reciprocal altruism.
4. We should expect to see a sexual division of labor, due to the different traits conducive for hunting vs. gathering.
5. Men should exploit status gains to be had from 'showing off' large hunting successes.
6. Men should have superior cognitive ability to navigate across large distances and perform 3D mental rotation tasks required for throwing spears and similar hunting acts. Women should have superior cognitive ability with spacial location memory and object arrays.

And as it turns out, *all* these predictions are correct.[7] (And no, evolutionary psychologists do not only offer 'postdictions' or 'just so' stories. Besides, probability

theory does not have separate categories for 'predictions' and 'postdictions'.)

Kin loyalty

Consider the intuition that we have more responsibility for the well-being of our close relatives than for the well-being of distant relatives or strangers. We would *expect* human evolution to produce exactly such an intuition given Hamilton's rule, which states that the reproductive cost to an agent is less than the genetic relatedness of the recipient to the agent multiplied by the additional reproductive benefit gained by the recipient of the altruistic act.

That's a mouthful, so instead let me illustrate the consequences of Hamilton's rule:

> Imagine that you pass by a river and notice that some of your genetic relatives are drowning in a ferocious current. You could jump in the water to save them, but you would pay with your own life. According to Hamilton's rule, selection will favor decision rules that, on average, result in your jumping into the water to save three of your brothers, but not one. You would be predicted *not* to sacrifice your own life for just *one* brother, because that would violate Hamilton's rule. Using the logic of Hamilton's rule, evolved decision rules should lead you to sacrifice your own life for five nieces or nephews, but you would have to save nine first cousins before you would sacrifice your own life.[8]

Hamilton's rule has indeed been observed at work in a wide variety of contexts.[9]

My intuition that I am more responsible for the well-being of my brother than my cousin, and more responsible for the well-being of my cousin than a stranger, looks like a good candidate for an evolved intuition.

Essentialism

Uneducated people around the world believe that organisms come in discrete packets, and that each species has an 'essence' that produces its form and abilities. The intuitive appeal of this *essentialism* often trumps the explicitly learned gradualism of biological evolution. Even someone who has read Richard Dawkins argue against essentialism might find himself the very next day stuck in essentialist thinking. Why? Many researchers have suggested that an evolved, intuitive 'folk biology' is responsible.[10]

These essentialist intuitions emerge early in life across all cultures we have studied.[11] For example, children may believe that

> ...if you remove the insides of a dog, it loses its 'essence' and is no longer really a dog anymore - it can't bark or bite. But if you remove its outsides or change its external appearance so that it doesn't look like a dog, children still believe that it has retained its essential 'dogness.'[12]

Many researchers think that essentialist intuitions evolved because it's useful for humans to respond to organisms in this way. With essentialist thinking, we can very

quickly drop organisms into categories concerning what we can and can't eat, what we can capture, what might capture us, and so on.

Essentialism has had a long-lasting hold on the minds of many philosophers, and greatly influenced their conclusions even after Darwin.

Heuristics and biases

Human reasoning is subject to a long list of biases. Why did we evolve such faulty thinking processes? Aren't false beliefs bad for survival and reproduction?

Many researchers suggest that while humans are poor at formal logic and Bayesian inference, humans display a kind of 'ecological rationality'.[13]

> Over evolutionary time, the human environment has had certain statistical regularities: Rain often followed thunder, violence sometimes followed angry shouts, sex sometimes followed prolonged eye contact, dangerous bites often followed getting too close to a snake, and so on. These statistical regularities are called *ecological structure*. Ecological rationality consists of evolved mechanisms containing design features that utilize ecological structure to facilitate adaptive problem solving.
>
> The shape and form of cognitive mechanisms, in other words, coordinate with the recurring statistical regularities of the ancestral environments in which humans evolved. We fear snakes and not electrical outlets...
>
> [Moreover], theories of formal logic that are content independent... are exceptionally poor at solving real adaptive problems. The world is full of logically arbitrary relationships: Dung happens to be potentially dangerous to humans, for example, but provides a hospitable home for dung flies. So applying formal logic cannot in principle solve the adaptive problem of avoiding dung. The only thing that can solve it is a content-specific mechanism, one that has been built over evolutionary time to capitalize on the recurring statistical regularities associated with dung as it interacted with our hominid ancestors.[14]

Conclusion

Our brains may have evolved intuition-generating mechanisms that worked for solving particular adaptive problems in the ancestral environment, but we may not have evolved psychological mechanisms that generate accurate intuitions useful for doing philosophy. For example, it seems unlikely that we evolved a mechanism that gives us reliable intuitions about the metaphysical possibility or impossibility of zombies.

Next post: Intuition and Unconscious Learning

Previous post: How You Make Judgments

Notes

[1] Recent introductions to the field include: Buss (2011); Workman & Reader (2008); Gaulin & McBurney (2003). It is also worth mentioning one of the major problems with evolutionary psychology. Evolutionary psychologists tend to focus on subjects that are difficult to test because they are *uniquely* human but also *universally* human, which is bad for testability (see here and here). For other difficulties, see Problems in Evolutionary Psychology.

[2] Birch (1999); Krebs (2009).

[3] Buss et al. (1990); Buss & Schmitt (1993); Khallad (2005); Gottschall et al. (2003); Gottschall et al. (2004); Kenrick et al. (1990); Gustavsson & Johnsson (2008); Wiederman (1993); Badahdah & Tiemann (2005); Marlowe (2004); Fisman et al. (2006); Asendorpf et al. (2010); Bokek-Cohen et al. (2007); Pettay et al. (2007).

[4] Signs of fertility that men prefer include youth (Buss 1989a; Kenrick & Keefe 1992; Kenrick et al. 1996), clear and smooth skin (Sugiyama 2005; Singh & Bronstad 1997; Fink & Neave 2005; Fink et al. 2008; Ford & Beach 1951; Symons 1995), facial femininity (Gangestad & Scheyd 2005; Schaefer et al. 2006; Rhodes 2006), long legs (Fielding et al. 2008; Sorokowski & Pawlowski 2008; Bertamini & Bennett 2009; Swami et al. 2006), and a low waist-to-hip ratio (Singh 1993, 2000; Singh & Young 1995; Jasienska et al. 2004; Singh & Randall 2007; Connolly et al 2000; Furnham et al 1997). Even men blind from birth prefer a low waist-to-hip ratio (Karremans et al. 2010). Note that standards for beautiful faces emerge before cultural can have much effect (Langlois et al. 1990) and that standards of beauty are relatively consistent across cultures (Cunningham et al. 1995; Cross & Cross 1971; Jackson 1992; Jones 1996; Thakerar & Iwawaki 1979).

[5] Buss (2011), pp. 92-94.

[6] Buss (2011), p. 85.

[7] Evidence cited by prediction number. 1: Hewlett (1991); Lee (1979). 2: Tooby & DeVore (1987). 3: Trivers (1971). 4: Roskraft et al. (2004); Tooby & DeVore (1987). 5: Hawkes (1991); Wiessner (2002). 6: Silverman & Philips (1998); Silverman et al. (2000); Eals & Silverman (1994); Silverman et al. (2007); New et al. (2007); Silverman & Choi (2005); Lippa et al. (2010).

[8] Buss (2011), p. 238-239.

[9] Buss (2011) calls Hamilton's theory of inclusive fitness (expressed in Hamilton's rule) "the single most important theoretical revision of Darwin's theory of natural selection in the past century" (p. 239). For a review of some of the evidence that supports Hamilton's rule, see Buss (2011), chapter 8.

[10] Atran (1998); Berlin (1992); Keil (1995); Medin & Atran (1999).

[11] Sperber & Hirschfeld (2004).

[12] Buss (2011), p. 73.

[13] Tooby & Cosmides (1998). Haselton et al. (2009) say humans are 'adaptively biased,' while Kenrick et al. (2009) say we are 'adaptively rational.'

[14] Buss (2011), pp. 396-397.

References

Asendorpf, Penke, & Back (2010). From dating to mating and relating: Predictors of initial and long-term outcomes of speed dating in a community sample. *European Journal of Personality*.

Atran (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences, 21*: 547-609.

Badahdah & Tiemann (2005). Mate selection criteria among Muslims living in America. *Evolution and Human Behavior, 26*: 432-440.

Berlin (1992). *Ethnobiological classification*. Princeton University Press.

Bertamini & Bennett (2009). The effect of leg length on perceived attractiveness of simplified stimuli. *Journal of Social, Evolutionary, and Cultural Psychology, 3*: 233-250.

Birch (1999). Development of food preferences. *Annual Review of Nutricion, 19*: 41-62.

Bokek-Cohen, Peres, & Kanazawa (2007). Rational choice and evolutionary psychology as explanations for mate selectivity. *Journal of Social, Evolutionary, and Cultural Psychology, 2*: 42-55.

Buss (1989). Sex differences in human mate preferences: Evolutionary hypotheses testing in 37 cultures. *Behavioral and Brain Sciences, 12*: 1-49.

Buss (2011). *Evolutionary Psychology: The New Science of Mind* (4th ed.). Prentice Hall.

Buss & Schmitt (1993). Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review, 100*: 204-232.

Buss, Abbott, Angleitner, Asherian, Biaggio, et al. (1990). International preferences in selecting mates: A study of 37 cultures. *Journal of Cross-Cultural Psychology, 21*: 5-47.

Connolly, Mealey, & Slaughter (2000). The development of waist-to-hip ratio preferences. *Perspectives in Human Biology, 5*: 19-29.

Cross & Cross (1971). Age, sex, race, and the perception of facial beauty. *Developmental Psychology, 5*: 433-439.

Cunningham, Roberts, Wu, Barbee, & Druen (1995). "Their ideas of beauty are, on the whole, the same as ours": Consistency and variability in the cross-cultural perception of female attractiveness. *Journal of Personality and Social Psychology, 68*: 261-279.

Eals & Silverman (1994). The hunter-gatherer theory of spatial sex differences: Proximate factors mediating the female advantage in recall of object arrays. *Ethology and Sociobiology, 15*: 95-105.

Fielding, Scholling, Adab, Cheng, Lao et al. (2008). Are longer legs associated with enhanced fertility in Chinese women? *Evolution and Human Behavior, 29*: 434-443.

Fink & Neave (2005). The biology of facial beauty. *Internal Journal of Cosmetic Science, 27*: 317-325.

Fink, Matts, Klingenberg, Kuntze, Weege, & Grammar (2008). Visual attention to variation in female skin color distribution. *Journal of Cosmetic Dermatology, 7*: 155-161.

Fisman, Iyengar, Kamenica, & Simonson (2006). Gender differences in mate selection: Evidence from a speed dating experiment. *The Quarterly Journal of Economics, 121*: 673-697.

Ford & Beach (1951). *Patterns of Sexual Behavior*. Harper & Row.

Furnham, Tan, & McManus (1997). Waist-to-hip ratio and preferences for body shape: A replication and extension. *Personality and Individual Differences, 22*: 539-549.

Gangestad & Scheyd (2005). The evolution of human physical attractiveness. *Annual Review of Anthropology, 34*: 523-548.

Gaulin & McBurney (2003). *Evolutionary Psychology* (2nd ed.) Prentice Hall.

Gottschall, Berkey, Cawson, Drown, Fleischner, et al. (2003). Patterns of characterization in folktales across geographic regions and levels of cultural complexity: Literature as a neglected source of quantitative data. *Human Nature, 14*: 365-382.

Gottschall, Martin, Quish, & Rea (2004). Sex differences in mate choice criteria are reflected in folktales from around the world and in historical European literature. *Evolution and Human Behavior, 25*: 102-112.

Gustavsson & Johnsson (2008). Mixed support for sexual selection theories of mate preferences in the Swedish population. *Evolutionary Psychology, 6*: 454-470.

Haselton , Bryant, Wilke, Frederick, Galperin, Franenhuis, & Moore (2009). Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition, 27*: 733-763.

Hawkes (1991). Showing off: Tests of another hypothesis about men's foraging goals. *Ethology and Sociobiology, 11*: 29-54.

Hewlett (1991). *Intimate Fathers: The nature and context of Aka pygmy paternal infant care*. University of Michigan Press.

Jackson (1992). *Physical appearance and gender: Sociobiological and sociocultural perspectives*. State University of New York Press.

Jasienska, Ziomkiewicz, Ellison, Lipson, & Thune (2004). Large breasts and narrow waists indicate high reproductive potential in women. *Proceedings of the Royal Society of London, B, 271*: 1213-1217.

Jones (1996). *Physical attractiveness and the theory of sexual selection*. University of Michigan Press.

Karremans, Frankenhuis, & Arons (2010). Blind men prefer a low waist-to-hip ratio. *Evolution and Human Behavior, 31*: 182-186.

Keil (1995). The growth of understandings of natural kinds. In Sperber, Premack, & Premack (eds.), *Causal cognition*. Clarendon Press.

Kenrick, Sadalla, Groth, & Trost (1990). Evolution, traits, and the stages of human courtship: Qualifying the parental investment model. *Journal of Personality, 58*: 97-116.

Kenrick, Keefe, Gabrielidis, & Cornelius (1996). Adolescents' age preferences for dating partners: Support for an evolutionary model of life-history strategies. *Child Development, 67*: 1499-1511.

Kenrick, Griskevicius, Sundie, Li, Li, & Neuberg (2009). Deep rationality: The evolutionary economics of decision making. *Social Cognition, 27*: 764-785.

Kenrick & Keefe (1992). Age preferences in mates reflect sex differences in reproductive strategies. *Behaivoral and Brain Sciences, 15*: 75-133.

Khallad (2005). Mate selection in Jordan: Effects of sex, socio-economic status, and culture. *Journal of Social and Personal Relationships, 22*: 155-168.

Krebs (2009). The gourmet ape: Evolution and human food preferences. *American Journal of Clinical Nutrition, 90*: 707S-711S.

Langlois, Roggman, & Reiser-Danner (1990). Infants' differential social responses to attractive and unattractive faces. *Developmental Psychology, 26*: 153-159.

Lee (1979). *The !Kung San: Men, women, and working in a foraging society*. Cambridge University Press.

Lippa, Collaer, & Peters (2010). Sex differences in mental rotation and line angle judgments are positively associated with gender equality and economic development across 53 nations. *Archives of Sexual Behavior, 39*: 990-997.

Marlowe (2004). Mate preferences among Hadza hunter-gatherers. *Human Nature, 4*: 365-376.

Medin & Atran, eds. (1999). *Folkbiology*. MIT Press.

New, Krasnow, Truxaw, & Gaulin (2007). Spatial adaptations for plant foraging: Women excel and calories count. *Proceedings of the Royal Society, B, 274*: 2679-2684.

Pettay, Helle, Jokela, & Lummaa (2007). Natural selection on female life-history traits in relation to socio-economic class in pre-industrial human populations. *Plos ONE*, July: 1-9.

Rhodes (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology, 57*: 199-226.

Roskraft, Hagen, Hagen, & Moksnes (2004). Patterns of outdoor recreation activities among Norwegians: An evolutionary approach. *Ann. Zool. Fennici, 41*: 609-618.

Schaefer, Fink, Grammar, Mitteroecker, Gunz, & Bookstein (2006). Female appearance: Facial and bodily attractiveness as shape. *Psychology Science, 48*: 187-205.

Silverman & Philips (1998). The evolutionary psychology of spatial sex differences. In Crawford & Krebs (eds.), *Handbook of evolutionary psychology* (pp. 595-612). Erlbaum.

Silverman & Choi (2005). Locating places. In Buss (ed.), *Handbook of evolutionary psychology* (pp. 177-199). Wiley.

Silverman, Choi, Mackewn, Fisher, Moro, & Olshanksy (2000). Evolved mechanisms underlying wayfinding: Further studies on the hunter-gatherer theory of spatial sex differences. *Evolution and Human Behavior, 21*: 201-213.

Silverman, Choi, & Peters (2007). On the universality of sex-related spatial competencies. *Archives of Human Sexuality, 36*: 261-268.

Singh (1993). Adaptive significance of waist-to-hip ratio and female physical attractiveness. *Journal of Personality and Social Psychology, 65*: 293-307.

Singh (2000). Waist-to-hip ratio: An indicator of female mate value. *International Research Center for Japanese Studies, International Symposium 16*: 79-99.

Singh & Randall (2007). Beauty is in the eye of the plastic surgeon: Waist-to-hip ratio (WHR) and women's attractiveness. *Personality and Individual Differences, 43*: 329-340.

Singh & Bronstad (1997). Sex differences in the anatomical locations of human body scarification and tattooing as a function of pathogen prevalence. *Evolution and Human Behavior, 18*: 403-416.

Singh & Young (1995). Body weight, waist-to-hip ratio, breasts, and hips: Role in judgments of female attractiveness and desirability for relationships. *Ethology and Sociobiology, 16*: 483-507.

Sperber & Hirschfeld (2004). The cognitive foundations of cultural stability and diversity. *Trends in Cognitive Science, 8*: 40-46.

Sorokowski & Pawlowski (2008). Adaptive preferences for leg length in a potential partner. *Evolution and Human Behavior, 29*: 86-91.

Sugiyama (2005). Physical attractiveness in adaptationist perspective. In Buss (ed.), *The handbook of evolutionary psychology* (pp. 292-342). Wiley.

Swami, Einon, & Furnham (2006). The leg-to-body ratio as a human aesthetic criterion. *Body Image, 3*: 317-323.

Symons (1995). Beauty is in the adaptations of the beholder: The evolutionary psychology of human female sexual attractiveness. In Abramson & Pinkerton (eds.), *Sexual nature, sexual culture* (pp. 80-118). University of Chicago Press.

Thakerar & Iwawaki (1979). Cross-cultural comparisons in interpersonal attraction of females toward males. *Journal of Social Psychology, 108*: 121-122.

Tooby & Cosmides (1998). *Ecological rationality and the multimodular mind: Grounding normative theories in adaptive problems*. Unpublished manuscript, University of California, Santa Barbara.

Tooby & DeVore (1987). The reconstruction of hominid behavioral evolution through strategic modeling. In Kinzey (ed.), *The evolution of human behavior* (pp. 183-237). State University of New York Press.

Trivers (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*: 35-57.

Wiederman (1993). Evolved gender differences in mate preferences: Evidence from personal advertisements. *Ethology and Sociobiology, 14*: 331-352.

Wiessner (2002). Hunting, healing, and *hzaro* exchange: A long-term perspective on !Kung large-game hunting. *Evolution and Human Behavior, 20*: 121-128.

Workman & Reader (2008). *Evolutionary Psychology: An Introduction* (2nd ed.). Cambridge University Press.

Intuition and Unconscious Learning

Part of the sequence: Rationality and Philosophy

We have already examined two sources of our intuitions: the attribute substitution heuristics and our evolved psychology. Today we look at a third source of our intuitions: *unconscious learning*.

Unconscious learning

The 'learning perspective' on intuition is compatible with the heuristics and biases literature and with evolutionary psychology, but adds a deeper understanding of what is going on 'under the hood.' The learning perspective says that many intuitions rely on representations that reflect the entirety of experiences stored in long-term memory. Such intuitions merely reproduce statistical regularities in long-term memory.[1]

An example will help explain:

> Assume you run into a man at the 20th anniversary party of your high school class graduation. You immediately sense a feeling of dislike. To avoid getting into a conversation, you signal and shout some words to a couple of old friends sitting at a distant table. While you are walking toward them, you try to remember the man's name, which pops into your mind after some time; and suddenly, you also remember that it was he who always did nasty things to you such as taking your secret letters and showing them to the rest of the class. You applaud the product of your intuition (the immediate feeling) that has helped you to make the right decision (avoiding interaction). Recall of prior experiences was not necessary to make this decision. The decision was solely based on a feeling, which reflected prior knowledge without awareness.[2]

Learning perspective theorists would suggest that your feeling of dislike - your intuition that you shouldn't talk to the man - came from something like an (unconscious) regularities analysis of your experiences with that man that were stored in long-term memory, and those experiences turned out to be mostly negative. As such, your intuition can make use of rapid parallel processing to draw on the whole sum of experiences in long-term memory, rather than using a slower, sequential-processing judgment algorithm.

It is difficult to track the source of any *particular* intuition (though we can try[3]), but there is evidence to suggest that unconscious learning is a common source of our intuitions.

Stock tickers

In a series of experiments,[4] researchers asked subjects to watch a series of advertisements. They warned subjects that a (fictional) stock ticker at the bottom of the screen would be added as a distractor (screenshot), and that they would be

quizzed on the advertisements later. After being quizzed on the advertisements, subjects were surprised by a quiz on their attitudes toward the fictional stocks. Post-experiment interviews confirmed that subjects had not intended to form attitudes toward the stocks.

Subjects watched 20 to 40 advertisements while the 'distractor' stock ticker displayed 70 to 140 return values for 4 to 8 shares. As the independent variable, researchers varied the return values (and thus their sum, average, frequency, and peaks).

When given the surprise quiz on their attitudes toward the fictional stocks, researchers found a perfect rank correlation between the subjects' mean evaluation of the shares and the sums of their returns. This was the case even though subjects had no concrete memories of the share returns, and could not remember the sum or average values. Subjects reported they had relied on their 'gut reaction' or 'intuitive feeling.'

Here, it does not seem that subjects were able to arrive at such accurate intuitions by way of a specific evolved intuition or an attribute substitution heuristic. Instead, they seem to have drawn upon their unconscious learning system without knowing that they were doing so.

Base rate neglect

Consider this problem:

> If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms or signs?[5]

Among 60 Harvard medical students and staff, almost half judged that the person has the disease with .95 probability, while only 18% got the correct answer: .02. This is an example of base rate neglect. Subjects based their judgment mostly on the evidence from the test, and ignored the strong evidence from the base rate (1/1000).

Base rate sensitivity improves when such problems are framed in terms of frequencies rather than probabilities, but even then base rate neglect occurs in about half of subjects.[6]

Subjects further improve their statistical judgments when they are allowed learn the distribution of a variable by their own sampling, and become even more sensitive to base rates.[7]

In a related study,[8] researchers had subjects perform several behaviors many times, and then asked them to estimate behavior frequency. Half of the subjects were asked to make spontaneous judgments, and half were asked to deliberate carefully about their judgments. In the deliberation condition, judgments were biased by the availability heuristic. Judgments from the spontaneous judgment condition were more accurate, and seemed to reflect unconscious recall of the totality of behaviors just performed, stored by unconscious learning.

Conclusion

These and many others studies[9] suggest that sometimes our feelings and intuitive judgments arise from unconscious parallel processing of all (or many) of the experiences relevant to a given judgment stored in our long-term memory.

Later we'll examine how this understanding of intuition (along with the perspectives from attribute substitution heuristics and evolutionary psychology) gives us some clues about how much trust we should put in our intuitions under particular conditions, and how we can train our intuitions to be more accurate.[10]

Next post: When Intuitions Are Useful

Previous post: Your Evolved Intuitions

Notes

[1] Betsch et al. (2004); Betsch & Haberstroh (2005); Betsch (2007); Klein (1999); Hogarth (2001, 2007); Epstein (2007). For an overview of the neuroscience of unconscious learning, see Volz & Cramon (2007). For an overview of the relation between emotion and intuition, see Zeelenberg et al. (2007).

[2] Betsch (2007), p. 6.

[3] Hamm (2007).

[4] Betsch et al. (2001, 2003, 2007).

[5] Tversky & Kahneman (1982), p. 154.

[6] Gigerenzer & Hoffrage (1995).

[7] Betsch et al. (1998); Fiedler et al. (2000).

[8] Haberstroh et al. (2006).

[9] Plessner et al. (2007); Raab & Johnson (2007); Glöckner (2007). Also see research on the 'sample size effect': Kaufmann & Betsch (2009).

[10] Hogarth (2001, 2007); Erev et al. (2007).

References

Betsch (2007). The nature of intuition and its neglect in research on judgment and decision making. In Plessner, Betsch, & Betsch (eds.), *Intuition in Judgment and Decision Making* (pp. 3-22). Psychology Press.

Betsch, Plessner, Schwieren, & Gütig (2001). I like it but I don't know why: A value-account approach to implicit attitude formation. *Personality and Social Psychology Bulletin, 27*: 242–253.

Betsch, Hoffmann, Hoffrage, & Plessner (2003). Intuition beyond recognition: When less familiar events are liked more. *Experimental Psychology, 50*: 49–54.

Betsch, Plessner, & Schallies (2004). The value-account model of attitude formation. In Haddock & Miao (eds.), *Contemporary perspectives on the psychology of attitudes* (pp. 251-274). Psychology Press.

Betsch & Haberstroh, eds. (2005). *The routines of decision making*. Psychology Press.

Betsch, Kaufmann, Lindow, Plessner, & Hoffmann (2006). Different principles of information integration in implicit and explicit attitude formation. *European Journal of Social Psychology, 36*: 887–905.

Betsch, Biel, Eddelbüttel, & Mock (1998). Natural sampling and base-rate neglect. *European Journal of Social Psychology, 28*: 269–273.

Epstein (2007). Intuition from the perspective of cognitive-experiential self-theory. In Plessner, Betsch, & Betsch (eds.), *Intuition in Judgment and Decision Making* (pp. 23-37). Psychology Press.

Erav, Shimonowitch, Schurr, & Hertwig (2007). Base rates: How to make the intuitive mind appreciate or neglect them. In Plessner, Betsch, & Betsch (eds.), *Intuition in Judgment and Decision Making* (pp. 135-148). Psychology Press.

Fiedler, Brinkmann, Betsch, & Wild (2000). A sampling approach to biases in conditional probability judgments: Beyond base-rate neglect and statistical format. *Journal of Experimental Psychology, General, 129*: 399–418.

Gigerenzer & Hoffrage (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*: 684–704.

Glöckner (2007). Does intuition beat fast and frugal heuristics? A systematic empirical analysis. In Plessner, Betsch, & Betsch (eds.), *Intuition in Judgment and Decision Making* (pp. 309-326). Psychology Press.

Haberstroh, Betsch, & Aarts (2006). When guessing is better than thinking: Multiple bases for frequency judgments. Unpublished manuscript.

Hamm (2007). Cue by hypothesis interactions in descriptive modeling of unconscious use of multiple intuitive judgment strategies. In Plessner, Betsch, & Betsch (eds.), *Intuition in Judgment and Decision Making* (pp. 55-70). Psychology Press.

Hogarth (2001). *Educating Intuition*. University of Chicago Press.

Hogarth (2007). On the learning of intuition. In Plessner, Betsch, & Betsch (eds.), *Intuition in Judgment and Decision Making* (pp. 91-106). Psychology Press.

Klein (1999). *Sources of power. How people make decisions*. MIT press.

Kaufmann & Betsch (2009). Origins of the sample-size effect in explicit evaluative judgments. *Experimental Psychology, 56*: 344-353.

Plessner, Betsch, Schallies, & Schwieren (2007). Automatic online formation of implicit attitudes toward politicians as a basis for intuitive voting behavior. In Plessner, Betsch, & Betsch (eds.), *Intuition in Judgment and Decision Making* (pp. 107-117). Psychology Press.

Raab & Johnson (2007). Implicit learning as a means to intuitive decision making in sports. In Plessner, Betsch, & Betsch (eds.), *Intuition in Judgment and Decision Making* (pp. 119-133). Psychology Press.

Tversky & Kahneman (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Volz & Cramon (2007). Can neuroscience tell a story about intuition? In Plessner, Betsch, & Betsch (eds.), *Intuition in Judgment and Decision Making* (pp. 71-87). Psychology Press.

Zeelenberg, Nelissen, & Pieters (2007). Emotion, motivation, and decision making: a feeling-is-for-doing approach. In Plessner, Betsch, & Betsch (eds.), *Intuition in Judgment and Decision Making* (pp. 173-189). Psychology Press.

When Intuitions Are Useful

Part of the sequence: Rationality and Philosophy

In this series, I have examined how intuitions work so that I can clarify how rationalists[1] *should* and *shouldn't* use their intuitions[2] when solving philosophical problems. Understanding the cognitive algorithms that generate our intuitions can dissolve traditional philosophical problems. As Brian Talbot puts it:

> ...where psychological research indicates that certain intuitions are likely to be inaccurate, or that whole categories of intuitions are not good evidence, this will overall benefit philosophy. This has the potential to resolve some problems due to conflicting intuitions, since some of the conflicting intuitions may be shown to be unreliable and not to be taken seriously; it also has the potential to free some domains of philosophy from the burden of having to conform to our intuitions, a burden that has been too heavy to bear in many cases...[3]

Knowing how intuitions work can also tell us something about how we can train them to make them render more accurate judgments.[4]

Problems with intuition

In most philosophy, intuitions play the role that observations do in science: they support and undermine various theories.[5] Conceptual analyses are rejected when intuitive counterexamples are presented. Moral theories are rejected when they lead to intuitively revolting results. Theories of mind and language and metaphysics rise and fall depending on how well they can be made to fit our intuitions, even in bizarre science fiction hypothetical scenarios.[6]

But why trust our intuitions? Our intuitions often turn out to contradict each other,[7] or they are contradicted by empirical evidence,[8] or they vary between people and between groups of people.[9] Compared to scientific methods, the philosopher's use of intuitions as his primary tool doesn't seem to have been very productive.[10] Also, we can't calibrate our intuitions, because wherever we *have* a non-intuition standard against which to calibrate our intuitions, we don't need to use intuition in the first place.[11] Moreover, philosophers have typically known very little about where their intuitions come from, and why they should trust them in the first place![12]

Defenders of intuitionist philosophy reply that we can't do philosophy without intuitions.[13] Others point out that we have similar worries about the reliability of of sense perception.[14] But these replies do not solve the problem. As Talbot says,[3] these responses "give us reasons to *want* to trust intuitions... but no evidence that they are particularly reliable or useful."

The way forward is not to give *a priori* arguments for or against the use of intuitions. The way forward is to explore what cognitive science can tell us about how our intuitions work (as we've been doing) so that we have some idea about when they work and when they don't.

What is intuition?

But first, what *is* this 'intuition' we're talking about? Definitions of 'intuition' abound.[15]

In 2008, Eliezer wrote a post about the 'intuitions' behind utilitarianism. He responded to a critic who used the word 'intuition' in a very broad sense - perhaps meaning *all thoughts and seemings.* But when we use the word so broadly, then the word is not so useful anymore - like the word 'god' after you've redefined it to mean 'a higher power'. When I talk about 'intuition', I want to talk about intuition in a more specific and useful way (as Eliezer would appreciate[16]).

But we don't need to argue about definitions. We can use stipulation. We can argue about the substance rather than the symbol.

For now, let's think of the thing we're investigating as the *seeming to be true* of some proposition due to an *opaque* mental process (and not memory or perception). After all, if intuitions were transparent, we could just point to *the things that ground them* as evidence, and the intuitions themselves would add no weight of their own to our evidence.[3]

When intuitions are useful

As we are discussing it, an intuition is a judgment that springs from the unconscious. And from where does the unconscious get its judgments? From evolution[17] and from unconscious learning[18] and from attribute substitution heuristics.[19]

Before considering how these sources of intuitions make them *unsuitable* for many of their popular uses in philosophy, let's acknowledge how effective intuitions are in *some* situations.

Familiarity with recent cognitive science has led many to conclude that "being more analytic and less intuitive should help you to develop more effective and rewarding solutions."[20] But recent investigations have located a few circumstances in which intuitions outperform considered judgments.

In one study, basketball experts asked to make spontaneous predictions about the outcomes of a basketball tournament made more accurate predictions than those asked to deliberate carefully about their predictions.[21] Other studies on intuition vs. deliberation have found intuition 'winning' on tests of certain kinds of face recognition,[22] route recognition,[23] and voice recognition,[24] while deliberation 'won' on tests of subadditivity probability judgments,[25] raffle-winning probability judgments,[26] quantity estimation,[27] picture recognition,[28] conjunctions and disjunctions,[29] and conditional inferences.[30]

Better-supported is a trend in research which finds that when selecting products to to take home with us, we end up feeling more satisfied with our choice if we made it using intuition rather than a conscious process of weighing pros and cons, costs and benefits.[31]

And if you're trying to avoid collisions or catch a baseball, you're better off acting on your split-second intuition than trying to calculate the physics of moving objects.[32]

Some authors have suggested other, very specific domains in which intuition may surpass the accuracy of considered judgment,[33] but these claims are not yet well substantiated.

You may have noticed that the domains in which intuition might excel are not particular relevant to solving philosophical problems. In the next post, we'll begin to examine the ways in which intuitions can lead us astray when doing philosophy.

Next post: Concepts Don't Work That Way

Previous post: Intuitions and Unconscious Learning

Notes

[1] I use the term 'rationalists' as Less Wrong uses the term, not as the mainstream philosophical community uses the term. As Less Wrong uses the term, a 'rationalist' is someone devoted to the craft of refining their rationality by counteracting known cognitive biases and trying to make their beliefs and decisions track with technically correct beliefs and decisions (defined with reference to, for example, Bayes' theorem and decision theory).

[2] In the preface of Plessner et al. (2009), the authors provide a handy list of search terms for those who wish to research the science of intuition on their own: "unconscious perceptions, 'blindsight,' pattern recognition, instinct, automatic processing, experiential knowing, tacit knowledge, religious experiences, emotional intelligence, nonverbal communication, clinical diagnoses, 'thin slices of behavior,' spontaneous trait inferences, the 'mere exposure' effect, the primacy of affect, 'thinking too much,' priming, feelings as information, implicit attitudes, expertise, creativity, and the 'sixth sense.'" They recommend the following sources as "excellent overviews" for studying these terms and ideas: Bastick (1982); Davis-Floyd & Arvidson (1992); Hogarth (2001); Myers (2002); Wilson (2002).

[3] Talbot (2009).

[4] Hogarth (2001, 2007).

[5] Cummins (1998); Talbot (2009).

[6] Talbot (2009) provides a nice little summary of how intuitions are used in philosophy (I've changed his citations to the original articles in some cases): "Intuitions about understanding Chinese are used by John Searle to argue against what he calls “strong AI” (Searle, 1980). In the philosophy of action, intuitions about playing video games are used by Michael Bratman to argue that we can try to do something without intending to do it (Bratman, 1987). Intuitions are used as counter-evidence against compatibilism (Bok, 1998). One of the most famous use of intuitions as counter evidence comes from epistemology: Gettier cases (Gettier, 1963). In metaphysics, intuitions are appealed to to argue for theories of causation (e.g., Lewis, 1973), and against them (by pointing out that they have counter-intuitive consequences, such as transitivity) (Hall, 2000). In ethics, Judith Jarvis Thomson uses intuitions about violinists and carpets to argue for her claim that abortion can be morally acceptable despite having a right to life (Thomson, 1971). Bernard Williams uses intuitions about killing rebels as counter-evidence against utilitarianism (Williams & Smart, 1973). In the philosophy of language, Tyler Burge uses intuitions about “arthritis” to argue for meaning externalism (Burge, 1979), and Saul Kripke uses intuitions about Gödel to argue against a descriptivist view of names (Kripke, 1972). This list goes on and on."

[7] Suppes (1984); Cummins (1998).

[8] Wisniewski (1998); Hastie & Dawes (2009); Gilovich (1991); Kahneman, Slovic, & Tversky (1982); Nisbett & Ross (1980); Stanovich (2009).

[9] Stich (1988); Weinberg, Nichols, & Stich (2001); Swain, Alexander, & Weinberg (2008).

[10] Bishop & Trout (2004); Miller (2000). Talbot (2009) explains: "Consider the state of philosophy, they say. There is little agreement on most key issues, we have produced few theories that have been very successful or survived criticism, and philosophy has accomplished little of practical significance in the last few hundred years. There is nothing about the subject matter of philosophy that makes these results inevitable; most of us believe that there are answers out there to be found, and at least some philosophical disciplines can produce useful results. This gives us reason to think that we are studying the right stuff but in the wrong way. Some aspects of our methodology –

logic and rigorous thought – are beyond criticism, and thus, they say, the blame for philosophy's lack of success falls on our use of intuitions."

[11] Cummins (1998); Talbot (forthcoming).

[12] Cummins (1998); Wisniewski (1998).

[13] Sosa (1998); Bealer (1998); BonJour (1998).

[14] Sosa (1998); Pust (2000).

[15] Bealer (1996) refers to intuitions as *a priori* seemings of the sort by which De Morgan's laws come to seem true to someone - intellectual seemings, not perceptions or imaginative seemings. Sosa (1998) defines 'intuition' as "noninferential belief due neither to perception nor introspection," but sees intuition as focused on abstract propositions: "At t, S has an intuition that p iff (a) if at t S were merely to understand fully enough the proposition that p (absent relevant perception, introspection, and reasoning), then S would believe that p; (b) at t, S does understand the proposition that p; (c) the proposition that p is an abstract proposition; and (d) at t, S thinks occurrently of the proposition that p (in propria persona, not just by description)." Williamson (2004) describes intuitions as "applications of our ordinary capacities for judgment" and says "when contemporary analytic philosophers run out of arguments, they appeal to intuition." Gopnik & Schwitzgebel (1998) say: "We will call any judgment an intuitive judgment, or more briefly an intuition, just in case that judgment is not made on the basis of some kind of explicit reasoning process that a person can consciously observe. Intuitions are judgments that grow, rather, out of an underground process... that cannot be directly observed." Goldman & Pust (1998) briefly refer to intuitions as "spontaneous moral judgments." Talbot (2009) calls an intuition "a relatively unreflective reaction that a proposition is true or false." Or, more precisely, he says "an intuition is the seeming to be true (although not necessarily acceptance of or belief in) of some proposition that is not a perceptual seeming, or due to conscious recollection, or due entirely to transparent mental processes." Hogarth (2001) says intuitions "are reached with little apparent effort and typically without conscious awareness. They involve little or no conscious deliberation." According to the 'associative learning' view of intuition, intuition draws on the whole stream of past experiences: Betsch et al. (2004); Betsch & Haberstroh (2005). Betsch (2007) offers a definition of intuition from this perspective: "Intuition is a process of thinking. The input to this process is mostly provided by knowledge stored in long-term memory that has been primarily acquired via associative learning. The input is processed automatically and without conscious awareness. The output of the process is a feeling that can serve as abasis for judgments and decisions." Note that the view of intuition from the heuristics and biases perspective (Kahneman & Frederick 2002, 2005) and from the associative learning view are generally not contradictory but rather complementary. Finally, also see surveys of opinion on the nature of intuition, for example Abernathy & Hamm (1995).

[16] In his post, Eliezer responds to his critic's broad definition of 'intuition' like this: "Now 'intuition' is not how I would describe the computations that underlie human morality and distinguish us, as moralists, from an ideal philosopher of perfect emptiness and/or a rock. But I am okay with using the word "intuition" as a term of art, bearing in mind that "intuition" in this sense is not to be contrasted to reason, but is, rather, the cognitive building block out of which both long verbal arguments and fast perceptual arguments are constructed."

[17] The field of evolutionary psychology is rich with candidates for evolutionarily adaptive psychological predispositions, some more thoroughly supported by the evidence than others. For an overview, see Buss (2011); Dunbar & Barrett (2007).

[18] Hogarth (2007); Sloman (1996); Sloman (2002); Evans & Over (1996); Stanovich (2004); Mercier & Sperber (2009); Betsch (2007); Epstein (2007).

[19] Kahneman & Frederick (2002, 2005).

[20] Kardes (2002), p. 402.

[21] Halberstadt and Levine (1999). For an overview of similar results, see Plessner & Czenna (2007). For older studies, see Ambady & Rosenthal (1992).

[22] Clare & Lewandowsky (2004); Fallshore & Schooler (1995); Halbertsadt (2005); Schooler & Engstler-Schooler (1990).

[23] Fiore & Schooler (2002).

[24] Perfect et al. (2002).

[25] Dougherty & Hunter (2003).

[26] Windschitl & Krizan (2005).

[27] Gilbert & Rappoport (1975).

[28] Klimesch (1980); Silverberg & Buchanan (2005).

[29] Kruglanski & Freund (1983).

[30] Schroyens et al. (2003).

[31] Dijksterhuis & van Olden (2005); Dijksterhuis et al. (2006).

[32] Gigerenzer (2007), ch. 1.

[33] Gigerenzer (2007); Gladwell (2005).

References

Abernathy & Hamm (1995). *Surgical intuition: What it is and how to get it*. Hanley & Belfus.

Ambady & Rosenthal (1992). Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis. *Psychological Bulletin, 111*: 256-274.

Bastick (1982). *Intuition: How we think and act*. Wiley.

Bealer (1996). A priori knowledge and the scope of philosophy. *Philosophical Studies, 81(2/3)*: 121-142.

Bealer (1998). Intuition and the autonomy of philosophy. In De Paul & Ramsey (eds.), *Rethinking Intuition*. Rowan and Littlefield.

Betsch (2007). The Nature of Intuition and Its Neglect in Research on Judgment and Decision Making. In Plessner, Betsch, & Betsch (eds.), *Intuition in Judgment and Decision Making* (pp. 3-22). Psychology Press.

Betsch & Haberstroh, eds. (2005). *The routines of decision making*. Lawrence Erlbaum Associates.

Betsch, Plessner, & Schallies (2004). The value-account model of attitude formation. In Maio & Haddock (eds.), *Contemporary perspectives on the psychology of attitudes* (pp. 252–273). Psychology Press.

Bishop & Trout (2004). *Epistemology and the Psychology of Human Judgment*. Oxford University Press.

Bok (1998). *Freedom and Responsibility*. Princeton University Press.

BonJour (1998). *In Defense of Pure Reason*. Cambridge University Press.

Bratman (1987). *Intentions, Plans, and Practical Reason*. Harvard University Press.

Burge (1979). Individualism and the mental. *Midwest Studies in Philosophy, 4*: 73-121.

Buss (2011). *Evolutionary psychology: The new science of mind (4th ed.)*. Prentice Hall.

Clare & Lewandowsky (2004). Verbalizing facial memory: Criterion effects in verbal overshadowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*: 739–755.

Cummins (1998). Reflection on reflective equilibrium. In De Paul & Ramsey (eds.), *Rethinking Intuition*. Rowan and Littlefield.

Davis-Floyd & Arvidson, eds. (1992). *Intuition: the inside story*. Routledge.

Dijksterhuis & van Olden (2005). On the benefits of thinking unconsciously: Unconscious thought can increase post-choice satisfaction. *Journal of Experimental Social Psychology, 42*: 627–631.

Dijksterhuis, Bos, Nordgren, & van Baaren (2006). On making the right choice: The dliberation-without-attention effect. Science, 17: 1005-1007.

Dougherty & Hunter (2003). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory and Cognition, 31*: 968–982.

Dunbar & Barrett, eds. (2007). *Oxford handbook of evolutionary psychology*. Oxford University Press.

Epstein (2007). Intuition From the Perspective of Cognitive-Experiential Self-Theory. In Plessner, Betsch, & Betsch (eds.), *Intuition in Judgment and Decision Making* (pp. 23-37). Psychology Press.

Evans & Over (1996). *Rationality and reasoning*. Psychology Press.

Fallshore & Schooler (1995). Verbal vulnerability of perceptual expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*: 1608–1623.

Fiore & Schooler (2002). How did you get here from there? Verbal overshadowing of spatial mental models. *Applied Cognitive Psychology, 16*: 897–909.

Gettier (1963). Is justified true belief knowledge? *Analysis, 23*: 121-123.

Gigerenzer (2007). *Gut Feelings: The Intelligence of the Unconscious*. Penguin.

Gilbert & Rappoport (1975). Categories of thought and variations in meaning: A demonstration experiment. *Journal of Phenomenological Psychology, 5*: 419–424.

Gilovich (1991). *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life*. Free Press.

Gladwell (2005). *Blink: The Power of Thinking Without Thinking*. Black Bay Books.

Goldman & Pust (1998). Philosophical theory and intuitional evidence. In De Paul & Ramsey (eds.), *Rethinking Intuition*. Rowan and Littlefield.

Gopnik & Schwitzgebel (1998). Whose concepts are they anyway? The role of philosophical intuition in empirical psychology. In De Paul & Ramsey (eds.), *Rethinking Intuition*. Rowan and Littlefield.

Hall (2000). Causation and the price of transitivity. *Journal of Philosophy, 97(4)*: 198-222.

Halbertsadt (2005). Featural shift in explanation-biased memory for emotional faces. *Journal of Personality and Social Psychology, 88*: 38–49.

Halberstadt and Levine (1999). Effects of reasons analysis on the accuracy of predicting basketball games. *Journal of Applied Social Psychology, 29*: 517–530.

Hastie & Dawes (2009). *Rational Choice in an Uncertain World, 2nd edition*. Sage.

Hogarth (2001). *Educating intuition*. University of Chicago Press.

Hogarth (2007). On the learning of intuition. In Plessner, Betsch, & Betsch (eds.), *Intuition in Judgment and Decision Making* (pp. 91-105). Psychology Press.

Kahneman & Frederick (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In Gilovich, Griffin, & Kahneman (eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49-81). Cambridge University Press.

Kahneman & Frederick (2005). A model of heuristic judgment. In Holyoak & Morrison (eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 267-294). Cambridge University Press.

Kahneman, Slovic, & Tversky (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Kardes (2002). *Consumer behavior and managerial decision making (2nd ed.)*. Prentice Hall.

Klimesch (1980). The effect of verbalization on memory performance for complex pictures. *Zeitschrift für Experimentelle und Angewandte Psychologie, 27*: 245–256.

Kripke (1972). *Naming and Necessity*. Harvard University Press.

Kruglanski & Freund (1983). The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology, 19*: 448–468.

Lewis (1973). Causation. *The Journal of Philosophy, 70(17)*: 556-567.

Mercier & Sperber (2009). Intuitive and reflective inferences. In Evans & Frankish (eds.), *In two minds: Dual processes and beyond*. Oxford University Press

Miller (2000). Without Intuitions. *Metaphilosophy, 31(3)*: 231-250.

Myers (2002). *Intuition: Its powers and perils*. Yale University Press.

Nisbett and Ross (1980). *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall.

Perfect, Hunt, & Harris (2002). Verbal overshadowing in voice recognition. *Applied Cognitive Psychology, 16*: 973–980.

Plessner, Betsch, & Betsch, eds. (2007). *Intuition in judgment and decision making*. Psychology Press.

Plessner & Czenna (2007). The benefits of intuition. In Plessner, Betsch, & Betsch (eds.), *Intuition in Judgment and Decision Making* (pp. 251-265). Psychology Press.

Pust (2000). Intuitions as Evidence. Garland.

Schooler & Engstler-Schooler (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22*: 36–71

Schroyens, Schaeken, & Handley (2003). In search of counter-examples: Deductive rationality in human reasoning. *Quarterly Journal of Experimental Psychology: A. Human Experimental Psychology, 56*: 1129–1145.

Searle (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*: 417-424.

Silverberg & Buchanan (2005). Verbal mediation and memory for novel figural designs: A dual interference study. *Brain and Cognition, 57*: 198–209

Sloman (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*: 3-22.

Sloman (2002). Two systems of reasoning. In Gilovich, Griffin, & Kahneman (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press.

Sosa (1998). Minimal Intuition. In De Paul & Ramsey (eds.), *Rethinking Intuition*. Rowman and Littlefield.

Stanovich (2004). *The robot's rebellion*. Chicago University Press.

Stanovich (2009). *How to Think Straight About Psychology, 9th edition*. Allyn & Bacon.

Stich (1988). Reflective equilibrium, analytic epistemology and the problem of cognitive diversity. *Synthese, 74(3)*: 391-413.

Suppes (1984). Conflicting intuitions about causality. *Midwest Studies in Philosophy, 9(1)*: 151-168.

Swain, Alexander, & Weinberg (2008). The instability of philosophical intuitions: Running hot and cold on truetemp. *Philosophy and Phenomenological Research, 76:* 138–155.

Talbot (2009). *How to Use Intuitions in Philosophy*. Dissertation. University of Southern California.

Talbot (forthcoming). The dilemma of calibrating intuitions.

Thomson (1971). A defense of abortion. *Philosophy and Public Affairs, 1(1)*: 47-66.

Weinberg, Nichols, & Stich (2001). Normativity and epistemic intuitions. Philosophical Topics, 29(1/2): 429-460.

Williams & Smart (1973). *Utilitarianism: For and Against*. Cambridge University Press.

Williamson (2004). Philosophical 'intuitions' and scepticism about judgement. *Dialectica, 58(1)*: 109-153.

Wilson (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Harvard University Press.

Windschitl & Krizan (2005). Contingent approaches to making likelihood judgments about polychotomous cases: The influence of task factors. *Journal of Behavioral Decision Making, 18*: 281–303.

Wisniewski (1998). The psychology of intuition. In De Paul & Ramsey (eds.), *Rethinking Intuition*. Rowan and Littlefield.

Concepts Don't Work That Way

Part of the sequence: Rationality and Philosophy

Philosophy in the Flesh, by George Lakoff and Mark Johnson, opens with a bang:

> The mind is inherently embodied. Thought is mostly unconscious. Abstract concepts are largely metaphorical.
>
> These are three major findings of cognitive science. More than two millennia of *a priori* philosophical speculation about these aspects of reason are over. Because of these discoveries, philosophy can never be the same again.
>
> When taken together and considered in detail, these three findings... are inconsistent with central parts of... analytic philosophy...
>
> This book asks: What would happen if we started with these empirical discoveries about the nature of mind and constructed philosophy anew?
>
> ...A serious appreciation of cognitive science requires us to rethink philosophy from the beginning, in a way that would put it more in touch with the reality of how we think.

So what *would* happen if we dropped all philosophical methods that were developed when we had a Cartesian view of the mind and of reason, and instead invented philosophy anew given what we now know about the physical processes that produce human reasoning?

> What emerges is a philosophy close to the bone. A philosophical perspective based on our empirical understanding of the embodiment of mind is a philosophy in the flesh, a philosophy that takes account of what we most basically *are* and *can be*.

Philosophy is a diseased discipline, but good philosophy can (and must) be done. I'd like to explore how one can do good philosophy, in part by taking cognitive science seriously.

Conceptual Analysis

Let me begin with a quick, easy example of how cognitive science can inform our philosophical methodology. The example below shouldn't surprise anyone who has read A Human's Guide to Words, but it does illustrate how misguided thousands of philosophical works can be due to an ignorance of cognitive science.

Consider what may be the central method of 20th century analytic philosophy: *conceptual analysis*. In its standard form, conceptual analysis assumes (Ramsey 1992) the "classical view" of concepts, that a "concept *C* has definitional structure in that it is composed of simpler concepts that express necessary and sufficient conditions for falling under *C*." For example, the concept bachelor has the constituents unmarried and man. Something falls under the concept bachelor if and only if it is an unmarried man.

Conceptual analysis, then, is the attempt to examine our intuitive concepts and arrive at definitions (in terms of necessary and sufficient conditions) that capture the meaning of those concepts. De Paul & Ramsey (1999) explain:

> Anyone familiar with Plato's dialogues knows how [conceptual analysis] is conducted. We see Socrates encounter someone who claims to have figured out the true essence of some abstract notion... the person puts forward a definition or analysis of the notion in the form of necessary and sufficient conditions that are thought to capture all and only instances of the concept in question. Socrates then refutes his interlocutor's definition of the concept by pointing out various counterexamples...
>
> For example, in Book I of the *Republic*, when Cephalus defines justice in a way that requires the returning of property and total honesty, Socrates responds by pointing out that it would be unjust to return weapons to a person who had gone mad or to tell the whole truth to such a person.... [The] proposed analysis is rejected because it fails to capture our intuitive judgments about the nature of justice.
>
> After a proposed analysis or definition is overturned by an intuitive counterexample, the idea is to revise or replace the analysis with one that is not subject to the counterexample. Counterexamples to the new analysis are sought, the analysis revised if any counterexamples are found, and so on...

The practice continues even today. Consider the conceptual analysis of *knowledge*. For centuries, knowledge was considered by most to be *justified true belief* (JTB). If Susan believed X but X wasn't true, then Susan couldn't be said to have knowledge of X. Likewise, if X was true but Susan didn't *believe* X, then she didn't have knowledge of X. And if Susan believed X and X was true but Susan had no justification for believing X, then she didn't really have "knowledge," she just had an accidentally true belief. But if Susan had justified true belief of X, then she *did* have knowledge of X.

And then Gettier (1963) offered some famous counterexamples to this analysis of knowledge. Here is a later counterexample, summarized by Zagzebski (1994):

> ...imagine that you are driving through a region in which, unknown to you, the inhabitants have erected three barn facades for each real barn in an effort to make themselves look more prosperous. Your eyesight is normal and reliable enough in ordinary circumstances to spot a barn from the road. But in this case the fake barns are indistinguishable from the real barns at such a distance. As you look at a real barn you form the belief 'That's a fine barn'. The belief is true and justified, but [intuitively, it isn't knowledge].

As in most counterexamples to the JTB analysis of knowledge, the counterexample to JTB arises due to "accidents" in the scenario:

> It is only an accident that visual faculties normally reliable in this sort of situation are not reliable in this particular situation; and it is another accident that you happened to be looking at a real barn and hit on the truth anyway... the [counter-example] arises because an accident of bad luck is cancelled out by an accident of good luck.

A cottage industry sprung up around these "Gettier problems," with philosophers proposing new sets of necessary and sufficient conditions for knowledge, and other

philosophers raising counter-examples to them. Weatherson (2003) described this circus as “the analysis of knowledge merry-go-round.”

My purpose here is not to examine Gettier problems in particular, but merely to show that the construction of conceptual analyses in terms of necessary and sufficient conditions is *mainstream philosophical practice*, and has been for a long time.

Now, let me explain how cognitive science undermines this mainstream philosophical practice.

Concepts in the Brain

The problem is that the brain doesn’t store concepts in terms of necessary and sufficient conditions, so philosophers have been using their intuitions to search for something that isn’t there. No wonder philosophers have, for over a century, failed to produce a single, successful, non-trivial conceptual analysis (Fodor 1981; Mills 2008).

How do psychologists know the brain doesn’t work this way? Murphy (2002, p. 16) writes:

> The groundbreaking work of Eleanor Rosch in the 1970s essentially killed the classical view, so that it is not now the theory of any actual [scientific] researcher...

But before we get to Rosch, let’s look at a different experiment:

> McCloskey and Glucksberg (1978)... found that when people were asked to make repeated category judgments such as ‘‘Is an olive a fruit?’’ or ‘‘Is a dog an animal?’’ there was a subset of items that individual subjects changed their minds about. That is, if you said that an olive was a fruit on one day, two weeks later you might give the opposite answer. Naturally, subjects did not do this for cases like ‘‘Is a dog an animal?’’ or ‘‘Is a rose an animal?’’ But they did change their minds on borderline cases, such as olive-fruit, and curtains-furniture. In fact, for items that were intermediate between clear members and clear nonmembers, McCloskey and Glucksberg’s subjects changed their mind 22% of the time. This may be compared to inconsistent decisions of under 3% for the best examples and clear nonmembers... Thus, the changes in subjects’ decisions do not reflect an overall inconsistency or lack of attention, but a bona fide uncertainty about the borderline members. In short, many concepts are not clear-cut. There are some items that... seem to be “kind of” members. (Mills 2002, p. 20)

Category-membership for concepts in the human brain is not a yes/no affair, as the “necessary and sufficient conditions” approach of the classical view assumes. Instead, category membership is *fuzzy*.

Another problem for the classical view is raised by *typicality effects*:

> Think of a fish, any fish. Did you think of something like a trout or a shark, or did you think of an eel or a flounder? Most people would admit to thinking of something like the first: a torpedo-shaped object with small fins, bilaterally symmetrical, which swims in the water by moving its tail from side to side. Eels are much longer, and they slither; flounders are also differently shaped, aren’t

> symmetrical, and move by waving their body in the vertical dimension. Although all of these things are technically fish, they do not all seem to be equally good examples of fish. The *typical* category members are the good examples — what you normally think of when you think of the category. The *atypical* objects are ones that are known to be members but that are unusual in some way... The classical view does not have any way of distinguishing typical and atypical category members. Since all the items in the category have met the definition's criteria, all are category members.
>
> ...The simplest way to demonstrate this phenomenon is simply to ask people to rate items on how typical they think each item is of a category. So, you could give people a list of fish and ask them to rate how typical each one is of the category fish. Rosch (1975) did this task for 10 categories and looked to see how much subjects agreed with one another. She discovered that the reliability of typicality ratings was an extremely high .97 (where 1.0 would be perfect agreement)... In short, people agree that a trout is a typical fish and an eel is an atypical one. (Mills 2002, p. 22)

So people agree that some items are more typical category members than others, but do these typicality effects manifest in normal cognition and behavior?

Yes, they do.

> Rips, Shoben, and Smith (1973) found that the ease with which people judged category membership depended on typicality. For example, people find it very easy to affirm that a robin is a bird but are much slower to affirm that a chicken (a less typical item) is a bird. This finding has also been found with visual stimuli: Identifying a picture of a chicken as a bird takes longer than identifying a pictured robin (Murphy and Brownell 1985; Smith, Balzano, and Walker 1978). The influence of typicality is not just in identifying items as category members — it also occurs with the production of items from a category. Battig and Montague (1969) performed a very large norming study in which subjects were given category names, like furniture or precious stone and had to produce examples of these categories. These data are still used today in choosing stimuli for experiments (though they are limited, as a number of common categories were not included). Mervis, Catlin and Rosch (1976) showed that the items that were most often produced in response to the category names were the ones rated as typical (by other subjects). In fact, the average correlation of typicality and production frequency across categories was .63, which is quite high given all the other variables that affect production.
>
> When people learn artificial categories, they tend to learn the typical items before the atypical ones (Rosch, Simpson, and Miller 1976). Furthermore, learning is faster if subjects are taught on mostly typical items than if they are taught on atypical items (Mervis and Pani 1980; Posner and Keele 1968). Thus, typicality is not just a feeling that people have about some items ("trout good; eels bad") — it is important to the initial learning of the category in a number of respects...
>
> Learning is not the end of the influence, however. Typical items are more useful for inferences about category members. For example, imagine that you heard that eagles had caught some disease. How likely do you think it would be to spread to other birds? Now suppose that it turned out to be larks or robins who caught the disease. Rips (1975) found that people were more likely to infer that other birds

> would catch the disease when a typical bird, like robins, had it than when an atypical one, like eagles, had it... (Murphy 2002, p. 23)

(If you want further evidence of typicality effects on cognition, see Murphy [2002] and Hampton [2008].)

The classical view of concepts, with its binary category membership, cannot explain typicality effects.

So the classical view of concepts must be rejected, along with any version of conceptual analysis that depends upon it. (If you doubt that many philosophers have done work dependent on the classical view of concepts, see here).

To be fair, quite a few philosophers have now given up on the classical view of concepts and the "necessary and sufficient conditions" approach to conceptual analysis. And of course there are other reasons that seeking definitions stipulated as necessary and sufficient conditions can be useful. But I wanted to begin with a clear and "settled" case of how cognitive science can undermine a particular philosophical practice and require that we ask and answer philosophical questions differently.

Philosophy by humans must respect the cognitive science of how humans reason.

Next post: Living Metaphorically

Previous post: When Intuitions Are Useful

References

Battig & Montague (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph, 80* (3, part 2).

Gettier (1963). Is justified true belief knowledge? *Analysis, 23*: 121-123.

De Paul & Ramsey (1999). Preface. In De Paul & Ramsey (eds.), *Rethinking Intuition*. Rowman & Littlefield.

Fodor (1981). The present status of the innateness controversy. In Fodor, *Representations: Philosophical Essays on the Foundations of Cognitive Science*. MIT Press.

Hampton (2008). Concepts in human adults. In Mareschal, Quinn, & Lea (eds.), *The Making of Human Concepts* (pp. 295-313). Oxford University Press.

McCloskey and Glucksberg (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition, 6*: 462–472.

Mervis, Catlin & Rosch (1976). Categorization of natural objects. *Annual Review of Psychology, 32*: 89–115.

Mervis & Pani (1980). Acquisition of basic object categories. *Cognitive Psychology, 12*: 496–522.

Mills (2008). Are analytic philosophers shallow and stupid? *The Journal of Philososphy, 105*: 301-319.

Murphy (2002). *The Big Book of Concepts*. MIT Press.

Murphy & Brownell (1985). Category differentiation in object recognition: Typicality constraints on the basic category advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*: 70–84.

Posner & Keele (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*: 353–363.

Rips (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior, 14*: 665–681.

Ramsey (1992). Prototypes and conceptual analysis. *Topoi 11*: 59-70.

Rips, Shoben, & Smith (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12*: 1–20.

Rosch (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*: 192–233.

Rosch, Simpson, & Miller (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance, 2*: 491–502.

Smith, Balzano, & Walker (1978). Nominal, perceptual, and semantic codes in picture categorization. In Cotton & Klatzky (eds.), *Semantic Factors in Cognition* (pp. 137–168). Erlbaum.

Weatherson (2003). What good are counterexamples? *Philosophical Studies, 115*: 1-31.

Living Metaphorically

Part of the sequence: Rationality and Philosophy

In my last post, I showed that the brain does not encode concepts in terms of necessary and sufficient conditions. So, any philosophical practice which assumes this — as much of 20th century conceptual analysis seems to do — is misguided.

Next, I want to show that human abstract thought is pervaded by metaphor, and that this has implications for how we think about the nature of philosophical questions and philosophical answers. As Lakoff & Johnson (1999) write:

> If we are going to ask philosophical questions, we have to remember that we are human... The fact that abstract thought is mostly metaphorical means that answers to philosophical questions have always been, and always will be, mostly metaphorical. In itself, that is neither good nor bad. It is simply a fact about the capacities of the human mind. But it has major consequences for every aspect of philosophy. Metaphorical thought is the principal tool that makes philosophical insight possible, and that constrains the forms that philosophy can take.

To understand how fundamental metaphor is to our thinking, we must remember that human cognition is *embodied*:

> We have inherited from the Western philosophical tradition a theory of faculty psychology, in which we have a "faculty" of reason that is separate from and independent of what we do with our bodies. In particular, reason is seen as independent of perception and bodily movement...
>
> The evidence from cognitive science shows that classical faculty psychology is wrong. There is no such fully autonomous faculty of reason separate from and independent of bodily capacities such as perception and movement. The evidence supports, instead, an evolutionary view, in which reason uses and grows out of such bodily capacities.

Consider, for example, the fact that as *neural* beings we *must categorize things*:

> We are neural beings. Our brains each have 100 billion neurons and 100 trillion synaptic connections. It is common in the brain for information to be passed from one dense ensemble of neurons to another via a relatively sparse set of connections. Whenever this happens, the pattern of activation distributed over the first set of neurons is too great to be represented in a one-to-one manner in the sparse set of connections. Therefore, the sparse set of connections necessarily groups together certain input patterns in mapping them across to the output ensemble. Whenever a neural ensemble provides the same output with different inputs, there is neural categorization.
>
> To take a concrete example, each human eye has 100 million light-sensing cells, but only about 1 million fibers leading to the brain. Each incoming image must therefore be reduced in complexity by a factor of 100. That is, information in each fiber constitutes a "categorization" of the information from about 100 cells.

Moreover, almost all our categorizations are determined by the unconscious associative mind — outside our control and even our awareness — as we interact with

the world. As Lakoff & Johnson note, "Even when we think we are deliberately forming new categories, our unconscious categories enter into our choice of possible conscious categories."

And because our categories are shaped not by a transcendent, universal faculty of reason but by the components of our sensorimotor system that process our interaction with the world, our concepts and categories end up being largely sensorimotor concepts and categories.

Here are some examples of metaphorical thought shaped by the sensorimotor system:

Important Is Big
Example: "Tomorrow is a *big* day."
Mapping: From importance to size.
Experience: As a child, finding that big things (e.g. parents) are important and can exert major forces on you and dominate your visual experience.

Intimacy Is Closeness
Example: "We've been *close* for years, but we're beginning to *drift apart*."
Mapping: From intimacy to physical proximity.
Experience: Being physically close to people you are intimate with.

Difficulties Are Burdens
Example: "She's *weighed down* by her responsibilities."
Mapping: From difficulty to muscular exertion.
Experience: The discomfort or disabling effect of lifting or carrying heavy objects.

More Is Up
Example: "Prices are high."
Mapping: From quantity to vertical orientation.
Experience: Observing the rise and fall of levels of piles and fluids as more is added or subtracted.

Categories Are Containers
Example: "Are tomatoes *in* the fruit or vegetable category?"
Mapping: From kinds to spatial location.
Experience: Observing that things that go together tend to be in the same bounded region.

Linear Scales Are Paths
Example: "John's intelligence *goes way beyond* Bill's."
Mapping: From degree to motion in space.
Experience: Observing the amount of progress made by an object.

Organization Is Physical Structure
Example: "How do the pieces of this theory fit together?"
Mapping: From abstract relationships to experience with physical objects.
Experience: Interacting with complex objects and attending to their structure.

States Are Locations
Example: "I'm *close* to being *in* a depression and the next thing that goes wrong will *send me over the edge*.
Mapping: From a subjective state to being in a bounded region of space.
Experience: Experiencing a certain state as correlated with a certain location (e.g. being cool under a tree, feeling secure in a bed).

Purposes Are Destinations
Example: "He'll ultimately be successful, but he isn't *there* yet."
Mapping: From achieving a purpose to reaching a destination in space.
Experience: Reaching destinations throughout everyday life and thereby achieving purposes (e.g. if you want food, you have to go to the fridge).

Actions Are Motions
Example: "I'm *moving* right along on the project."
Mapping: From action to moving your body through space.
Experience: The common action of moving yourself through space, especially in the early years of life when that is to some degree the *only* kind of action you can take.

Understanding Is Grasping
Example: "I've never been able to grasp transfinite numbers."
Mapping: From comprehension to object manipulation.
Experience: Getting information about an object by grasping and manipulating it.

As a neural being interacting with the world, you can't help but build up such "primary" metaphors:

> If you are a normal human being, you inevitably acquire an enormous range of primary metaphors just by going about the world constantly moving and perceiving. Whenever a domain of subjective experience or judgment is coactivated regularly with a sensorimotor domain, permanent neural connections are established via synaptic weight changes. Those connections, which you have unconsciously formed by the thousands, provide inferential structure and qualitative experience activated in the sensorimotor system to the subjective domains they are associated with. Our enormous metaphoric conceptual system is thus built up by a process of neural selection. Certain neural connections between the activated source- and target-domain networks are randomly established at first and then have their synaptic weights increased through their recurrent firing. The more times those connections are activated, the more the weights are increased, until permanent connections are forged.

Primary metaphors are combined to build complex metaphors. For example, Actions Are Motions and Purposes Are Destinations are often combined to form a new metaphor:

A Purposeful Life is a Journey
Example: "She seems lost, without direction. She's fallen off track. She needs to find her purpose and get moving again."

Can we think *without* metaphor, then? Yes. Our concepts of so-called "basic level" objects (that we interact with in everyday experience) are often literal, as are sensorimotor concepts. Our concepts of "tree" (the thing that grows in dirt), "grasp" (holding an object), and "in" (in the spatial sense) are all literal. But when it comes to abstract reasoning or subjective judgment, we tend to think in metaphor. We can't help it.

Implications for philosophical method

What happens when we fail to realize that our thinking is metaphorical? Let's consider a famous example: Zeno's paradox of the arrow.

Zeno described time as a sequence of points along a timeline. Now, consider an arrow in flight. At any point on the timeline, the arrow is at some particular fixed location. At a later point on the timeline, the arrow is at a different location. But since the arrow is located at a single fixed place at every point in time, then where is the motion?

> Suppose, Zeno argues, that time really is a sequence of points constituting a time line. Consider the flight of an arrow. At any point in time, the arrow is at some fixed location. At a later point, it is at another fixed location. The flight of the arrow would be like the sequence of still frames that make up a movie. Since the arrow is located at a single fixed place at every time, where, asks Zeno, is the motion?

The puzzle arises when you take the metaphor of time as discrete *points* along the *space* of a time*line* as being literal:

> Zeno's brilliance was to concoct an example that forced a contradiction upon us: [a contradiction between] literal motion and motion metaphorically conceptualized as a sequence of fixed locations at fixed points in time.

Moral concepts as metaphors

For a more detailed illustration of the philosophical implications of metaphorical thought, let's examine the metaphors that ground our moral concepts:

> Morality is fundamentally seen as the enhancing of well-being, especially of others. For this reason, ...basic folk theories of what constitutes fundamental well-being form the grounding for systems of moral metaphors around the world. For example, since most people find it better to have enough wealth to live comfortably than to be impoverished, we are not surprised to find that well-being is conceptualized as wealth...
>
> We all conceptualize well-being as wealth. We understand an increase in well-being as a *gain* and a decrease of well-being as a *loss* or a *cost*. We speak of *profiting* from an experience, of having a *rich* life, of *investing* in happiness, and of *wasting* our lives... If you do something good for me, then I *owe* you something, I am in your *debt*. If I do something equally good for you, then I have *repaid* you and we are *even*. The books are balanced.

Well-Being Is Wealth is not the only metaphor behind our moral thinking. Here are a few others:

Being Moral Is Being Upright; Being Immoral Is Being Low; Evil Is a Force
Example: "He's an *upstanding* citizen. She's on the *up and up*. She's as *upright* as they come. That was a *low* thing to do. He's *under*handed. I would never *stoop* to such a thing. She *fell* from grace. She succumbed to the *floods* of emotion and the *fires* of passion. She didn't have enough moral *backbone* to *stand up to* evil."

How does the metaphorical nature of our moral concepts constrain moral philosophy? Let us contrast a traditional view of moral concepts with the view of moral concepts emerging from cognitive science:

> The traditional view of moral concepts and reasoning says the following: Human reasoning is compartmentalized, depending on what aspects of experience it is directed to. There are scientific judgments, technical judgments, prudential

> judgments, aesthetic judgments, and ethical judgments. For each type of judgment, there is a corresponding distinct type of literal concept. Therefore, there exists a unique set of concepts that pertain only to ethical issues. These ethical concepts are literal and must be understood only "in themselves" or by virtue of their relations to other purely ethical concepts. Moral rules and principles are made up from purely ethical concepts like these, concepts such as *good*, *right*, *duty*, *justice*, and *freedom*. We use our reason to apply these ethical concepts and rules to concrete, actual situations in order to decide how we ought to act in a given case.
>
> … [But] there is no set of pure moral concepts that could be understood "in themselves" or "on their own terms." Instead, we understand morality via mappings of structures from other aspects and domains of our experience: wealth, balance, order, boundaries, light/dark, beauty, strength, and so on. If our moral concepts are metaphorical, then their structure and logic come primarily from the source domains that ground the metaphors. We are thus understanding morality by means of structures drawn from a broad range of dimensions of human experience, including domains that are never considered by the traditional view to be "ethical" domains. In other words, the constraints on our moral reasoning are mostly imported from other conceptual domains and aspects of experience...

An explosion of productivity in moral psychology since Lakoff & Johnson's book was published has confirmed these claims. The convergence of evidence suggests that multiple competing systems contribute to our moral reasoning, and they engage many processes not unique to moral reasoning.

Once again, knowledge of cognitive science constrains philosophy:

> This view of moral concepts as metaphoric profoundly calls into question the idea of a "pure" moral reason... [Moreover,] we do not have a monolithic, homogeneous, consistent set of moral concepts. For example, we have different, inconsistent, metaphorical structurings of our notion of well-being, and these are employed in moral reasoning.

Next post: Intuitions Aren't Shared That Way

Previous post: Concepts Don't Work That Way

Intuitions Aren't Shared That Way

Part of the sequence: Rationality and Philosophy

Consider these two versions of the famous trolley problem:

> **Stranger**: A train, its brakes failed, is rushing toward five people. The only way to save the five people is to throw the switch sitting next to you, which will turn the train onto a side track, thereby preventing it from killing the five people. However, there is a stranger standing on the side track with his back turned, and if you proceed to thrown the switch, the five people will be saved, but the person on the side track will be killed.
>
> **Child**: A train, its brakes failed, is rushing toward five people. The only way to save the five people is to throw the switch sitting next to you, which will turn the train onto a side track, thereby preventing it from killing the five people. However, there is a 12-year-old boy standing on the side track with his back turned, and if you proceed to throw the switch, the five people will be saved, but the boy on the side track will be killed.

Here it is: a standard-form philosophical thought experiment. In standard analytic philosophy, the next step is to engage in *conceptual analysis* — a process in which we use our intuitions as evidence for one theory over another. For example, if your intuitions say that it is "morally right" to throw the switch in both cases above, then these intuitions may be counted as evidence for consequentialism, for moral realism, for agent neutrality, and so on.

Alexander (2012) explains:

> Philosophical intuitions play an important role in contemporary philosophy. Philosophical intuitions provide data to be explained by our philosophical theories [and] evidence that may be adduced in arguments for their truth... In this way, the role... of intuitional evidence in philosophy is similar to the role... of perceptual evidence in science...
>
> Is knowledge simply justified true belief? Is a belief justified just in case it is caused by a reliable cognitive mechanism? Does a name refer to whatever object uniquely or best satisfies the description associated with it? Is a person morally responsible for an action only if she could have acted otherwise? Is an action morally right just in case it provides the greatest benefit for the greatest number of people all else being equal? When confronted with these kinds of questions, philosophers often appeal to philosophical intuitions about real or imagined cases...
>
> ...there is widespread agreement about the role that [intuitions] play in contemporary philosophical practice... We advance philosophical theories on the basis of their ability to explain our philosophical intuitions, and appeal to them as evidence that those theories are true...

In particular, notice that philosophers do not appeal to their intuitions as merely an exercise in *autobiography*. Philosophers are not merely trying to map the contours of *their own* idiosyncratic concepts. That could be interesting, but it wouldn't be worth decades of publicly-funded philosophical research. Instead, philosophers appeal to

their intuitions as evidence for what is *true in general* about a concept, or true about the world.

In this sense,

> We [philosophers] tend to believe that our philosophical intuitions are more or less universally shared... We... appeal to philosophical intuitions, when we do, because we anticipate that others share our intuitive judgments.

But anyone with more than a passing familiarity with cognitive science might have bet in advance that this basic underlying assumption of a core philosophical method is... *incorrect*.

For one thing, philosophical intuitions show *gender diversity*. Consider again the *Stranger* and *Child* versions of the Trolley problem. It turns out that men are less likely than women to think it is morally acceptable to throw the switch in the *Stranger* case, while women are less likely than men to think it is morally acceptable to throw the switch in the *Child* case (Zamzow & Nichols 2009).

Or, consider a thought experiment meant to illuminate the much-discussed concept of *knowledge*:

> Peter is in his locked apartment and is reading. He decides to have a shower. He puts his book down on the coffee table. Then he takes off his watch, and also puts it on the coffee table. Then he goes into the bathroom. As Peter's shower begins, a burglar silently breaks into Peter's apartment. The burglar takes Peter's watch, puts a cheap plastic watch in its place, and then leaves. Peter has only been in the shower for two minutes, and he did not hear anything.

When presented with this vignette, only 41% of men say that Peter "knows" there is a watch on the table, while 71% of women say that Peter "knows" there is a watch on the table (Starman & Friedman 2012). According to Buckwalter & Stich (2010), Starmans & Friedman ran another study using a slightly different vignette with a female protagonist, and that time only 36% of men said the protagonist "knows," while 75% of women said she "knows."

The story remains the same for intuitions about free will. In another study reported in Buckwalter & Stich (2010), Geoffrey Holtman presented subjects with this vignette:

> Suppose scientists figure out the exact state of the universe during the Big Bang, and figure out all the laws of physics as well. They put this information into a computer, and the computer perfectly predicts everything that has ever happened. In other words, they prove that everything that happens has to happen exactly that way because of the laws of physics and everything that's come before. In this case, is a person free to choose whether or not to murder someone?

In this study, only 35% of men, but 63% of women, said a person in this world could be free to choose whether or not to murder someone.

Intuitions show not only gender diversity but also *cultural diversity*. Consider another thought experiment about *knowledge* (you can punch me in the face, later):

> Bob has a friend Jill, who has driven a Buick for many years. Bob therefore thinks that Jill drives an American car. He is not aware, however, that her Buick has recently been stolen, and he is also not aware that Jill has replaced it with a

> Pontiac, which is a different kind of American car. Does Bob really know that Jill drives an American car, or does he only believe it?

Only 26% of Westerners say that Bob "knows" that Jill drives an American car, while 56% of East Asian subjects, and 61% of South Asian subjects, say that Bob "knows."

Now, consider a thought experiment meant to elicit semantic intuitions:

> Suppose that John has learned in college that Gödel is the man who proved... the incompleteness of arithmetic. John is quite good at mathematics and he can give an accurate statement of the incompleteness theorem, which he attributes to Gödel as the discoverer. But this is the only thing that he has heard about Gödel. Now suppose that Gödel was not the author of this theorem. A man called "Schmidt"... actually did the work in question. His friend Gödel somehow got a hold of the manuscript and claimed credit for the work, which was thereafter attributed to Gödel... Most people who have heard the name "Gödel" are like John; the claim that Gödel discovered the incompleteness theorem is the only thing that they have ever heard about Gödel.

When presented with this vignette, East Asians are more likely to take the "descriptivist" view of reference, believing that John "is referring to" Schmidt — while Westerners are more likely to take the "causal-historical" view, believing that John "is referring to" Gödel (Machery et al. 2004).

Previously, I asked:

> What would happen if we dropped all philosophical methods that were developed when we had a Cartesian view of the mind and of reason, and instead invented philosophy anew given what we now know about the physical processes that produce human reasoning?

For one thing, we would never assume that people of all kinds would share our intuitions.

Next post: Philosophy Needs to Trust Your Rationality Even Though It Shouldn't

Previous post: Living Metaphorically

Philosophy Needs to Trust Your Rationality Even Though It Shouldn't

Part of the sequence: Rationality and Philosophy

> Philosophy is notable for the extent to which disagreements with respect to even those most basic questions persist among its most able practitioners, despite the fact that the arguments thought relevant to the disputed questions are typically well-known to all parties to the dispute.
>
> Thomas Kelly

> The goal of philosophy is to uncover certain truths... [But] philosophy continually leads experts with the highest degree of epistemic virtue, doing the very best they can, to accept a wide array of incompatible doctrines. Therefore, philosophy is an unreliable instrument for finding truth. A person who enters the field is highly unlikely to arrive at true answers to philosophical questions.
>
> Jason Brennan

After millennia of debate, philosophers remain heavily divided on many core issues. According to the largest-ever survey of philosophers, they're split 25-24-18 on deontology / consequentialism / virtue ethics, 35-27 on empiricism vs. rationalism, and 57-27 on physicalism vs. non-physicalism.

Sometimes, they are even divided on *psychological* questions that psychologists have *already answered*: Philosophers are split evenly on the question of whether it's possible to make a moral judgment without being motivated to abide by that judgment, even though we already know that this *is* possible for some people with damage to their brain's reward system, for example many Parkinson's patients, and patients with damage to the ventromedial frontal cortex (Schroeder et al. 2012).[1]

Why are physicists, biologists, and psychologists more prone to reach consensus than philosophers?[2] One standard story is that "the method of science is to amass such an enormous mountain of evidence that... scientists cannot ignore it." Hence, *religionists* might still argue that Earth is flat or that evolutionary theory and the Big Bang theory are "lies from the pit of hell," and *philosophers* might still be divided about whether somebody can make a moral judgment they aren't themselves motivated by, but *scientists* have reached consensus about such things.

In its dependence on masses of evidence and definitive experiments, science doesn't trust your rationality:

> Science is built around the assumption that you're too stupid and self-deceiving to just use [probability theory]. After all, if it was that simple, we wouldn't need a social process of science... [Standard scientific method] doesn't trust your rationality, and it doesn't rely on your ability to use probability theory as the arbiter of truth. It wants you to set up a definitive experiment.

Sometimes, you can answer philosophical questions with mountains of evidence, as with the example of moral motivation given above. But or many philosophical problems, overwhelming evidence simply isn't *available*. Or maybe you can't afford to wait a decade for definitive experiments to be done. Thus, "if you would rather not waste ten years trying to prove the wrong theory," or if you'd like to get the right answer without overwhelming evidence, "you'll need to [tackle] the vastly more difficult problem: listening to evidence that doesn't shout in your ear."

This is why philosophers need rationality training even more desperately than scientists do. Philosophy asks you to get the right answer *without* evidence that shouts in your ear. The less evidence you have, or the harder it is to interpret, the more rationality you need to get the right answer. (As likelihood ratios get smaller, your priors need to be better and your updates more accurate.)

Because it tackles so many questions that *can't* be answered by masses of evidence or definitive experiments, philosophy needs to trust your rationality even though it shouldn't: we generally *are* as "stupid and self-deceiving" as science assumes we are. We're "predictably irrational" and all that.

But hey! Maybe philosophers are prepared for this. Since philosophy is so much more demanding of one's rationality, perhaps the field has built top-notch rationality training into the standard philosophy curriculum?

Alas, it doesn't seem so. I don't see much Kahneman & Tversky in philosophy syllabi — just light-weight "critical thinking" classes and lists of informal fallacies. But even classes in human bias might not improve things much due to the sophistication effect: someone with a sophisticated knowledge of fallacies and biases might just have more ammunition with which to attack views they don't like. So what's really needed is regular habits training for genuine curiosity, motivated cognition mitigation, and so on.

(Imagine a world in which Frank Jackson's famous reversal on the knowledge argument *wasn't news* — because established philosophers changed their minds all the time. Imagine a world in which philosophers were fine-tuned enough to reach consensus on 10 bits of evidence rather than 1,000.)

We might also ask: How well do philosophers perform on standard tests of rationality, for example Frederick (2005)'s CRT? Livengood et al. (2010) found, via an internet survey, that subjects with *graduate-level* philosophy training had a mean CRT score of 1.32. (The best possible score is 3.)

A score of 1.32 isn't radically different from the mean CRT scores found for psychology undergraduates (1.5), financial planners (1.76), Florida Circuit Court judges (1.23), Princeton Undergraduates (1.63), and people who happened to be sitting along the Charles River during a July 4th fireworks display (1.53). It is also noticeably *lower* than the mean CRT scores found for MIT students (2.18) and for attendees to a LessWrong.com meetup group (2.69).

Moreover, several studies show that philosophers are just as prone to particular biases as laypeople (Schulz et al. 2011; Tobia et al. 2012), for example order effects in moral judgment (Schwitzgebel & Cushman 2012).

People are typically excited about the Center for Applied Rationality because it teaches thinking skills that can improve one's happiness and effectiveness. That excites me, too. But I hope that in the long run CFAR will also help produce *better*

philosophers, because it looks to me like we need top-notch philosophical work to secure a desirable future for humanity.[3]

Next post: Train Philosophers with Pearl and Kahneman, not Plato and Kant

Previous post: Intuitions Aren't Shared That Way

Notes

[1] Clearly, many philosophers have advanced versions of motivational internalism that are directly contradicted by these results from psychology. However, we don't know exactly which version of motivational internalism is defended by each survey participant who said they "accept" or "lean toward" motivational internalism. Perhaps many of them defend weakened versions of motivational internalism, such as those discussed in section 3.1 of May (forthcoming).

[2] Mathematicians reach even stronger consensus than physicists, but they don't appeal to what is usually thought of as "mountains of evidence." What's going on, there? Mathematicians and philosophers almost always agree about whether a proof or an argument is valid, given a particular formal system. The difference is that a mathematician's premises consist in axioms and in theorems already strongly proven, whereas a philosopher's premises consist in substantive claims about the world for which the evidence given is often very weak (e.g. that philosopher's intuitions).

[3] Bostrom (2000); Yudkowsky (2008); Muehlhauser (2011).

Train Philosophers with Pearl and Kahneman, not Plato and Kant

Part of the sequence: Rationality and Philosophy

> Hitherto the people attracted to philosophy have been mostly those who loved the big generalizations, which were all wrong, so that few people with exact minds have taken up the subject.
>
> Bertrand Russell

I've complained before that philosophy is a diseased discipline which spends far too much of its time debating definitions, ignoring relevant scientific results, and endlessly re-interpreting old dead guys who didn't know the slightest bit of 20th century science. Is that *still* the case?

You bet. There's *some* good philosophy out there, but much of it is bad enough to make CMU philosopher Clark Glymour suggest that on tight university budgets, philosophy departments could be defunded unless their work is useful to (cited by) scientists and engineers — just as his own work on causal Bayes nets is now widely used in artificial intelligence and other fields.

How did philosophy get this way? Russell's hypothesis is not too shabby. Check the syllabi of the undergraduate "intro to philosophy" classes at the world's top 5 U.S. philosophy departments — NYU, Rutgers, Princeton, Michigan Ann Arbor, and Harvard — and you'll find that they spend a *lot* of time with (1) old dead guys who were wrong about almost everything because they knew nothing of modern logic, probability theory, or science, and with (2) 20th century philosophers who were way too enamored with cogsci-ignorant armchair philosophy. (I say more about the reasons for philosophy's degenerate state here.)

As the CEO of a philosophy/math/compsci research institute, I think many philosophical problems are important. But the field of philosophy doesn't seem to be very good at answering them. What can we do?

Why, come up with better philosophical methods, of course!

Scientific methods have improved over time, and so can philosophical methods. Here is the *first* of my recommendations...

More Pearl and Kahneman, less Plato and Kant

Philosophical training should begin with the latest and greatest formal methods ("Pearl" for the probabilistic graphical models made famous in Pearl 1988), and the latest and greatest science ("Kahneman" for the science of human reasoning reviewed in Kahneman 2011). Beginning with Plato and Kant (and company), as most universities do today, both (1) filters for inexact thinkers, as Russell suggested, and

(2) teaches people to have too much respect for failed philosophical methods that are out of touch with 20th century breakthroughs in math and science.

So, I recommend we teach young philosophy students:

more Bayesian rationality, heuristics and biases, & debiasing,	*less*	informal "critical thinking skills";
more mathematical logic & theory of computation,	*less*	term logic;
more probability theory & Bayesian scientific method,	*less*	pre-1980 philosophy of science;
more psychology of concepts & machine learning,	*less*	conceptual analysis;
more formal epistemology & computational epistemology,	*less*	pre-1980 epistemology;
more physics & cosmology,	*less*	pre-1980 metaphysics;
more psychology of choice,	*less*	philosophy of free will;
more moral psychology, decision theory, and game theory,	*less*	intuitionist moral philosophy;
more cognitive psychology & cognitive neuroscience,	*less*	pre-1980 philosophy of mind;
more linguistics & psycholinguistics,	*less*	pre-1980 philosophy of language;
more neuroaesthetics,	*less*	aesthetics;
more causal models & psychology of causal perception,	*less*	pre-1980 theories of causation.

(In other words: train philosophy students like they do at CMU, but even "more so.")

So, my own "intro to philosophy" mega-course might be guided by the following core readings:

1. Stanovich, *Rationality and the Reflective Mind* (2010)
2. Hinman, *Fundamentals of Mathematical Logic* (2005)
3. Russell & Norvig, *Artificial Intelligence: A Modern Approach* (3rd edition, 2009) — contains chapters which briefly introduce probability theory, probabilistic graphical models, computational decision theory and game theory, knowledge representation, machine learning, computational epistemology, and other useful subjects
4. Sipser, *Introduction to the Theory of Computation* (3rd edition, 2012) — relevant to lots of philosophical problems, as discussed in Aaronson (2011)
5. Howson & Urbach, *Scientific Reasoning: The Bayesian Approach* (3rd edition, 2005)
6. Holyoak & Morrison (eds.), *The Oxford Handbook of Thinking and Reasoning* (2012) — contains chapters which briefly introduce the psychology of knowledge representation, concepts, categories, causal learning, explanation, argument, decision making, judgment heuristics, moral judgment, behavioral game theory, problem solving, creativity, and other useful subjects
7. Dolan & Sharot (eds.), *Neuroscience of Preference and Choice* (2011)
8. Krane, *Modern Physics* (3rd edition, 2012) — includes a brief introduction to cosmology

(There are many prerequisites to these, of course. I think philosophy should be a Highly Advanced subject of study that requires lots of prior training in maths and the sciences, like string theory but hopefully more productive.)

Once students are equipped with some of the latest math and science, *then* let them tackle The Big Questions. I bet they'd get farther than those raised on Plato and Kant instead.

You might also let them read 20th century analytic philosophy at that point — hopefully their training will have inoculated them from picking up bad thinking habits.

Previous post: Philosophy Needs to Trust Your Rationality Even Though It Shouldn't