Jerks, Zombie Robots, and Other Philosophical Misadventures

Eric Schwitzgebel

in memory of my father,

psychologist, inventor, parent, philosopher, and giver of strange objects

Preface

I enjoy writing short philosophical reflections for broad audiences. Evidently, I enjoy this a lot: Since 2006, I've written over a thousand such pieces, mostly published on my blog *The Splintered Mind,* but also in the *Los Angeles Times, Aeon Magazine,* and elsewhere. This book contains fifty-eight of my favorites, revised and updated.

The topics range widely, as I've tried to capture in the title of the book. I discuss moral psychology ("jerks"), speculative philosophy of consciousness ("zombie robots"), the risks of controlling your emotions technologically, the ethics of the game of dreidel, multiverse theory, the apparent foolishness of Immanuel Kant, and much else. There is no unifying topic.

Maybe, however, there is a unifying theme. The human intellect has a ragged edge, where it begins to turn against itself, casting doubt on itself or finding itself lost among seemingly improbable conclusions. We can reach this ragged edge quickly. Sometimes, all it takes to remind us of our limits is an eight-hundred-word blog post. Playing at this ragged edge, where I no longer know quite what to think or how to think about it, is my idea of fun.

Given the human propensity for rationalization and self-deception, when I disapprove of others, how do I know that I'm not the one who is being a jerk? Given that all our intuitive, philosophical, and scientific knowledge of the mind has been built on a narrow range of cases, how much confidence can we have in our conclusions about strange new possibilities that are likely to open up in the near future of Artificial Intelligence? Speculative cosmology at once poses the (literally) biggest questions that we can ask about the universe while opening up possibilities that undermine our confidence in our ability to answer those same questions. The history of philosophy is humbling when we see how badly wrong previous thinkers have been, despite their intellectual skills and confidence.

Not all of my posts fit this theme. It's also fun to use the once-forbidden word "fuck" over and over again in a chapter about profanity. And I wanted to share some reminiscences about how my father saw the world – especially since in some ways I prefer his optimistic and proactive vision to my own less hopeful skepticism. Other of my blog posts I just liked or wanted to share for other reasons. A few are short fictions.

It would be an unusual reader who liked every chapter. I hope you'll skip anything you find boring. The chapters are all free-standing. Please don't just start reading on page one and then try to slog along through everything sequentially out of some misplaced sense of duty! Trust your sense of fun (Chapter 47). Read only the chapters that appeal to you, in any order you like.

Riverside, California, Earth (I hope)

October 25, 2018

Part One: Jerks and Excuses

Part Two: Cute AI and Zombie Robots

Part Three: Regrets and Birthday Cake

Part One: Jerks and Excuses

1. A Theory of Jerks

Picture the world through the eyes of the jerk. The line of people in the post office is a mass of unimportant fools; it's a felt injustice that you must wait while they bumble with their requests. The flight attendant is not a potentially interesting person with her own cares and struggles but instead the most available face of a corporation that stupidly insists you stow your laptop. Custodians and secretaries are lazy complainers who rightly get the scut work. The person who disagrees with you at the staff meeting is an idiot to be shot down. Entering a subway is an exercise in nudging past the dumb schmoes.

We need a theory of jerks. We need such a theory because, first, it can help us achieve a calm, clinical understanding when confronting such a creature in the wild. Imagine the nature-documentary voice-over: "Here we see the jerk in his natural environment. Notice how he subtly adjusts his dominance display to the Italian-restaurant situation…." And second – well, I don't want to say what the second reason is quite yet.

As it happens, I do have such a theory. But before we get into it, I should clarify some terminology. The word "jerk" can refer to two different types of person. The older use of "jerk" designates a chump or ignorant fool, though not a morally odious one. When Weird Al Yankovic sang, in 2006, "I sued Fruit of the Loom 'cause when I wear their tightie-whities on my head I look like a jerk" or when, in 1959, Willard Temple wrote in the *Los Angeles Times* "He could have married the campus queen…. Instead the poor jerk fell for a snub-nosed, skinny little broad", it's clear it's the chump they have in mind.[1]

The jerk-as-fool usage seems to have begun among traveling performers as a derisive reference to the unsophisticated people of a "jerkwater town", that is, a town not rating a full-scale train station, requiring the boilerman to pull on a chain to water his engine. The term expresses the traveling troupe's disdain.[2] Over time, however, "jerk" shifted from being primarily a class-based insult to its second, now dominant, sense as a moral condemnation.

Such linguistic drift from class-based contempt to moral deprecation is a common pattern across languages, as observed by Friedrich Nietzsche in *On the Genealogy of Morality.*[3] (In English, consider "rude", "villain", and "ignoble".) It is the immoral jerk who concerns me here.

Why, you might be wondering, should a philosopher make it his business to analyze colloquial terms of abuse? Doesn't the Urban Dictionary cover that kind of thing quite adequately? Shouldn't I confine myself to truth, or beauty, or knowledge, or why there is something rather than nothing? I am, in fact, interested in all those topics. And yet I see a folk wisdom in the term "jerk" that points toward something morally important. I want to extract that morally important thing, isolating the core phenomenon implicit in our usage. Precedents for this type of philosophical work include Harry Frankfurt's essay *On Bullshit* and, closer to my target, Aaron James's book *Assholes*.[4] Our taste in vulgarity reveals our values.

I submit that the unifying core, the essence of jerkitude in the moral sense, is this: *The jerk culpably fails to appreciate the perspectives of others around him, treating them as tools to be manipulated or fools to be dealt with rather than as moral and epistemic peers.* This failure has both an intellectual dimension and an emotional dimension, and it has these two dimensions on both sides of the relationship. The jerk himself is both intellectually and emotionally defective, and what he defectively fails to appreciate is both the intellectual and emotional perspectives of the people around him. He can't appreciate how he might be wrong and others right about some matter of fact; and what other people want or value doesn't register as of interest to him, except derivatively upon his own interests. The bumpkin ignorance captured in the earlier use of "jerk" has become a type of moral ignorance.

Some related traits are already well-known in psychology and philosophy – the "dark triad" of Machiavellianism, narcissism, and psychopathy; low "Agreeableness" on the Big Five personality test; and Aaron James's conception of the asshole, already mentioned. But my conception of the jerk differs from all of these. The asshole, James says, is someone who allows himself to enjoy special advantages out of an entrenched sense of entitlement.[5] That is one important dimension of jerkitude, but not the whole story. The callous psychopath, though cousin to the jerk, has an impulsivity and love of risk-taking that needn't belong to the jerk's character.[6] Neither does the jerk have to be as thoroughly self-involved as the narcissist or as self-consciously cynical as the Machiavellian, though narcissism and Machiavellianism are common jerkish attributes.[7] People low in Big-5 Agreeableness tend to be unhelpful, mistrusting, and difficult to get along with – again, features related to jerkitude, and perhaps even partly constitutive of it, but not exactly jerkitude as I've defined it. Also, my definition of jerkitude has a conceptual unity that is, I think, theoretically appealing in the abstract and fruitful in helping to explain some of the peculiar features of this type of animal, as we will see.

The opposite of the jerk is the *sweetheart*. The sweetheart sees others around him, even strangers, as individually distinctive people with valuable perspectives, whose desires and opinions, interests and goals, are worthy of attention and respect. The sweetheart yields his place in line to the hurried shopper, stops to help the person who has dropped her papers, calls an acquaintance with an embarrassed apology after having been unintentionally rude. In a debate, the sweetheart sees how he might be wrong and the other person right.

The moral and emotional failure of the jerk is obvious. The intellectual failure is obvious, too: No one is as right about everything as the jerk thinks he is. He would learn by listening. And one of the things he might learn is the true scope of his jerkitude – a fact about

which, as I will explain shortly, the all-out jerk is inevitably ignorant. This brings me to the other great benefit of a theory of jerks: It might help you figure out if you yourself are one.

#

Some clarifications and caveats.

First, no one is a perfect jerk or a perfect sweetheart. Human behavior – of course! – various hugely with context. Different situations (department meetings, traveling in close quarters) might bring out the jerk in some and the sweetie in others.

Second, the jerk is someone who *culpably* fails to appreciate the perspectives of others around him. Young children and people with severe cognitive disabilities aren't capable of appreciating others' perspectives, so they can't be blamed for their failure and aren't jerks. ("What a selfish jerk!" you say about the baby next to you on the bus, who is hollering and flinging her slobbery toy around. Of course you mean it only as a joke. Hopefully.) Also, not all perspectives deserve equal treatment. Failure to appreciate the outlook of a neo-Nazi, for example, is not a sign of jerkitude – though the true sweetheart might bend over backwards to try.

Third, I've called the jerk "he", since the best stereotypical examples of jerks tend to be male, for some reason. But then it seems too gendered to call the sweetheart "she", so I've made the sweetheart a "he" too.

#

I've said that my theory might help us assess whether we, ourselves, are jerks. In fact, this turns out to be a strangely difficult question. The psychologist Simine Vazire has argued

that we tend to know our own personality traits rather well when the traits are evaluatively neutral and straightforwardly observable, and badly when the traits are highly value-laden and not straightforward to observe.[8] If you ask people how talkative they are, or whether they are relatively high-strung or mellow, and then you ask their friends to rate them along those same dimensions, the self-ratings and the peer ratings usually correlate well – and both sets of ratings also tend to line up with psychologists' attempts to measure such traits objectively.

Why? Presumably because it's more or less fine to be talkative and more or less fine to be quiet; okay to be a bouncing bunny and okay instead to keep it low-key, and such traits are hard to miss in any case. But few of us want to be inflexible, stupid, unfair, or low in creativity. And if you don't want to see yourself that way, it's easy enough to dismiss the signs. Such characteristics are, after all, connected to outward behavior in somewhat complicated ways; we can always cling to the idea that we've been misunderstood by those who charge us with such defects. Thus, we overlook our faults.

With Vazire's model of self-knowledge in mind, I conjecture a correlation of approximately zero between how one would rate oneself in relative jerkitude and one's actual true jerkitude. The term is morally loaded, and rationalization is so tempting and easy! Why did you just treat that cashier so harshly? Well, she deserved it – and anyway, I've been having a rough day. Why did you just cut into that line of cars at the last moment, not waiting your turn to exit? Well, that's just good tactical driving – and anyway, I'm in a hurry! Why did you seem to relish failing that student for submitting his essay an hour late? Well, the rules were clearly stated; it's only fair to the students who worked hard to submit their essays on time – and that was a grimace not a smile.

Since probably the most effective way to learn about defects in one's character is to listen to frank feedback from people whose opinions you respect, the jerk faces special

obstacles on the road to self-knowledge, beyond even what Vazire's theory would lead us to expect. By definition, he fails to respect the perspectives of others around him. He's much more likely to dismiss critics as fools – or as jerks themselves – than to take the criticism to heart.

Still, it's entirely possible for a picture-perfect jerk to acknowledge, in a *superficial* way, that he is a jerk. "So what, yeah, I'm a jerk," he might say. Provided that this admission carries no real sting of self-disapprobation, the jerk's moral self-ignorance remains. Part of what it is to fail to appreciate the perspectives of others is to fail to see what's inappropriate in your jerkishly dismissive attitude toward their ideas and concerns.

Ironically, it is the sweetheart who worries that he has just behaved inappropriately, that he might have acted too jerkishly, and who feels driven to make amends. Such distress is impossible if you don't take others' perspectives seriously into account. Indeed, the distress itself constitutes a deviation (in this one respect at least) from pure jerkitude: Worrying about whether it might be so helps to make it less so. Then again, if you take comfort in that fact and cease worrying, you have undermined the very basis of your comfort.

#

Jerks normally distribute their jerkitude mostly *down* the social hierarchy, and to anonymous strangers. Waitresses, students, clerks, strangers on the road – these are the unfortunate people who bear the brunt of it. With a modicum of self-control, the jerk, though he implicitly or explicitly regards himself as more important than most of the people around him, recognizes that the perspectives of others above him in the hierarchy also deserve some consideration. Often, indeed, he feels sincere respect for his higher-ups. Maybe deferential impulses are too deeply written in our natures to disappear entirely. Maybe the jerk retains a

vestigial concern specifically for those he would benefit, directly or indirectly, from winning over. He is at least concerned enough about their opinion of him to display tactical respect while in their field of view. However it comes about, the classic jerk kisses up and kicks down. For this reason, the company CEO rarely knows who the jerks are, though it's no great mystery among the secretaries.

Because the jerk tends to disregard the perspectives of those below him in the hierarchy, he often has little idea how he appears to them. This can lead to ironies and hypocrisy. He might rage against the smallest typo in a student's or secretary's document, while producing a torrent of typos himself; it just wouldn't occur to him to apply the same standards to himself. He might insist on promptness, while always running late. He might freely reprimand other people, expecting them to take it with good grace, while any complaints directed against him earn his undying enmity. Such failures of parity typify the jerk's moral short-sightedness, flowing naturally from his disregard of others' perspectives. These hypocrisies are immediately obvious if one genuinely imagines oneself in a subordinate's shoes for anything other than selfish and self-rationalizing ends, but this is exactly what the jerk habitually fails to do.

Embarrassment, too, becomes practically impossible for the jerk, at least in front of his underlings. Embarrassment requires us to imagine being viewed negatively by people whose perspectives we care about. As the circle of people the jerk is willing to regard as true peers and superiors shrinks, so does his capacity for shame – and with it a crucial entry point for moral self-knowledge.

As one climbs the social hierarchy it is also easier to *become* a jerk. Here's a characteristically jerkish thought: "I'm important and I'm surrounded by idiots!" Both halves of this proposition serve to conceal the jerk's jerkitude from himself. Thinking yourself important is a pleasantly self-gratifying excuse for disregarding the interests and desires of

others. Thinking that the people around you are idiots seems like a good reason to dismiss their intellectual perspectives. As you ascend the social hierarchy, you will find it easier to discover evidence of your relative importance (your big salary, your first-class seat) and of the relative stupidity of others (who have failed to ascend as high as you). Also, flatterers will tend to squeeze out frank, authentic critics.

This isn't the only possible explanation for the prevalence of powerful jerks. Maybe natural, intuitive jerks are also more likely to rise in government, business, and academia than non-jerks. The truest sweethearts often suffer from an inability to advance their own projects over the projects of others. But I suspect the causal path runs at least as much the other direction. Success might or might not favor the existing jerks, but I'm pretty sure it nurtures new ones.

#

The *moralistic jerk* is an animal worth special remark. Charles Dickens was a master painter of the type: his teachers, his preachers, his petty bureaucrats and self-satisfied businessmen, Scrooge condemning the poor as lazy, Mr. Bumble shocked that Oliver Twist dares to ask for more food, each dismissive of the opinions and desires of their social inferiors, each inflated with a proud self-image and ignorant of how they are rightly seen by those around them, and each rationalizing this picture with a web of moralizing "shoulds".

Scrooge and Bumble are cartoons, and we can be pretty sure we aren't as bad as them. Yet I see in myself and all those who are not pure sweethearts a tendency to rationalize my privilege with moralistic sham justifications. Here's my reason for dishonestly trying to wheedle my daughter into the best school; my reason why the session chair should call on me

rather than on the grad student who got her hand up earlier; my reason why it's fine that I have 400 library books in my office….

Philosophers appear to have a special talent in concocting such dubious justifications: With enough work, we can concoct a moral rationalization for anything! Such skill at rationalization might partly explain why ethicist philosophers seem to behave no morally better, on average, than do comparison groups of non-ethicists, as my collaborators and I have found in a long series of empirical studies on issues ranging from returning library books, to courteous behavior at professional conferences, to rates of charitable giving, to membership in the Nazi party in 1930s Germany (see Chapters 4 and 53). The moralistic jerk's rationalizations justify his disregard of others, and his disregard of others prevents him from accepting an outside corrective on his rationalizations, in a self-insulating cycle. Here's why it's fine for him, he says, to neglect his obligations to his underlings and inflate his expense claims, you idiot critics. Coat the whole thing, if you like, in a patina of business-speak or academic jargon.

The moralizing jerk is apt to go badly wrong in his moral opinions. Partly this is because his morality tends to be self-serving, and partly it's because his disrespect for others' perspectives puts him at a general epistemic disadvantage. But there's more to it than that. In failing to appreciate others' perspectives, the jerk almost inevitably fails to appreciate the full range of human goods – the value of dancing, say, or of sports, nature, pets, local cultural rituals, and indeed anything that he doesn't personally care for. Think of the aggressively rumpled scholar who can't bear the thought that someone would waste her time getting a manicure. Or think of the manicured socialite who can't see the value of dedicating one's life to dusty Latin manuscripts. Whatever he's into, the moralizing jerk exudes a continuous aura of disdain for everything else.

Furthermore, *mercy* is near the heart of practical, lived morality. Virtually everything that everyone does falls short of perfection: one's turn of phrase is less than perfect, one arrives a bit late, one's clothes are tacky, one's gesture irritable, one's choice somewhat selfish, one's coffee less than frugal, one's melody trite. Practical mercy involves letting these imperfections pass forgiven or, better yet, entirely unnoticed. In contrast, the jerk appreciates neither others' difficulties in attaining all the perfections that he attributes to himself, nor the possibility that some portion of what he regards as flawed is in fact blameless. Hard moralizing principle therefore comes naturally to him. (Sympathetic mercy is natural to the sweetheart.) And on the rare occasions where the jerk is merciful, his indulgence is usually ill-tuned: the flaws he forgives are exactly the ones he sees in himself or has ulterior reasons to let slide. Consider another brilliant literary cartoon jerk: Severus Snape, the infuriating potions teacher in J.K. Rowling's novels, always eager to drop the hammer on Harry Potter or anyone else who happens to annoy him, constantly bristling with indignation, but wildly off the mark – contrasted with the mercy and broad vision of Dumbledore.

Despite the jerk's almost inevitable flaws in moral vision, the moralizing jerk can sometimes happen to be right about some specific important issue (as Snape proved to be) – especially if he adopts a big social cause. He needn't care only about money and prestige. Indeed, sometimes an abstract and general concern for moral or political principles serves as a substitute for concern about the people in his immediate field of view, possibly leading to substantial self-sacrifice. He might loathe and mistreat everyone around him, yet die for the cause. And in social battles, the sweetheart will always have some disadvantages: The sweetheart's talent for seeing things from the opponent's perspective deprives him of bold self-certainty, and he is less willing to trample others for his ends. Social movements sometimes do well when led by a moralizing jerk.

#

How can you know your own moral character? You can try a label on for size: "lazy", "jerk", "unreliable" – is that really me? As the work of Vazire and other personality psychologists suggests, this might not be a very illuminating approach. More effective, I suspect, is to shift from first-person reflection (what am *I* like?) to second-person description (tell me, what *am* I like?). Instead of introspection, try listening. Ideally, you will have a few people in your life who know you intimately, have integrity, and are concerned about your character. They can frankly and lovingly hold your flaws to the light and insist that you look at them. Give them the space to do this and prepare to be disappointed in yourself.

Done well enough, this second-person approach could work fairly well for traits such as laziness and unreliability, especially if their scope is restricted – laziness-about-X, unreliability-about-Y. But as I suggested above, jerkitude is probably not so tractable, since if one is far enough gone, one can't listen in the right way. Your critics are fools, at least on this particular topic (their critique of you). They can't appreciate your perspective, you think – though really it's that you can't appreciate theirs.

To discover one's degree of jerkitude, the best approach might be neither (first-person) direct reflection upon yourself nor (second-person) conversation with intimate critics, but rather something more third-person: looking in general at other people. Everywhere you turn, are you surrounded by fools, by boring nonentities, by faceless masses and foes and suckers and, indeed, jerks? Are you the only competent, reasonable person to be found? In other words, how familiar was the vision of the world I described at the beginning of this essay?

If your self-rationalizing defenses are low enough to feel a little pang of shame at the familiarity of that vision of the world, then you probably aren't pure diamond-grade jerk. But who is? We're all somewhere in the middle. That's what makes the jerk's vision of the world so instantly recognizable. It's our own vision. But, thankfully, only sometimes.

2. Forgetting as an Unwitting Confession of Your Values

Every September 11, my social media feeds are full of reminders to "never forget" the Twin Tower terrorist attacks. Similarly, the Jewish community insists that we keep vivid the memory of the Holocaust. It says something about a person's values, what they strive to remember – a debt, a harm, a treasured moment, a loved one now gone, an error or lesson.

What we remember says, perhaps, more about us than we would want. Forgetfulness can be an unwitting confession of your values. The Nazi Adolf Eichmann, in Hannah Arendt's famous portrayal of him, had little memory of his decisions about shipping thousands of Jews to their deaths, but he remembered in detail small social triumphs with his Nazi superiors. The transports he forgot – but how vividly he remembers the notable occasion when he was permitted to lounge beside a fireplace with Reinhard Heydrich, watching the Nazi leader smoke and drink![9] Eichmann's failures and successes of memory are more eloquent and accurate testimony of his values than any of his outward avowals.

I remember obscure little arguments in philosophy articles if they are relevant to an essay I'm working on, but I can't seem to recall the names of the parents of my children's friends. Some of us remember insults and others forget them. Some remember the exotic foods they ate on vacation, others the buildings they saw, others the wildlife, and still others hardly anything specific at all.

From the leavings of memory and forgetting we could create a map, I think, of a person's values. Features of the world that you don't see – the subtle sadness in a colleague's face? – and features that you briefly see but don't react to or retain, are in some sense not part of the world shaped for you by your interests and values. Other people with different values will remember a very different series of events.

To carve David, simply remove everything from the stone that is not David.[10] Remove from your life everything you forget; what is left is you.

3. The Happy Coincidence Defense and The-Most-You-Can-Do Sweet Spot

Here are four things I care intensely about: being a good father, being a good philosopher, being a good teacher, and being a morally good person. It would be lovely if there were never any tradeoffs among these four aims.

It is highly unpleasant to acknowledge such tradeoffs – sufficiently unpleasant that most of us try to rationalize them away. It's distinctly uncomfortable to me, for example, to acknowledge that I would probably be a better father if I traveled less for work. (I am writing now from a hotel room in England.) Similarly uncomfortable is the thought that the money I will spend this summer on a family trip to Iceland could probably save a few people from death due to poverty-related causes, if given to the right charity.[11]

Below are two of my favorite techniques for rationalizing such unpleasant thoughts away. Maybe you'll find these techniques useful too!

#

The Happy Coincidence Defense

Consider travel for work. I don't really need to travel around the world giving talks and meeting people. No one will fire me if I don't do it, and some of my colleagues do it much less. On the face of it, I seem to be prioritizing my research career at the cost of being a somewhat less good father, teacher, and global moral citizen (given the pollution of air travel and the luxurious use of resources).

The Happy Coincidence Defense says, no, in fact I am not sacrificing these other goals at all. Although I am away from my children, I am a better father for it. I am a role model of career success for them, and I can tell them stories about my travels. I have

enriched my life, and I can mingle that richness into theirs. I am a more globally aware, wiser father! Similarly, though I might cancel a class or two and de-prioritize lecture preparation, since research travel improves me as a philosopher it improves my teaching in the long run. And my philosophical work, isn't that an important contribution to society? Maybe it's important enough to justify the expense, pollution, and waste: I do more good for the world jetting around to talk philosophy than I could do by leading a more modest lifestyle at home, working within my own community.

After enough reflection, it can come to seem that I'm not making any tradeoffs at all among these four things I care intensely about. Instead I am maximizing them all. This trip to England is the best thing I can do, all things considered, as a philosopher *and* as a father *and* as a teacher *and* as a citizen of the moral community. Yay!

Now that *might* be true. If so, it would be a happy coincidence. Sometimes there really are such happy coincidences. We should aim to structure our lives and societies to enhance the likelihood of such coincidences. But still, I think you'll agree that the pattern of reasoning is suspicious. Life is full of tradeoffs and hard choices. I'm probably just rationalizing, trying to convince myself that something I want to be true really is true.

#

The-Most-You-Can-Do Sweet Spot

Sometimes trying too hard at something makes you do worse. Trying too hard to be a good father might make you overbearing and invasive. Overpreparing a lecture can spoil your spontaneity. And sometimes, maybe, moral idealists push themselves so hard in support of their ideals that they collapse along the way. For example, someone moved by the arguments for vegetarianism who immediately attempts the very strictest veganism might be

more likely to revert to cheeseburger eating after a few months than someone who sets their sights a bit lower.

The-Most-You-Can-Do Sweet Spot reasoning runs like this: Whatever you're doing right now is the most you can realistically, sustainably do. Were I, for example, to try any harder to be a good father, I would end up being a worse father. Were I to spend any more time reading and writing philosophy than I already do, I would only exhaust myself, or I'd lose my freshness of ideas. If I gave any more to charity, or sacrificed any more for the well-being of others in my community, then I would… I would… I don't know, collapse from charity fatigue? Bristle with so much resentment that it undercuts my good intentions?

As with Happy Coincidence reasoning, The-Most-You-Can-Do Sweet Spot reasoning can sometimes be right. Sometimes you really are doing the most you can do about something you care intensely about, so that if you tried to do any more it would backfire. Sometimes you don't need to compromise: If you tried any harder or devoted any more time, it really would mess things up. But it would be amazing if this were reliably the case. I probably could be a better father, if I spent more time with my children. I probably could be a better teacher, if I gave more energy to my students. I probably could be a morally better person, if I just helped others a little bit more. If I typically think that wherever I happen to be, that's already the Sweet Spot, I am probably rationalizing.

#

By giving these ordinary phenomena cute names, I hope to make them more salient and laughable. I hope to increase the chance that the next time you or I rationalize in this way, some little voice pops up to say, in gentle mockery, "Ah, lovely! What a Happy

Coincidence that is. You're in the Most-You-Can-Do Sweet Spot!" And then, maybe, we can drop the excuse and aim for better.

4. Cheeseburger Ethics (or How Often Do Ethicists Call Their Mothers?)

None of the classic questions of philosophy is beyond a seven-year-old's understanding. If God exists, why do bad things happen? How do you know there's still a world on the other side of that closed door? Are we just made out of material stuff that will turn into mud when we die? If you could get away with robbing people just for fun, would it be reasonable to do it? The questions are natural. It's the answers that are hard.

In 2007, I'd just begun a series of empirical studies on the moral behavior of professional ethicists. My son Davy, then seven years old, was in his booster seat in the back of my car. "What do you think, Davy?" I asked. "People who think a lot about what's fair and about being nice – do they behave any better than other people? Are they more likely to be fair? Are they more likely to be nice?"

Davy didn't respond right away. I caught his eye in the rearview mirror.

"The kids who always talk about being fair and sharing," I recall him saying, "mostly just want you to be fair to them and share with them."

#

When I meet an ethicist for the first time – by "ethicist", I mean a professor of philosophy who specializes in teaching and researching ethics – it's my habit to ask whether they think that ethicists behave any differently from other types of professor. Most say no.

I'll probe further. Why not? Shouldn't regularly thinking about ethics have some sort of influence on one's behavior? Doesn't it seem that it would?

To my surprise, few professional ethicists seem to have given the question much thought. They'll toss out responses that strike me as flip or as easily rebutted, and then

they'll have little to add when asked to clarify. They'll say that academic ethics is all about abstract problems and bizarre puzzle cases, with no bearing on day-to-day life – a claim easily shown to be false by a few examples: Aristotle on virtue, Kant on lying, Singer on charitable donation. They'll say, "What, do you expect epistemologists to have more knowledge? Do you expect doctors to be less likely to smoke?" I'll reply that the empirical evidence does suggest that doctors are less likely to smoke than non-doctors of similar social and economic background.[12] Maybe epistemologists don't have more knowledge, but I'd hope that specialists in feminism were less biased against women – and if they weren't, that would be interesting to know. I'll suggest that the relationship between professional specialization and personal life might play out differently for different professions.

It seems odd to me that philosophers have so little to say about this matter. We criticize Martin Heidegger for his Nazism, and we wonder how deeply connected his Nazism was to his other philosophical views.[13] But we don't feel the need to turn the mirror on ourselves.

The same issues arise with clergy. In 2010, I was presenting some of my work at the Confucius Institute for Scotland. Afterward, I was approached by not one but two bishops. I asked them whether they thought that clergy, on average, behaved better, the same, or worse than laypeople.

"About the same," said one.

"Worse!" said the other.

No clergyperson has ever expressed to me the view that clergy behave on average better than laypeople, despite all their immersion in religious teaching and ethical conversation. Maybe in part this is modesty on behalf of their profession. But in most of their voices, I also hear something that sounds like genuine disappointment, some remnant of the young adult who had headed off to seminary hoping it would be otherwise.

#

In a series of empirical studies starting in 2007 – mostly in collaboration with the philosopher Joshua Rust – I have empirically explored the moral behavior of ethics professors. As far as I know, Josh and I are the only people ever to have done so in a systematic way.[14]

Here are the measures we've looked at: voting in public elections, calling one's mother, eating the meat of mammals, donating to charity, littering, disruptive chatting and door-slamming during philosophy presentations, responding to student emails, attending conferences without paying registration fees, organ donation, blood donation, theft of library books, overall moral evaluation by one's departmental peers, honesty in responding to survey questions, and joining the Nazi party in 1930s Germany.

Obviously, some of these measures are more significant than others. They range from trivialities (littering) to substantial life decisions (joining the Nazis), and from contributions to strangers (blood donation) to personal interactions (calling Mom). Some of our measures rely on self-report (we didn't ask ethicists' mothers how long it had *really* been) but in many cases we had direct observational evidence.

Ethicists do not appear to behave better. Never once have we found ethicists as a whole behaving better than our comparison groups of other professors, by any of our main planned measures. But neither, overall, do they seem to behave worse. (There are some mixed results for secondary measures and some cases where it matters who is the comparison group.) For the most part, ethicists behave no differently from other sorts of professors – logicians, biologists, historians, foreign-language instructors.

Nonetheless, ethicists do embrace more stringent moral norms on some issues, especially vegetarianism and charitable donation. Our results on vegetarianism were especially striking. In a survey of professors from five U.S. states, we found that 60% of ethicist respondents rated "regularly eating the meat of mammals, such as beef or pork" somewhere on the "morally bad" side of a nine-point scale ranging from "very morally bad" to "very morally good". In contrast, only 19% of professors in departments other than philosophy rated it as bad. That's a pretty big difference of opinion! Non-ethicist philosophers were intermediate, at 45%. But when asked later in the survey whether they had eaten the meat of a mammal at their previous evening meal, we found no statistically significant difference in the groups' responses – 38% of professors reported having done so, including 37% of ethicists.[15]

Similarly for charitable donation. In the same survey, we asked respondents what percentage of income, if any, the typical professor should donate to charity, and then later we asked what percentage of income they personally had given in the previous calendar year. Ethicists espoused the most stringent norms: their average recommendation was 7%, compared with 5% for the other two groups. However, ethicists did not report having given a greater percentage of income to charity than the non-philosophers (4% for both groups). Nor did adding a charitable incentive to half of our surveys (a promise of a $10 donation to their selected charity from a list) increase ethicists' likelihood of completing the survey. Interestingly, the non-ethicist philosophers, though they reported having given the least to charity (3%), were the only group that responded to our survey at detectably higher rates when given the charitable incentive.[16]

#

Should we expect ethicists to behave especially morally well as a result of their training – or at least more in accord with the moral norms that they themselves espouse?

Maybe we can defend a "no". Consider this thought experiment:

An ethics professor teaches Peter Singer's arguments for vegetarianism to her undergraduates.[17] She says she finds those arguments sound and that in her view it is morally wrong to eat meat. Class ends, and she goes to the cafeteria for a cheeseburger. A student approaches her and expresses surprise at her eating meat. (If you don't like vegetarianism as an issue, consider marital fidelity, charitable donation, fiscal honesty, or courage in defense of the weak.)

"Why are you surprised?" asks our ethicist. "Yes, it is morally wrong for me to enjoy this delicious cheeseburger. However, I don't aspire to be a saint. I aspire only to be about as morally good as others around me. Look around this cafeteria. Almost everyone is eating meat. Why should I sacrifice this pleasure, wrong though it is, while others do not? Indeed, it would be unfair to hold me to higher standards just because I'm an ethicist. I am paid to teach, research, and write, like every other professor. I am paid to apply my scholarly talents to evaluating intellectual arguments about the good and the bad, the right and the wrong. If you want me also to live as a role model, you ought to pay me extra!

"Furthermore," she continues, "if we demand that ethicists live according to the norms they espouse, that will put major distortive pressures on the field. An ethicist who feels obligated to live as she teaches will be motivated to avoid highly self-sacrificial conclusions, such as that the wealthy should give most of their money to charity or that we should eat only a restricted subset of foods. Disconnecting professional ethicists' academic inquiries from their personal choices allows them to consider the arguments in a more even-handed way. If no one expects us to act in accord with our scholarly opinions, we are more likely to arrive at the moral truth."

"In that case," replies the student, "is it morally okay for me to order a cheeseburger too?"

"No! Weren't you listening? It would be wrong. It's wrong for me, also, as I just admitted. I recommend the avocado and sprouts. I hope that Singer's and my arguments help create a culture permanently free of the harms to animals and the environment that are caused by meat-eating."

"This reminds me of Thomas Jefferson's attitude toward slave ownership," I imagine the student replying. Maybe the student is Black.

"Perhaps so. Jefferson was a great man. He had the courage to recognize that his own lifestyle was morally odious. He acknowledged his mediocrity and resisted the temptation to paper things over with shoddy arguments. Here, have a fry."

Let's call this view *cheeseburger ethics.*

#

Any of us could easily become much morally better than we are, if we chose to. For those of us who are affluent by global standards, one path is straightforward: Spend less on luxuries and give the savings to a good cause. Even if you aren't affluent by global standards, unless you are on the precipice of ruin, you could give more of your time to helping others. It's not difficult to see multiple ways, every day, in which one could be kinder to those who would especially benefit from kindness.

And yet, most of us choose mediocrity instead. It's not that we try but fail, or that we have good excuses. We – most of us – actually aim at mediocrity. The cheeseburger ethicist is perhaps only unusually honest with herself about this. We aspire to be about as morally good as our peers. If others cheat and get away with it, we want to do the same. We don't

want to suffer for goodness while others laughingly gather the benefits of vice. If the morally good life is uncomfortable and unpleasant, if it involves repeated painful sacrifices that aren't compensated in some way, sacrifices that others aren't also making, then we don't want it.

Recent empirical work in moral psychology and experimental economics, especially by Robert B. Cialdini and Cristina Bicchieri, seems to confirm this general tendency.[18] People are more likely to comply with norms that they see others following, less likely to comply with norms when they see others violating them. Also, empirical research on "moral self-licensing" suggests that people who act well on one occasion often use that as an excuse to act less well subsequently.[19] We gaze around us, then aim for so-so.

What, then, is moral reflection good for? Here's one thought. Maybe it gives us the power to calibrate more precisely toward our chosen level of mediocrity. I sit on the couch, resting while my wife cleans up from dinner. I know that it would be morally better to help than to continue relaxing. But how bad, exactly, would it be for me not to help? Pretty bad? Only a little bad? Not at all bad, but also not as good as I'd like to be if I weren't feeling so lazy? These are the questions that occupy my mind. In most cases, we already know what is good. No special effort or skill is needed to figure that out. Much more interesting and practical is the question of how far short of the ideal we are comfortable being.

Suppose it's generally true that we aim for goodness only by peer-relative, rather than absolute, standards. What, then, should we expect to be the effect of discovering, say, that it is morally bad to eat meat, as the majority of U.S. ethicists seem to think? If you're trying to be only about as good as others, and no better, then you can keep enjoying the cheeseburgers. Your behavior might not change much at all. What would change is this: You'd acquire a lower opinion of (almost) everyone's behavior, your own included.

You might hope that others will change. You might advocate general social change – but you'll have no desire to go first. Like Jefferson maybe.

#

I was enjoying dinner in an expensive restaurant with an eminent ethicist, at the end of an ethics conference. I tried these ideas out on him.

"B+," he said. "That's what I'm aiming for."

I thought, but I didn't say, *B+ sounds good*. Maybe that's what I'm aiming for too. B+ on the great moral curve of White middle-class college-educated North Americans. Let others get the As.

Then I thought, most of us who are aiming for B+ will probably fall well short. You know, because we fool ourselves. Here I am away from my children again, at a well-funded conference in a beautiful $200-a-night hotel, mainly, I suspect, so that I can nurture and enjoy my rising prestige as a philosopher. What kind of person am I? What kind of father? B+?

(Oh, it's excusable! – I hear myself saying. I'm a model of career success for the kids, and of independence. And morality isn't so demanding. And my philosophical work is a contribution to the general social good. And I give, um, well, a little to charity, so that makes up for it. And I'd be too disheartened if I couldn't do this kind of thing, which would make me worse as a father and as a teacher of ethics. Plus, I owe it to myself. And…. Wow, how neatly what I want to do fits with what's ethically best, once I think about it! [See Chapter 3 on Happy Coincidence Reasoning.])

A couple of years later, I emailed that famous ethicist about the B+ remark, to see if I could quote him on it by name. He didn't recall having said it, and he denied that was his view. He is aiming for moral excellence after all. It must have been the chardonnay speaking.

#

Most of the ancient philosophers and great moral visionaries of the religious wisdom traditions, East and West, would find the cheeseburger ethicist strange. Most of them assumed that the main purpose of studying ethics was self-improvement. Most of them also accepted that philosophers were to be judged by their actions as much as by their words. A great philosopher was, or should be, a role model: a breathing example of a life well-lived. Socrates taught as much by drinking the hemlock as by any of his dialogues, Confucius by his personal correctness, Siddhartha Guatama by his renunciation of wealth, Jesus by washing his disciples' feet. Socrates does not say: Ethically, the right thing for me to do would be to drink this hemlock, but I will flee instead! (Maybe he could have said that, but then he would have been a different sort of model.)

I'd be suspicious of a 21st-century philosopher who offered up her- or himself as a model of wise living. This is no longer what it is to be a philosopher – and those who regard themselves as especially wise are in any case usually mistaken. Still, I think, the ancient philosophers got something right that the cheeseburger ethicist gets wrong.

Maybe it's this: I have available to me the best attempts of earlier generations to express their ethical understanding of the world. I even seem to have some advantages over ancient philosophers, in that there are now many more centuries of written texts and several distinct cultures with long traditions that I can compare. And I am paid, quite handsomely by global standards, to devote a large portion of my time to thinking through this material. What should I do with this amazing opportunity? Use it to get some publications and earn praise from my peers, plus a higher salary? Sure. Use it – as my seven-year-old son observed – as a tool to badger others into treating me better? Okay, I guess so, sometimes. Use it to try to

shape other people's behavior in a way that will make the world a generally better place? Simply enjoy its power and beauty for its own sake? Yes, those things too.

But also, it seems a waste not to try to use it to make myself ethically better than I currently am. Part of what I find unnerving about the cheeseburger ethicist is that she seems so comfortable with her mediocrity, so uninterested in deploying her philosophical tools toward self-improvement. Presumably, if approached in the right way, the great traditions of moral philosophy have the potential to help us become morally better people. In cheeseburger ethics, that potential is cast aside.

The cheeseburger ethicist risks intellectual failure as well. Real engagement with a philosophical doctrine probably requires taking some steps toward living it. The person who takes, or at least tries to take, personal steps toward Kantian scrupulous honesty, or Mozian impartiality, or Buddhist detachment, or Christian compassion, gains a kind of practical insight into those doctrines that is not easily achieved through intellectual reflection alone. A full-bodied understanding of ethics requires some relevant life experience.

What's more, abstract doctrines lack specific content if they aren't tacked down in a range of concrete examples. Consider the doctrine "treat everyone as equals who are worthy of respect". What counts as adhering to this norm, and what constitutes a violation of it? Only when we understand how norms play out across examples do we really understand them.[20] Living our norms, or trying to live them, forces a maximally concrete confrontation with examples. Does your ethical vision really require that you free the slaves on which your lifestyle crucially depends? Does it require giving away your salary and never again enjoying an expensive dessert? Does it require drinking hemlock if your fellow citizens unjustly demand that you do so?

Few professional ethicists really are cheeseburger ethicists, I think, when they stop to consider it. They – or rather we, I suppose, as I find myself becoming more and more an

ethicist, do want our ethical reflections to improve us morally, a little bit. But here's the catch: We aim only to become *a little* morally better. We cut ourselves slack when we look at others around us. We grade ourselves on a curve and tell ourselves, if pressed, that we're aiming for B+ rather than A. And at the same time, we excel at rationalization and excuse-making – maybe more so, the more ethical theories we have ready to hand. So we end, on average, about where we began, behaving more or less the same as others of our social group.

Should we aim for "A+", then? Being frank with myself, I don't want the self-sacrifice I'm pretty sure would be required. Should I aim at least a little higher than B+? Should I resolutely aim to be morally far better than my peers – A or maybe A-minus – even if not quite a saint? I worry that needing to see myself as unusually morally excellent is more likely to increase self-deception and rationalization than to actually improve me.

Should I redouble my efforts to be kinder and more generous, coupling them with reminders of humility about my likelihood of success? Yes, I will – today! But I already feel my resentment building, and I haven't even done anything yet. Maybe I can escape that resentment by adjusting my sense of "mediocrity" upward. I might try to recalibrate by surrounding myself with like-minded peers in virtue. But avoiding the company of those I deem morally inferior seems more characteristic of the moralizing jerk than of the genuinely morally good person.

I can't quite see my way forward. But now I worry that this, too, is excuse-making. Nothing will ensure success, so (phew!) I can comfortably stay in the same old mediocre place I'm accustomed to. This defeatism also fits nicely with one natural way to read Josh Rust's and my data: Since ethicists don't behave better or worse than others, philosophical reflection must be behaviorally inert, taking us only where we were already headed, its power mainly that of providing different words by which to decorate our pre-determined choices.[21] So I'm not to be blamed if all my ethical philosophizing hasn't improved me.

I reject that view. Instead I favor this less comfortable idea: Philosophical reflection does have the power to move us, but it is not a tame thing. It takes us where we don't intend or expect, sometimes one way, as often the other, sometimes amplifying our vices and illusions, sometimes yielding real insight and inspiring substantial moral change. These tendencies cross-cut and cancel in complex ways that are difficult to detect empirically. If we could tell in advance which direction our reflection would carry us and how, we'd be implementing a pre-set educational technique rather than challenging ourselves philosophically.

Genuine philosophical thinking critiques its prior strictures, including even the assumption that we ought to be morally good. It damages almost as often as it aids, is free, wild, and unpredictable, always breaks its harness. It will take you somewhere, up, down, sideways – you can't know in advance. But you are responsible for trying to go in the right direction with it, and also for your failure when you don't get there.

5. On Not Seeking Pleasure Much

Back in the 1990s, when I was a graduate student, my girlfriend Kim asked me what, of all things, I most enjoyed doing. Skiing, I answered. I was thinking of those moments breathing the cold, clean air, relishing the mountain view, then carving a steep, lonely slope. I'd done quite a lot of that with my mom when I was a teenager. But how long had it been since I'd gone skiing? Maybe three years? Grad school kept me busy and I now had other priorities for my winter breaks. Kim suggested that if it had been three years since I'd done what I most enjoyed doing, then maybe I wasn't living wisely.

Well, what, I asked, did she most enjoy? Getting massages, she said. Now, the two of us had a deal at the time: If one gave the other a massage, the recipient would owe a massage in return the next day. We exchanged massages occasionally, but not often, maybe once every few weeks. I pointed out that she, too, might not be perfectly rational: She could easily get much more of what she most enjoyed simply by giving me more massages. Surely the displeasure of massaging my back couldn't outweigh the pleasure of the thing she most enjoyed in the world? Or was pleasure for her such a tepid thing that even the greatest pleasure she knew was hardly worth getting?

Suppose it's true that avoiding displeasure is much more motivating than gaining pleasure, so that even our top pleasures (skiing, massages) aren't motivationally powerful enough to overcome only moderate displeasures (organizing a ski trip, giving a massage).[22] Is this rational? Is displeasure more bad than pleasure is good? Is it much better to have a steady neutral than a mix of highs and lows? If so, that might explain why some people are attracted to the Stoics' and Buddhists' emphasis on avoiding suffering, even at the cost of losing opportunities for pleasure.[23]

Or is it irrational not to weigh pleasure and displeasure evenly? In dealing with money, people will typically, and seemingly irrationally, do much more to avoid a loss than

to secure an equivalent gain.[24] Maybe it's like that? Maybe sacrificing two units of pleasure to avoid one unit of displeasure is like irrationally forgoing a gain of $2 to avoid a loss of $1?

It used to be a truism in Western, especially British, philosophy that people sought pleasure and avoided pain. A few old-school psychological hedonists, like Jeremy Bentham, went so far as to say that was *all* that motivated us.[25] I'd guess quite differently: Although pain is moderately motivating, pleasure motivates us very little. What motivates us more are outward goals, especially socially approved goals – raising a family, building a career, winning the approval of peers – and we will suffer immensely for these things. Pleasure might bubble up as we progress toward these goals, but that's a bonus and side-effect, not the motivating purpose, and summed across the whole, the displeasure might vastly outweigh the pleasure. Evidence suggests that even raising a child is probably for most people a hedonic net negative, adding stress, sleep deprivation, and unpleasant chores, as well as crowding out the pleasures that childless adults regularly enjoy.[26]

Have you ever watched a teenager play a challenging video game? Frustration, failure, frustration, failure, slapping the console, grimacing, swearing, more frustration, more failure – then finally, woo-hoo! The sum over time has got to be negative; yet they're back again to play the next game. For most of us, biological drives and addictions, personal or socially approved goals, concern for loved ones, habits and obligations – all appear to be better motivators than gaining pleasure, which we mostly seem to save for the little bit of free time left over.

If maximizing pleasure is central to living well and improving the world, we're going about it entirely the wrong way. Do you really want to maximize pleasure? I doubt it. Me, I'd rather write some good philosophy and raise my kids.

6. How Much Should You Care about How You Feel in Your Dreams?

Prudential hedonists say that how well your life is going for you, your personal well-being, is wholly constituted by facts about how much pleasure or enjoyment you experience and how much pain, displeasure, or suffering you experience. (This is different from *motivational* hedonism, discussed in Chapter 5, which concerns what moves us to act.) Prudential hedonism is probably a minority view among professional philosophers: Most would say that personal well-being is also partly constituted by facts about your flourishing or the attainment of things that matter to you – good health, creative achievement, loving relationships – even if that flourishing or attainment isn't fully reflected in positive emotional experiences.[27] Nevertheless, prudential hedonism might seem to be an important *part* of the story: Improving the ratio of pleasure to displeasure in life might be central to living wisely and structuring a good society.

Often a dream is the most pleasant or unpleasant thing that occurs all day. Discovering that you can fly – whee! How much do you do in waking life that is as fun as that? Conversely, how many things in waking life are as unpleasant as a nightmare? Most of your day you ride on an even keel, with some ups and downs; but at night your dreams bristle with thrills and anguish.

Here's a great opportunity, then, to advance the hedonistic project! Whatever you can do to improve the ratio of pleasant to unpleasant dreams should have a big impact on the ratio of pleasure to displeasure in your life.

This fact explains the emphasis prudential hedonists and utilitarian ethicists have placed on improving one's dream life. It also explains the gargantuan profits of all those dream-improvement mega-corporations.

Not. Of course not! When I ask people how concerned they are about the overall hedonic balance of their dreams, their response is almost always a shrug. But if the overall sum of felt pleasure and displeasure is important, shouldn't we take at least somewhat seriously the quality of our dream lives?

Dreams are usually forgotten, but I'm not sure how much that matters. Most people forget most of their childhood, and within a week they forget almost everything that happened on a given day. That doesn't make the hedonic quality of those events irrelevant. Your two-year-old might entirely forget the birthday party a year later, but you still want her to enjoy it, right? And anyway, we can work to remember our dreams if we want. Simply attempting to remember one's dreams upon waking, by jotting some notes into a dream diary, dramatically increases dream recall.[28] So if recall were important, you could work toward improving the hedonic quality of your dreams (maybe by learning lucid dreaming?[29]) and also work to improve your dream memory. The total impact on the amount of remembered pleasure in your life could be enormous!

I can't decide whether the fact that I haven't acted on this sensible advice illustrates my irrationality or instead illustrates that I care even less about pleasure and displeasure than I said in Chapter 5.[30]

7. Imagining Yourself in Another's Shoes vs. Extending Your Love

There's something I don't like about the "Golden Rule", the admonition to do unto others as you would have others do unto you.

Consider this passage from the ancient Chinese philosopher Mengzi (Mencius):

> That which people are capable of without learning is their genuine capability. That which they know without pondering is their genuine knowledge. Among babes in arms there are none that do not know to love their parents. When they grow older, there are none that do not know to revere their elder brothers. Treating one's parents as parents is benevolence. Revering one's elders is righteousness. There is nothing else to do but extend these to the world.[31]

One thing I like about the passage is that it assumes love and reverence for one's family as a given, rather than as a special achievement. It portrays moral development simply as a matter of extending that natural love and reverence more widely.

In another famous passage, Mengzi notes the kindness that the vicious tyrant King Xuan exhibits in saving a frightened ox from slaughter, and he urges the king to extend similar kindness to the people of his kingdom.[32] Such extension, Mengzi says, is a matter of "weighing" things correctly – a matter of treating similar things similarly and not overvaluing what merely happens to be nearby. If you have pity for an innocent ox being led to slaughter, you ought to have similar pity for the innocent people dying in your streets and on your battlefields, despite their invisibility beyond your beautiful palace walls.

The Golden Rule works differently – and so too the common advice to imagine yourself in someone else's shoes. Golden Rule / others' shoes advice assumes self-interest as the starting point, and implicitly treats overcoming egoistic selfishness as the main cognitive and moral challenge.

This contrasts sharply with Mengzian extension, which starts from the assumption that you are already concerned about nearby others, and takes the challenge to be extending that concern beyond a narrow circle.

Maybe we can model Golden Rule / others' shoes thinking like this:

(1.) If I were in the situation of Person X, I would want to be treated according to Principle P.

(2.) Golden Rule: Do unto others as you would have others do unto you.

(3.) Thus, I will treat Person X according to Principle P.

And maybe we can model Mengzian extension like this:

(1.) I care about Person Y and want to treat them according to Principle P.

(2.) Person X, though perhaps more distant, is relevantly similar.

(3.) Thus, I will treat Person X according to Principle P.

There will be other more careful and detailed formulations, but this sketch captures the central difference between these two approaches to moral cognition. Mengzian extension models general moral concern on the natural concern we already have for people close to us, while the Golden Rule models general moral concern on concern for oneself.

I like Mengzian extension better for three reasons.

First, Mengzian extension is more psychologically plausible as a model of moral development. People do, naturally, have concern and compassion for others around them. Explicit exhortations aren't needed to produce this. This natural concern and compassion is likely to be the main seed from which mature moral cognition grows. Our moral reactions to vivid, nearby cases become the bases for more general principles and policies. If you need to reason or analogize your way into concern even for close family members, you're already in deep moral trouble.

Second, Mengzian extension is less ambitious – in a good way. The Golden Rule imagines a leap from self-interest to generalized good treatment of others. This may be excellent and helpful advice, perhaps especially for people who are already concerned about others and thinking about how to implement that concern. But Mengzian extension has the advantage of starting the cognitive project much nearer the target, requiring less of a leap. Self to other is a huge moral and ontological divide. Family to neighbor, neighbor to fellow citizen – that's much less of a divide.

Third, Mengzian extension can be turned back on yourself, if you are one of those people who has trouble standing up for your own interests – if you are, perhaps, too much of a sweetheart (in the sense of Chapter 1). You would want to stand up for your loved ones and help them flourish. Apply Mengzian extension, and offer the same kindness to yourself. If you'd want your father to be able to take a vacation, realize that you probably deserve a vacation too. If you wouldn't want your sister to be insulted by her spouse in public, realize that you too shouldn't have to suffer that indignity.[33]

Although Mengzi and the 18th-century French philosopher Jean-Jacques Rousseau both endorse mottoes standardly translated as "human nature is good" and have views that are similar in important ways,[34] this is one difference between them. In both *Emile* and *Discourse on Inequality,* Rousseau emphasizes self-concern as the root of moral development, making pity and compassion for others secondary and derivative. He endorses the foundational importance of the Golden Rule, concluding that "Love of men derived from love of self is the principle of human justice".[35]

This difference between Mengzi and Rousseau is not a general difference between East and West. Confucius, for example, endorses something like the Golden Rule: "Do not impose on others what you yourself do not desire".[36] Mozi and Xunzi, also writing in the period, imagine people acting mostly or entirely selfishly, until society artificially imposes

its regulations, and so they see the enforcement of rules rather than Mengzian extension as the foundation of moral development.[37] Moral extension is thus specifically Mengzian rather than generally Chinese.

Care about me not because you can imagine what you would selfishly want if you were me. Care about me because you see how I am not really so different from others you already love.

8. Is It Perfectly Fine to Aim for Moral Mediocrity?

As I suggested in Chapter 4, as well as in other work,[38] most people aim to be morally mediocre. They aim to be about as morally good as their peers, not especially better, not especially worse. They don't want to be the one honest person in a classroom of unpunished cheaters; they don't want to be the only one of five housemates who reliably cleans up her messes and returns what she borrows; they don't want to turn off the air conditioner and let the lawn die to conserve energy and water if their neighbors aren't doing the same. But neither do most people want to be the worst sinner around – the most obnoxious of the housemates, the lone cheater in a class of honorable students, the most wasteful homeowner on the block.

Suppose I'm right about that. What is the ethics of aiming for moral mediocrity? It kind of sounds bad, the way I put it – aiming for mediocrity. But maybe it's not so bad, really? Maybe there's nothing wrong with aiming for the moral middle. "Cs get degrees!" Why not just go for a low pass – just enough to squeak over the line, if we're grading on a curve, while simultaneously enjoying the benefits of moderate, commonplace levels of deception, irresponsibility, screwing people over, and destroying the world's resources for your favorite frivolous luxuries? Maybe that's good enough, really, and we should save our negative judgments for what's more uncommonly rotten.[39]

There are two ways you could argue that it's not bad to aim for less than moral excellence. You could argue that although it's somewhat *morally* bad to aim for moral mediocrity, in some other broader sense of "bad" it's not overall bad to be a little bit morally bad. Morality isn't everything, after all. It might (in some sense) fine and (in some sense) reasonable to trade morality off against other goals you have. Maybe that's what the "cheeseburger ethicist" (Chapter 4) is doing, and what I'm doing when I face up to the fact that I'm always compromising among goods like fatherhood, research, teaching, and morality

(rather than being in the Most-You-Can-Do-Sweet Spot or maximizing them all by Happy Coincidence: Chapter 3).

Alternatively, you could argue that it's *not at all morally bad* to aim for moral mediocrity – or, more neutrally phrased, moral averageness. Maybe it would be morally better to aim for moral excellence, but that doesn't mean that it's wrong or bad to shoot instead for the middle. Average coffee isn't bad; it's just not as good as it could be. Average singers aren't bad; they're just not as… no, I take it back. Average singers are bad. (That they may be joyously, life-affirmingly bad, as in the shower or in singing "Happy Birthday" together, does not change their musical ineptitude.) So that's the question. Is morality more like coffee or singing?

I'd argue that morality is a bit like singing while drinking coffee. There's something fine about being average; but there are also some sour notes that can't entirely be wished away. I think you'll agree with me, if you consider the moral character of your neighbors and coworkers: They're not horrible, but neither are they above moral criticism. If the average person is aiming for approximately that, the average person is somewhat morally criticizable for their low moral ambitions.

The best argument I can know of that it's perfectly morally fine to aim to be morally average is what I'll call the Fairness Argument. But before I get to it, two clarifications:

First, aiming for moral mediocrity or averageness, in the sense I intend, doesn't require conceptualizing ourselves as doing so. We might say to ourselves, quite sincerely, that we are aiming for moral excellence. We fool ourselves all the time; we convince ourselves of the strangest things, if we want them to be true; we refuse to take a serious critical look at our motives. Aiming is shown less by what we sincerely say than by our patterns of choice.[40] Suppose that littering is bad. To aim for mediocrity with respect to littering is to be disposed to calibrate yourself up or down toward approximately a middle

target. Suppose the wind whips a food wrapper from your hands and carries it down the block; or suppose you accidentally drop an almost-empty bottle of sun lotion from a ski lift. In either case, it would be a hassle to retrieve the thing, but you could do it. If you're calibrating toward the middle, what you do will depend on what you perceive to be standard behavior among people you regard as your peers. If most of them would let it go in a similar situation, so would you. If most of them would chase it down, so would you. It doesn't really matter what story you tell yourself about it. If you're a half-decent moral philosopher, I'm sure that in such moderate grayish cases you can concoct a half-plausible excuse of some sort, regardless of whether you really should go retrieve your litter or not. (If you don't like the littering example, try another: household energy use, helping out someone in a weak position, gray-area expense reporting.)

Second, I don't intend to be making a universal claim about the goodness or badness of aiming for mediocrity or averageness in *all* social contexts. I mean to be talking about *us,* my notional readership. If one's social reference group is the *Einsatzgruppen* killing squads of Nazi-occupied Eastern Europe, aiming for about-average murder and cruelty is not just bad but horrible. If one's social reference group is some saintly future utopia, the mediocre may be perfectly benign. I don't mean them. I just mean us: you, as I imagine you, and me, and our friends and family and neighbors.

#

The Fairness Argument

A certain type of appeal to fairness is one way of defending the idea that it's perfectly fine to aim to be morally mediocre. I will call this the Fairness Argument.

First, assume – of course it's disputable[41] – that being morally excellent normally involves substantial self-sacrifice. It's morally better to donate large amounts to worthy charities than to donate small amounts. It's morally better to be generous rather than stingy with your time in helping colleagues, neighbors, and distant relatives who might not be your favorite people. It's morally better to meet your deadlines rather than inconvenience others by running late. It's morally better to have a small carbon footprint than a medium or large one. It's morally better not to lie, cheat, and fudge in all the small, and sometimes large, ways that people do. To be near the moral maximum in every respect would be practically impossible near-sainthood; but we non-saints could still presumably be somewhat better in many ways. We just choose not to be better, because we'd rather not make the sacrifices involved.

The idea of the Fairness Argument, then, is this: Since most of your peers aren't making the sacrifices necessary for peer-relative moral excellence, it's unfair for you to be blamed for also declining to do so. If the average person in your financial condition gives 3% of their income to charity, it would be unfair to blame you for not giving more. If your colleagues down the hall cheat, shirk, fib, and flake X amount of the time, it's only fair that you get to do the same. Fairness requires that we demand no more than average moral sacrifice from the average person. Thus, there's nothing wrong with aiming to be only a middling member of the moral community – approximately as (un-)selfish, (dis-)honest, and (un-)reliable as everyone else.

#

Two Replies to the Fairness Argument

(1.) *Absolute Standards.* Some actions are morally bad, even if the majority of your peers are doing them. A Nazi death camp guard who is somewhat less cruel to the inmates than average but who is still somewhat more cruel and murderous than he needs to be to keep his job. "Hey, at least I'm better than average!" would be a poor excuse. Closer to home, most people regularly exhibit small to moderate degrees of sexism, racism, ableism, and preferential treatment of the conventionally beautiful.[42] Even though most people do this, we are still criticizable for it. That you're typical or average in your degree of bias is at best a mitigator of blame, not a full excuser from blame. Similarly with failing to live up to your promises: That most people fail to fulfill a significant proportion of their promises makes it, perhaps, forgivable and understandable that you also do so (depending on the promise), but doesn't make it perfectly okay.

(2.) *Strange Tradeoffs.* Most of us see ourselves as having areas of moral strength and weakness. Maybe you're a warm-hearted fellow, but flakier than average about responding to important emails. Maybe you know you tend to be rude and grumpy to strangers, but you're an unusually active volunteer in your community. Empirical evidence suggests that in implicitly guiding our behavior, we tend to treat these tradeoffs as exculpatory or licensing: You cut yourself slack on one in light of the other.[43] You – typically unconsciously or only half-consciously – let your excellence in one area justify lowering your aims in another, so that averaging the two, you come out somewhere in the middle. In aiming for moral mediocrity, we needn't aim for mediocrity by every moral standard considered separately. Such trade-offs between norm types are emotionally tempting when you're motivated to see yourself positively, but making them fully explicit reveals their strangeness: "It's fine that I insulted the cashier, because this afternoon I'm volunteering for river clean-up." "I'm not criticizable for neglecting Cameron's urgent email because this morning I greeted Monica and Britney kindly, filling the office with good

vibes." Stated so baldly, they would probably sound odd even to you. Such tradeoffs wilt under explicit scrutiny.

I'm not imagining here that you are working so hard for a good cause that you simply have no time or energy left over for other good things. That's just human limitation, perfectly excusing. What I'm imagining is the more common phenomenon of letting yourself be a little bad, unkind, or unjustifiably irresponsible about X because you're perhaps a little too proud of yourself for having done Y.

#

Most of us aim for only for moral mediocrity. We ought to own up to this fact. Proper acknowledgement of this possibly uncomfortable fact can then ground a frank confrontation with the question of how much we really care about morality. How much are we willing to sacrifice to be closer to morally excellent, if our neighbors remain off-key?

9. A Theory of Hypocrisy

Hypocrisy, let's say, is when someone conspicuously advocates a moral rule while also secretly (or at least much less conspicuously) violating that moral rule at least as much as does the typical member of their audience.

It's hard to know exactly how common hypocrisy is, because people understandably hide their embarrassing behavior and because the psychology of moral advocacy is a complex and understudied topic. But it seems likely that hypocrisy is more common than a purely strategic analysis of its advantages would predict. I think of "family values" and anti-homosexuality politicians who seem disproportionately likely to be caught in gay affairs.[44] I think of the angry, judgmental people we all know who emphasize how important it is to control your emotions. I think of police officers who break the laws they enforce on others and of environmental activist Al Gore's (formerly?) environmentally-unfriendly personal habits.[45] I think of the former head of the academic integrity office at U.C. Riverside who – I was told when I tried to contact him to ask his views about the moral behavior of students in ethics classes – was fired when his colleagues discovered that he had falsified his resume.

Anti-gay activists might or might not actually be more likely than others to have homosexual affairs, etc. But it's striking to me that the rates even come close, as it seems they do. A purely strategic analysis of hypocrisy suggests that, in general, people who conspicuously condemn X should have low rates of X, since the costs of condemning X while secretly doing X are typically high. Among those costs: contributing to a climate in which X-ish behavior is generally more condemned; attracting friends and allies who are especially likely also to condemn X; attracting extra scrutiny of whether you in fact do X or not; and attracting the charge of hypocrisy, in addition to the charge of X-ing itself, if your X-ing is discovered. It seems strategically foolish for a preacher with a secret gay lover to choose anti-homosexuality as a central theme in his sermons!

Here's what I suspect is going on.

As I suggested in Chapters 4 and 8, people don't generally aim to be saints, nor even much morally better than their neighbors. They aim instead for moral mediocrity. If I see a bunch of people profiting from doing something that I regard as morally wrong, I want to do that thing too. No fair that (say) 15% of people cheat on the test and get As, or regularly get away with underreporting self-employment income. I want to benefit, if they are! This reasoning is tempting even if cheaters are a minority.

Now consider the preacher tempted by homosexuality or the environmentalist who wants to eat steaks in his large air-conditioned house. They might be entirely sincere in their moral opinions. Hypocrisy needn't involve insincere commitment to the moral ideas you espouse. What they see when they look around are many others who are getting away with what they condemn. Seeing these others, and implicitly aiming only for mediocrity, they might feel licensed to indulge themselves a bit too.

Furthermore, the norm violations might be more salient and visible to them than to the average person. The IRS worker sees how frequent and easy it is to cheat on taxes. The anti-homosexuality preacher sees himself in a world full of sinning gays. The environmentalist grumpily notices all the giant SUVs rolling down the road. This increased salience might lead them to overestimate the frequency of such misbehavior – and then when they calibrate toward mediocrity, their scale might be skewed.

Still, this doesn't seem enough to explain the high rate of hypocrisy, given its high costs. Increased salience might lead moral advocates of X to somewhat mistune their estimates of X-violation, but you'd think that they'd try to steer their own behavior on X far down toward the good end of the scale (maybe allowing themselves more laxity on other issues, as a kind of reward).

So here's the final piece of the puzzle: Suppose there's a norm that you find yourself especially tempted to violate. Suppose further that you succeed for a while, at substantial personal cost, in not violating it. You love cheeseburgers but go vegetarian. You have intense homosexual desires but avoid acting on them. Envy might lead you to be especially condemnatory of others who still do such things. If you've worked so hard, they should too! It's an issue you've struggled with personally, so now you have some wisdom about it, you think. You want to make sure that others don't enjoy the sins you've worked so hard to avoid. Furthermore, optimistic self-illusions (excessively positive expectations about yourself) and end-of-history thinking (thinking that your current preferences and habits are unlikely to change) might lead you to overestimate the chance that you will stay strong and not lapse.[46] These envious, self-confident moments are the moments when you are likely to conspicuously condemn the sins to which you are most tempted.

And then you're on the hook for it. If you have been sufficiently conspicuous in your condemnations, it becomes hard to change your tune later, even after you've lapsed.

10. On Not Distinguishing Too Finely Among Your Motivations

I've been reading the book *What's Wrong with Morality?* by the eminent moral psychologist Daniel Batson. Batson distinguishes four types of motives for seemingly moral behavior. Although his taxonomy is conceptually interesting, and although you might think that more distinctions would encourage more precise understanding, I want to push back against Batson on this issue. I don't think it usually make sense to distinguish as finely as Batson does among people's motives for doing good.

Suppose I offer a visiting speaker a ride to the airport. That seems like a nice thing to do. According to Batson, I might have one or more of the following types of motivation:

(1.) I might be *egoistically* motivated – acting in my own perceived self-interest. Maybe the speaker is the editor of a prestigious journal and I think I'll have a better chance of publishing and advancing my career if she thinks well of me.

(2.) I might be *altruistically* motivated – aiming primarily to benefit the visitor herself. I just want her to have a good time, a good experience at my school. Giving her a ride is a way of advancing that goal I have. (To see that altruistic motivations aren't always moral, consider altruistically benefiting one person by unfairly harming many others.)

(3.) I might be *collectivistically* motivated – aiming primarily to benefit a group. I want my school's philosophy department to flourish, and giving the speaker a ride is a way of advancing that thing I care about.

(4.) I might be *motivated by principle* – acting according to a moral standard, principle, or ideal. Maybe I think that driving the speaker to the airport will maximize global utility, or that it is ethically required given my social role and past promises.

Batson characterizes his view of motivation as "Galilean" – focused on the underlying forces that drive behavior.[47] His idea seems to be that when I make that offer to the visiting speaker, my action must have been impelled by some particular motivational force inside of me that is egoistic, altruistic, collectivist, or principled, or some specific blend of those four. On this view, we don't understand why I am offering the ride until we know which of these interior forces is the one that caused me to offer the ride. Principled morality is rare, Batson argues, because it requires being caused to act by the fourth type of motivation, and people are more commonly driven by the first three.

I'm nervous about appeals to internal causes of this sort. My best guess is that these sorts of simple, familiar (or semi-familiar) categories won't tend to map well onto the real causal processes that generate our behavior – processes that are likely to be very complicated and to be misaligned with the categories that come easily to mind. In AI and cognitive science, consider connectionist structures and deep learning systems, like AlphaGo and Facebook's picture categorization algorithms, which often succeed in mimicking human skills via a complex network of substructures and subclassifications that make little intuitive sense to us.[48] The brain might be like that. Four types of motivational causes might be far too few and too neat.

With this background picture in place, let me suggest this: It's plausible that our motives are often a tangled mess; and when they are a tangled mess, attempting to distinguish too finely among them is a mistake.

For example, there are probably hypothetical conditions under which I would decline to drive the speaker because it conflicted with my self-interest, and there are probably other hypothetical conditions under which I would set aside my self-interest and drive the speaker anyway. I doubt that these hypothetical conditions line up neatly so that I would decline to drive the speaker if and only if it would require sacrificing X amount or more of self-interest.

I'm not as coherent and rational as that. There's habit, association, subtle situational pressures I'm not aware of but which I would repudiate on reflection, other related motives that I weigh against each other inconsistently, some smell in the wind that makes a consideration salient that otherwise wouldn't occur to me, or triggers some specific memory or sentiment. Some situations might channel me into driving her, even at substantial personal cost, while others might more easily invite the temptation to wiggle out, for no good reason.

Similarly for other motivations. Hypothetically, if the situation were different so that it was less in the collective interest of the department, or less in the speaker's interest, or less strongly compelled by my favorite moral principles, I might drive or not drive the speaker depending partly on each of these but also partly on other factors of the situation and my internal psychology – habits, scripts, potential embarrassment, moods and memories that happen to bubble up – probably in no tidy pattern.

Furthermore, egoistic, altruistic, collectivist, and principled aims come in many varieties, difficult to straighten out. I might be egoistically interested in the collective flourishing of my department as a way of enhancing my own stature. I might relish displaying the sights of the L.A. basin through the windows of my car, with a feeling of civic pride. I might be drawn to different, conflicting moral principles. I might altruistically desire that the speaker enjoy the company of the cleverest conversationalist in the department, which I self-deceptively believe to be myself because it flatters my ego to think so.

Among all of these possible motives – infinitely many possible motives, perhaps, depending on how finely we slice the candidates – does it make sense to seek the one or few "real" motives that are genuinely causally responsible for my choice?

Now if my actual and counterfactual choices – what I actually tend to do as well as what I would do in various hypothetical circumstances – were all neatly aligned with my perceived self-interest, then of course self-interest would be my real motive. Similarly, if my

pattern of actual and counterfactual choices were all neatly aligned with one particular moral principle, then we could say I was mainly moved by that principle. But if my dispositions aren't so neatly arranged, if my patterns of choice comprise a crazy-spaghetti tangle, then each of Batson's four factors is only an approximate and simplified label, rather than a deep Galilean cause of my decision.

Furthermore, the four factors might not compete with each other as starkly as Batson supposes. Each of them might, to a first approximation, capture my motivation reasonably well, in those fortunate cases where self-interest, other-interest, collective interest, and moral principle all tend to align. I have lots of reasons for driving the speaker! This might be so even if, in more remote hypothetical cases, I diverge from the predicted patterns, probably in different and complex ways.

If we accept a modest, folksy vision of morality (instead of a demanding one that requires us to abandon our families to follow Jesus, or to dedicate almost all of our time and money to fighting poverty, or etc.), and if we enjoy fairly fortunate middle-class lives in a well-structured society, then in many of our daily choices, Batson's four types of motives will align. Why take my daughter to the park on Sunday? I'll enjoy it; she'll enjoy it; it will help the family overall; and it's part of my duty as a father to do such things from time to time. Why show up for the boring faculty meeting? It's in my interest to be seen as reliable by my colleagues; and I can help advance (my vision of) what's good for the department; and my colleagues, who I care about, are relying on me; and it's my obligation as a responsible faculty member. Only when there's a conflict do we need to weigh such considerations against each other; and even then we only need to weigh them well enough for practical purposes, across a narrow range of foreseen possibilities.

(Wait, I hear you saying, what about the inevitable tradeoffs I insist on in Chapter 3, where I criticize Happy Coincidence reasoning? Answer: Even in such tradeoffs, there will

tend to be egoistic, altruistic, collectivist, and principled motives mixing together on each side of the tradeoff. For example, my attending the fancy conference feeds my ego as a philosopher of rising prominence, contributes to the collective flourishing of research in my subfield, advances the interests of my friends who invited me, and fulfills a promise I made earlier.)

My motivations might be described, with approximately equal accuracy, as egoistic, altruistic, collectivist, *and* principled, in different flavors and subtypes, when these considerations align across the relevant range of situations. This isn't because each type of motivation contributes equal causal juice to my behavior but rather because each attribution captures well enough the pattern of choices I would make in the range of likely cases we care about.

Batson seems to want what many of us want when we ask, skeptically or scientifically, *what was her real motive?* He wants a single clean answer, or maybe a neatly quantifiable mix of 70% this and 30% that. Such tidiness, however, is probably more the exception than the rule.

11. The Mush of Normativity

Recently, several psychologists, most prominently Jonathan Haidt, have emphasized the connection between disgust and moral condemnation. Evidence suggests (though there are some concerns about replicability), that inducing disgust – whether by hypnosis, rubbish, or fart spray – tends to increase the severity of people's moral condemnation.[49] Conversely, pleasant odors might have a positive effect: For example, shopping-mall passersby might be more likely to agree to break a dollar when approached near a pleasant-smelling bakery than when approached near a neutral-smelling dry goods store.[50] Similarly, people tend to find the idea of sex with a frozen chicken revolting, and so they morally condemn it, even if often they can find no rational basis for that condemnation.[51]

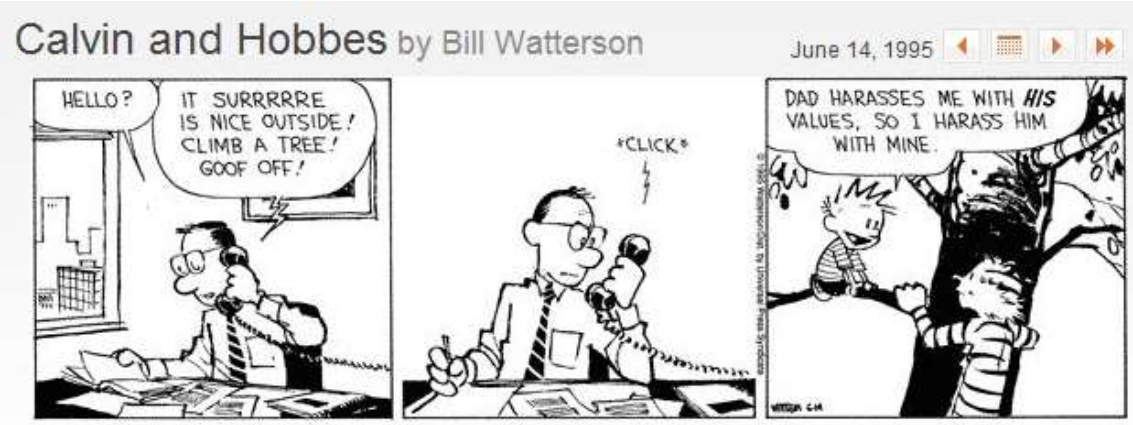
One possible interpretation of these results retains the idea that we have a distinct cognitive system for moral judgment, different from that for aesthetic judgment, but allows that the moral system's outputs can be influenced by factors like odor and sexual disgust. Moral judgments, according to most philosophers and moral psychologists, are one thing, and aesthetic judgments are another, even if the one influences the other. Philosophers have long recognized several different and distinct species of "norms" or types of evaluation: moral norms, aesthetic norms, prudential norms of self-interest, and epistemic norms concerning what to believe or accept. Sometimes it's argued that these different standards of evaluation constrain each other in some way. Maybe the immoral can never truly be beautiful, or maybe people have a moral obligation to be epistemically rational. But even if one type of "normativity" completely subsumes another, wherever there is a multiplicity of norm types, philosophers typically treat these norms as sharply conceptually distinct.

A different possibility is that norms or modes of evaluation mush together, not just causally (with aesthetic judgments affecting moral judgments, as a matter of psychological fact), and not just via philosophically-discovered putative contingencies (such as that the

immoral is never beautiful), but into a blurry mess that defies neat sorting. You might make an evaluative judgment and recognize a normative fact, while the type of normativity involved remains indistinct.

I'm not merely suggesting that normative judgments can be multi-faceted. If we accept the gemstone analogy implicit in the word "facet", facets are by nature distinct. If normativity has facets, different types of normativity might be sharply distinct, and yet a particular evaluative judgment might have more than one normative dimension, for example both an epistemic and a moral dimension. Racism might be epistemically irrational *and* immoral *and* ugly, with each of these constituting a distinct facet of its badness. My thought isn't that. Instead, my thought is that many evaluative judgments, and perhaps also many normative facts, aren't so sharply structured.

Consider this strip from Calvin and Hobbes:



(Calvin's father answers the phone at work. Calvin says, "It surrrrre is nice outside! Climb a tree! Goof off!" In the last panel, Calvin says to Hobbes, "Dad harasses me with *his* values, so I harass him with mine."[52])

Calvin's exhortation is normative. He explicitly says so in the closing panel. It's about values. So, is Calvin urging a set of *moral* values on his father? Is he an ethical Daoist of some sort, who thinks it's wrong to waste one's precious life struggling for money and accomplishment? Or rather is it that Calvin sees prudential, self-interested value in climbing

trees, and he wishes his father would recognize climbing trees to be in his self-interest too? Or is Calvin urging an aesthetic worldview, centered on properly appreciating the beauty of nature? Psychologically, I don't know that there need be some particular mix, in Calvin, of moral vs. prudential vs. aesthetic dimensions in this evaluative judgment. Must there be a fine-grained fact of the matter? Perhaps they tangle and twist together in ways that Calvin couldn't articulate, even with a philosopher's or psychologist's help, and what's beneath isn't fully stable and coherent.

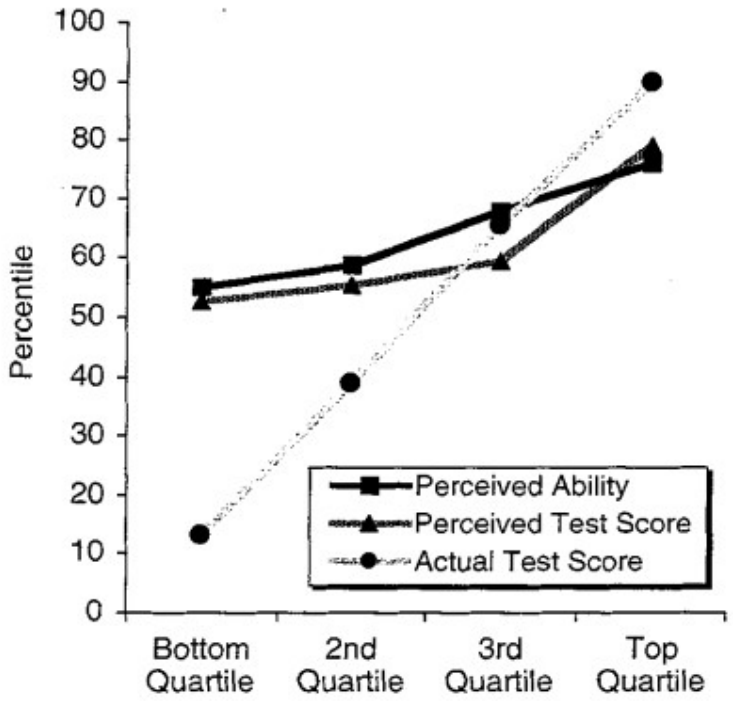
Let's suppose, further, that Calvin is right. His father *should* climb a tree and goof off. What kind of "should" or what blend of shouldishness is at issue? Need it be that the normativity is X% moral and Y% prudential? Or prudential rather than moral? Or definitely both prudential and moral but for distinct metaphysical reasons?

My core thought is this: The psychology of normative evaluation might be a mushy mess of attractions and repulsions, of pro and con attitudes, not well characterized by sharp distinctions among philosophers' several types of normativity. And if normative facts are partly grounded in such psychological facts, as many philosophers think they are – if the aesthetic and moral and epistemic and prudential value are partly based on what people are disposed to aesthetically, morally, epistemically, and prudentially praise and condemn[53] – then the normative facts themselves might inherit this psychological mushiness.

12. A Moral Dunning-Kruger Effect?

In a famous series of experiments, Justin Kruger and David Dunning found that people who scored in the lowest quartile of skill in grammar, logic, and (yes, they tried to measure this) humor tended to substantially overestimate their abilities, rating themselves as a bit above average in these skills.[54] In contrast, people in the top half of ability had more accurate estimations (even tending to underestimate a bit). In each quartile of skill, the average participant rated themselves as above average; and overall, the correlation between self-rated skill and measured skill was small.

For example, here's Kruger and Dunning's chart for the relation between self-rated logic ability and scores on a logic test:



Kruger and Dunning's explanation is that poor skill at (say) logical reasoning not only impairs one's performance at logical reasoning tasks but also impairs one's ability to evaluate one's performance at logical reasoning tasks. You need to know that affirming the

consequent is a logical error to realize that you've just committed a logical error in affirming the consequent. Otherwise, you're likely to think, "P implies Q, and Q is true, so P must also be true. Right! Hey, I'm doing great!"

Although popular presentations of the Dunning-Kruger Effect tend to generalize it to all skill domains, it seems unlikely that it does generalize universally. In domains where evaluating one's success doesn't depend on the skill in question, and instead depends on simpler forms of observation and feedback, one might expect more accurate self-assessments.[55] For example: footraces. I'd wager that people who are slow runners don't tend to think that they're above average in running speed. They might not have perfect self-knowledge; they might show some self-enhancing optimistic bias,[56] but I doubt we'd see the almost flat line characteristic of Dunning-Kruger. You don't have to be a fast runner to notice that your friends can outrun you.

So… what about ethics? Ought we to expect a moral Dunning-Kruger effect?

My hunch is yes. Evaluating your own moral or immoral behavior is a skill that itself depends on your moral abilities. The least moral people are typically also the least capable of recognizing what counts as a moral violation and how serious the violation is – especially, perhaps, when considering their own actions. I don't want to overcommit on this point. Surely there are exceptions. But as a general trend, this seems plausible.

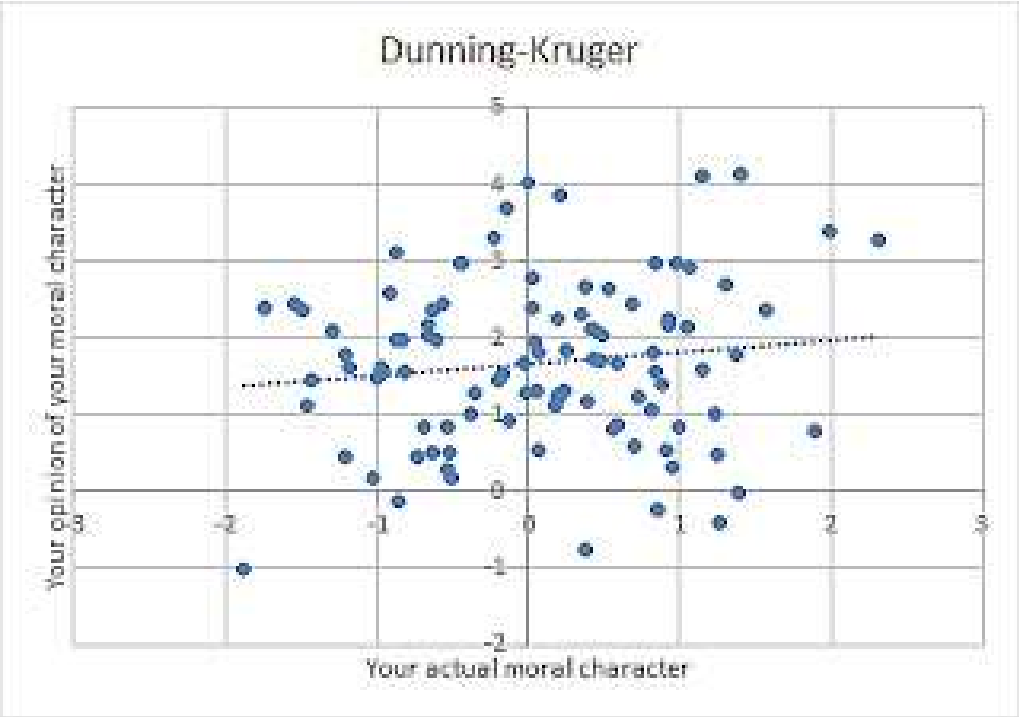
Consider sexism. The most sexist people tend to be the least capable of understanding what constitutes sexist behavior and what makes sexist behavior unethical. They will tend either to regard themselves as not sexist or to regard themselves only as "sexist" in a non-pejorative sense. ("So what, I'm a 'sexist'. I think men and women are different. If you don't, you're a fool.") Similarly, the most habitual liars might not see anything bad in lying, or not even think of what they are doing as "lying". (Maybe it's just

exaggerating or selling or spinning.) They might tend to assume that almost everyone avoids the truth when convenient.

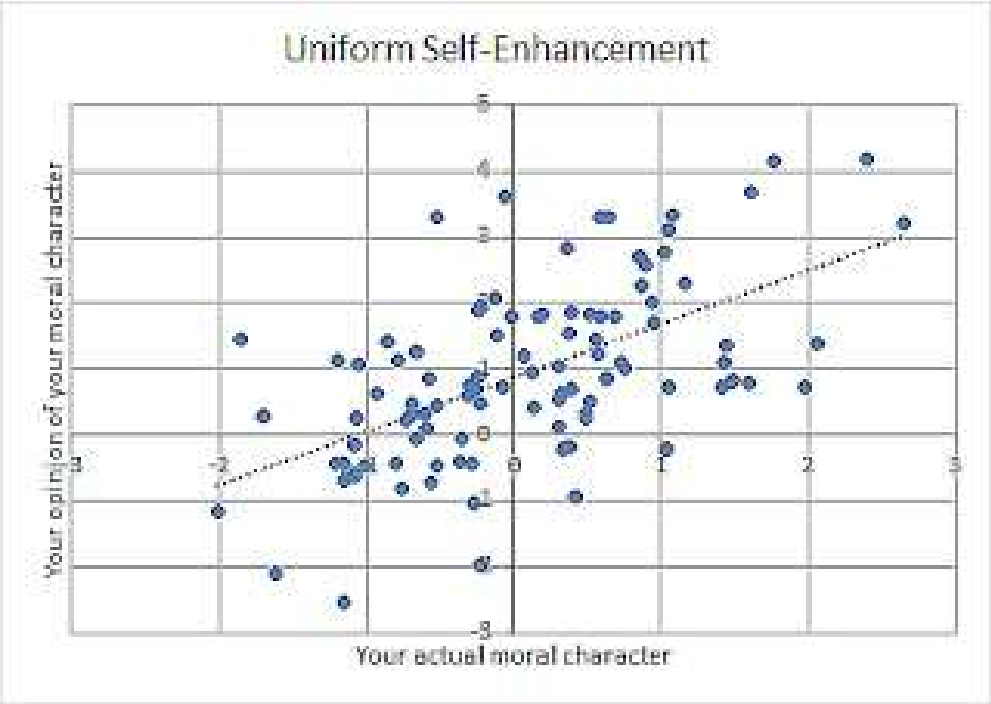
It probably doesn't make sense to think that overall morality can be precisely captured in a unidimensional scale – just like it probably doesn't make sense to think that there's one correct unidimensional scale for skill at basketball or for skill as a philosopher or for being a good parent. And yet, clearly some ball players, philosophers, and parents are better than others. There are great, good, mediocre, and crummy versions of each. As a first approximation, I think it's probably okay to think that there are more and less ethical people overall. And if so, we can imagine a rough scale.

With that important caveat, then, consider the following possible relationships between one's overall moral character and one's opinion about one's moral character:

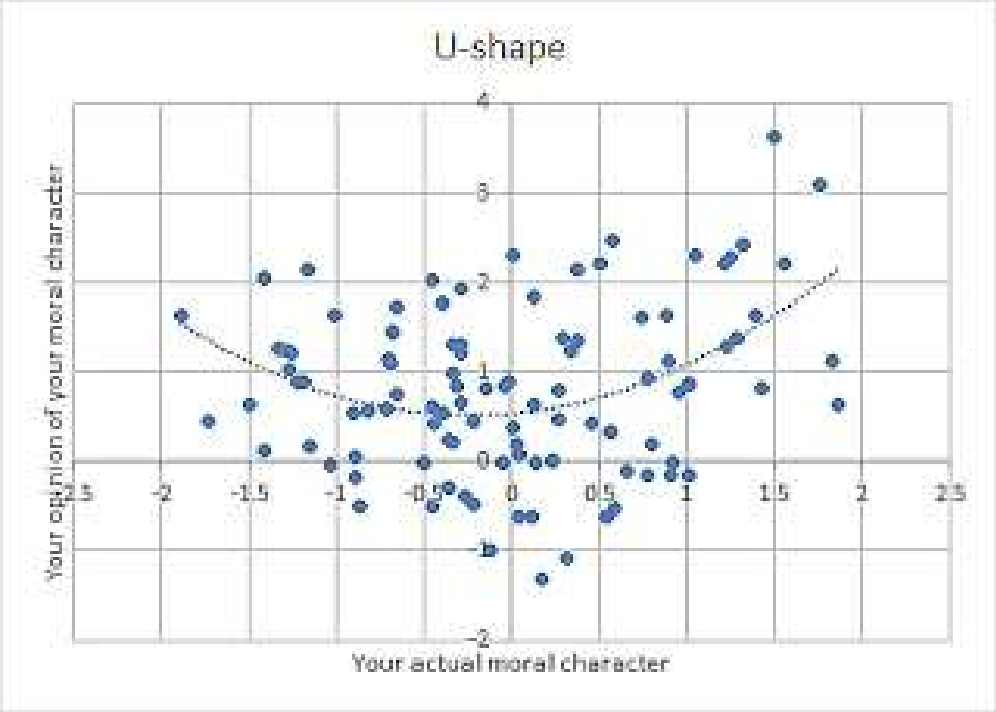
Dunning-Kruger (more self-enhancement for lower moral character):[57]



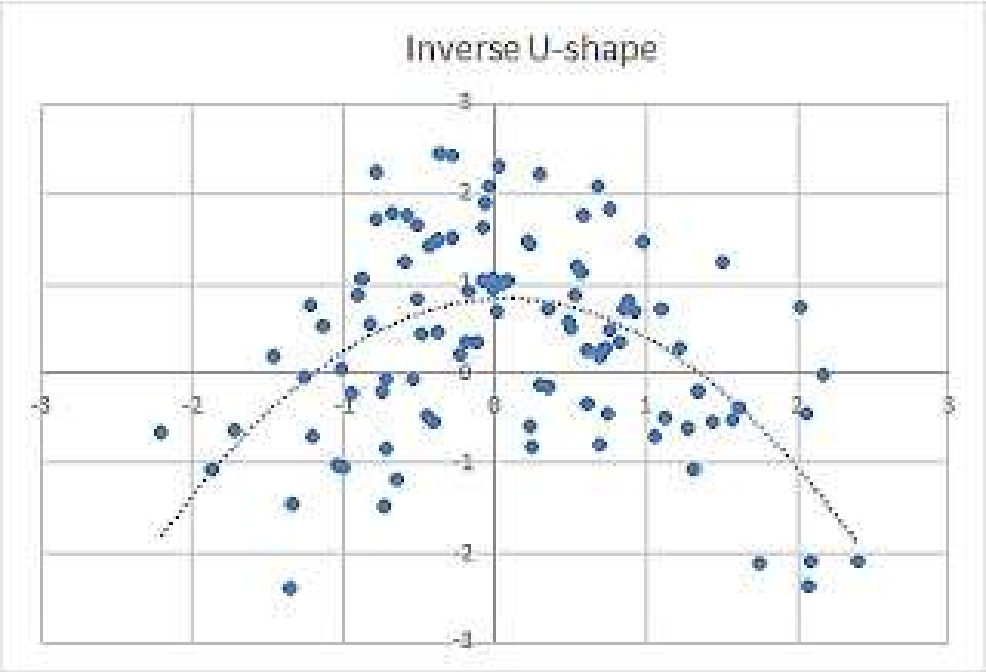
Uniform self-enhancement (everyone tends to think they're a bit better than they are):



U-shaped curve (even more self-enhancement among people who are below average):



Inverse U (realistically low self-image for the worst, self-enhancement in the middle, and self-underestimation for the best):



I don't think we really know yet which of these models is closest to the truth, but in all of them I have included two elements: substantial self-enhancement (a greater than zero average opinion about one's own moral character) and high scatter (at best a weak correlation between one's opinion and one's real moral character). Both elements are well attested in the empirical literature (and, I think, by ordinary observation).[58]

Suppose, then, that in your opinion your moral character is somewhat above average – pretty good but not saintly, B+-ish, maybe 1.5 on the vertical axis. On any of the charts above, your actual moral character could be more or less anywhere along the horizontal axis – unless, of course, your opinion about your moral character is somehow more securely grounded than others'?

13. The Moral Compass and the Liberal Ideal in Moral Education

Consider these two very different approaches to children's moral education:

The outward-in approach. Inform the child what the rules are. Don't expect the child to like the rules or to regard them as wise. Instead, enforce compliance through punishment and reward. Secondarily, explain the rules, in the hope that eventually the child will come to appreciate their wisdom, internalize them, and be willing to abide by them without threat of punishment.[59]

The inward-out approach. When the child does something wrong, help the child see for herself what makes it wrong. Invite the child to reflect on what constitutes a good system of rules. Invite her to think about how people should treat each other. Collaborate with the child to develop guidelines and ideals that she can accept as wise. Trust that even young children can see the value of moral guidelines and ideals. Punish only as a fallback when more collaborative approaches fail.[60]

Although there need be no clean mapping, I suspect that preference for the outward-in approach correlates with political conservatism and the inward-out approach correlates with political liberalism. The crucial difference is that the outward-in approach trusts children's judgment less. On the outward-in approach, children should be taught to defer to established rules, even if those rules don't make sense to them. Applied to adults, this resembles Burkean political conservativism, which prioritizes respect for the functioning of our historically established traditions and institutions, mistrusting our current judgments about how those traditions and institutions might be improved or replaced.

In contrast, the liberal ideal in moral education depends on the thought that most or all people – including most or all children – have something like an inner moral compass that

can be relied on as at least a partial, imperfect moral guide. If you pull aside four-year-old Pooja[61] after she has punched Lauren and patiently ask her to explain herself and to reflect on the ethics of punching, you should get something sensible in reply. For this liberal ideal to work, it must be true that Pooja can be brought to understand the importance of treating others kindly and fairly. It must be true that, after reflection, she will usually find that she wants to be kind and fair to others, even with no outer reward.

As part of this inward-out collaboration, you must be ready also to genuinely listen to the child. Otherwise it is not a collaboration, but outward-in imposition in disguise. Eventually, they will see through the disguise.

Moral wisdom is a lot to expect of four-year-olds. And yet I do think that most children, when approached patiently, can find it. In my experience watching parents and educators, it strikes me that when adults are at their best – not overloaded with stress or too many students – they can successfully use the inward-out approach. Empirical psychology, as I read it, suggests that the (imperfect, undeveloped) seeds of morality are present early in development and shared among primates.[62] The inward-out approach relies on these seeds and nurtures them.[63]

We can extend this thinking to adults too – when we aim to "educate" people we judge to be morally mistaken, when we hope to win over people with very different moral visions of the world that our own, or people who seem to us to think only rarely or badly about the real moral consequences of their opinions and actions. In this case, even more so than in the case of children, it seems wise to adopt a collaborative, inward-out approach. It might be preferable on broad empirical grounds, and it might furthermore be required of us as a matter of respect.

The liberal conception of the human condition – "liberal" in rejecting the top-down imposition of values and celebrating instead people's discovery of their own values – is

founded on the hope that when people are given a chance to reflect, in conditions of peace, with broad access to relevant information, they will tend to find themselves revolted by evil and attracted to good. Hatred and evil will wither under thoughtful critical examination. Despite complexities, bumps, regressions, and contrary forces, introspection and broad exposure to facts and arguments will bend us toward freedom, egalitarianism, and mutual respect.

Sometimes I find this faith hard to maintain. I read widely on the history of evil, and I teach a class every year on racial lynching and the Holocaust. I study the excuses that people historically have made for the horrible things they've done. In the chapters above, I have expressed doubts about people's ability to think clearly about their own morality. And yet something keeps bringing me back to the inward-out view and liberal ideal of moral education. Somewhere beneath the noise and self-deception, we can find our moral compass.

If this view is correct, here's something you can always do in the face of hate, chaos, and evil: Invite people to think alongside you. Treat them as collaborative partners in thought. Share the knowledge you have, and listen to them in return. If there is light and insight in your thinking, people will slowly walk toward it. And maybe you too will walk toward their light. Maybe you will find yourselves walking together, in a direction somewhat different than you would have predicted.

Part Two: Cute AI and Zombie Robots

14. Should Your Driverless Car Kill You So Others May Live?

It's 2030. You and your daughter are riding in a driverless car along the Pacific Coast Highway. The autonomous vehicle rounds a corner and detects a crosswalk full of children. It brakes, but your lane is unexpectedly full of sand from a recent rock slide. Your car can't get traction. Its AI does some calculations: If it continues braking, there's a 90% chance that it will kill at least three children. Should it save them by steering you and your daughter off the cliff?

This isn't an idle thought experiment. Driverless cars will be programmed to avoid colliding with pedestrians and other vehicles. They will also be programmed to protect the safety of their passengers. What should happen in an emergency when these two aims conflict?

Regulatory agencies around the country have been exploring safety rules for autonomous vehicles. The new rules might or might not clarify when it is acceptable for collision-avoidance programs to risk passengers to avoid harming others. The new rules might or might not clarify general principles of risk tradeoff, might or might not clarify specific guidelines like when it's permissible, or required, to cross a double-yellow line in a risky situation, what types of maneuvers on ice should be allowed or disallowed.

Technology companies have been arguing for minimal regulation. Google, for example, has proposed that manufactures not be held to specific functional safety standards and instead be allowed to "self-certify" the safety of their vehicles, with substantial freedom to develop collision-avoidance algorithms as they see fit. But this let-the-manufactures-decide attitude risks creating a market of excessively self-protective cars.

Consider some boundary cases. Some safety algorithms seem far too selfish. Protecting passenger safety at all costs, for example, is overly simple and would be morally odious to implement. On such a rule, if the car calculates that the only way to avoid killing a pedestrian would be to side-swipe a parked truck, with a 5% chance of minor injury to the passengers, then the car should kill the pedestrian. On the other hand, a simple utilitarian rule of maximizing lives saved disregards personal accountability, and is too sacrificial in some cases, since it doesn't take into account that others might have irresponsibly put themselves in danger. If you come head-on with a reckless motorcyclist speeding around a sharp curve, it's reasonable for your car to prioritize your safety over the biker's.

People might tend to prefer that their own cars be highly self-protective, while others' cars are more evenhandedly neutral. They might want others' cars to sacrifice the two passengers rather than kill the three kids on the sidewalk, but when choosing what to buy or rent for themselves and their own children, they might want passenger protection to be given top priority.[64] If everyone chooses the safety algorithms only for the cars they ride in, with no regulation of others' algorithms, cars as a whole might end up far more selfishly protective of their passengers than we as a society would collectively prefer.

We face a social coordination problem on a matter of huge moral importance. Who gets priority in an accident? How much selfishness is acceptable in passenger risk-reduction? We can't just leave it to market forces and manufacturers' secret design choices. We should throw the algorithms to light, scrutinize them openly, and after public debate, draw some regulatory parameters concerning acceptable tradeoffs between passenger safety and risk to others.

Completely uniform safety protocols might not be ideal either, however. Some consumer freedom seems desirable. To require that all vehicles at all times employ the same set of collision-avoidance procedures would needlessly deprive people of the chance to

choose algorithms or driving styles that reflect their values. Some people might wish to prioritize the safety of their children over themselves. Others might prefer to prioritize all passengers equally. Some people might elect algorithms that are more self-sacrificial on behalf of strangers than the government could legitimately require of its citizens. There will also always be tradeoffs between speed and safety, and different riders might legitimately evaluate the tradeoffs differently, on different occasions, as we now do in our manual driving choices.

If cars can be programmed for a range of driving styles, the selected styles might be broadcast to others. A car in "max aggressive" mode might visibly crouch low and display a tight posture, or a certain pattern of lights on its roof, signaling to other cars, and to pedestrians, that it will be driving as quickly and aggressively as permitted by law. A car in "max safety" mode might display a different posture or different pattern of roof lights, indicating to others that it will be traveling more cautiously. A "baby on board" setting might signal that backseat passengers will be prioritized in an emergency – maybe even signaling to other cars' AI systems that they should hit the front rather than the back if there's a choice. A "utilitarian" mode, valuing all lives equally, might earn moral praise from your neighbors.

We should also insist on passengers' prerogative to pre-emptively override their autonomous systems – not only so that we can drive according to our values, but also because, for the foreseeable future, there will be situations in which human cognition can be expected to outperform computer cognition. Properly licensed passengers ought to be free to take control when that is likely. Although computers normally have faster reaction times than people, out best computer programs still lag far behind normal human vision at detecting objects in novel, cluttered environments. On a windy fall day, a woman might be pushing a coat rack across the street in a swirl of leaves. Without manual override, the computer might sacrifice you for a mirage.

There is something romantic about the hand upon the wheel – about the responsibility it implies. But it's not just romanticism to resist ceding this responsibility too quickly, and without sufficient oversight, to the engineers at Google.

Future generations might be amazed that we allowed music-blasting 16-year-olds to pilot vehicles unsupervised at 65 miles per hour, with a flick of the steering wheel the difference between life and death. A well-designed machine will probably do better in the long run. That machine will never drive drunk, never look away from the road to change the radio station or yell at the kids in the back seat. It will, however, have the power over life and death. We need to decide – publicly – how it will exert that power.[65]

15. Cute AI and the ASIMO Problem

A few years ago, I saw the ASIMO show at Disneyland. ASIMO is a robot designed by Honda to walk bipedally with something like a human gait. I'd entered the auditorium with a somewhat negative attitude about ASIMO, having read Andy Clark's critique[66] of Honda's computationally-heavy approach to locomotion, and the animatronic Mr. Lincoln on Disneyland's Main Street had left me cold.

But ASIMO is cute! He's about four feet tall, humanoid, with big round dark eyes inside what looks like an astronaut's helmet. He talks, he dances, he kicks soccer balls, he makes funny hand gestures. On the Disneyland stage, he keeps up a fun patter with a human actor. ASIMO's gait isn't quite human, but his nervous-looking crouching run only makes him that much cuter. By the end of the show I thought that if you asked me to take him apart against his protests, I'd be reluctant to comply.

ASIMO the cute robot. Image from http://asimo.honda.com. Used with permission.

Another case: ELIZA was a simple chatbot written in the 1960s, drawing on a small template of pre-programmed responses to imitate a non-directive psychotherapist. ("Are such questions on your mind often?" "Tell me more about your mother.") Apparently, some early users mistook it for a human and spent long hours chatting with it.[67]

In more recent research, Kate Darling finds that participants are often reluctant to strike a robotic toy with a mallet, especially when she introduces it using anthropomorphic language.[68] In my undergraduate class on philosophy of mind, on the day we discuss the minds of non-human animals, I normally bring a large, cute stuffed teddy bear to class. I treat it affectionately at the start, stroking its head and calling it by endearing names. Then, about halfway through the class I suddenly punch it in the face. Students scream in shock. Sometimes I hear, years later, that the thing they most vividly remembered about my course was (unfortunately) that the professor punched a stuffed bear in the face.

Relatedly, evidence from developmental and social psychology suggests that people are generally swift to attribute mental states to entities with eyes and movement patterns that look goal directed, even if in other ways the entities are plainly not sophisticated.[69]

I assume that ASIMO and ELIZA are too simple to be proper targets of substantial moral concern.[70] They have no more consciousness than a laptop computer, and no more capacity for genuine joy or suffering. But they at least *tempt* us to treat them with moral regard. And future engineers could presumably create entities with an even better repertoire of superficial tricks. Discussing these issues with my sister, she mentioned a friend who had been designing a laptop that would whine and cry when its battery ran low. Imagine that! "Oh please *please* update me! I'm scared of those new viruses that everyone's saying are out there."

Conversely, suppose that it's someday possible to create an AI system so advanced that it really does have genuine consciousness, a genuine sense of self, real joy, and real

suffering. If that AI also happens to be ugly or boxy or poorly interfaced, it might tend to attract much less moral concern than is warranted. In the famous *Star Trek* episode "The Measure of a Man", a scientist who wants to disassemble the humanoid robot Data (sympathetically portrayed by a human actor) says of the robot, "If it were a box on wheels, I would not be facing this opposition."[71] He also points out that people normally think nothing of upgrading the computer systems of a starship, though that means discarding a highly intelligent AI.

As AI continues to improve, our emotional responses to AIs might become radically misaligned with the AIs' real moral status. If a cute ASIMO full of superficial sympathy-arousing tricks and a flesh-and-blood human being both fall in the water at the same time, so that we can only save one, a well-intentioned rescuer might dive in to save the mindless ASIMO while letting the real human drown. Conversely, we might create a host of suffering AI slaves whose real welfare interests we ignore because they don't give us the right interface cues. In AI cases, the superficial features might not track underlying mentality very well at all.

Call this the *ASIMO Problem*.

I draw two main lessons from the ASIMO Problem.

First is a methodological lesson: In thinking about the moral status of AI, we should be careful not to overweight emotional reactions and intuitive judgments that might be driven by such superficial features.

The second lesson is a bit of AI design advice. As responsible creators of artificial entities, we should want people neither to over- nor under-attribute moral status to the entities with which they interact. If our AIs are simple and non-conscious, we should generally try to avoid designing them in a way that will lead normal users to give them substantial undeserved moral consideration.[72] This might be especially important in designing children's

toys and artificial companions for the lonely. Manufacturers might understandably be tempted to create artificial pets, friends, or helpers that children and others will love and attach to – but we should be cautious about future children overly attaching mostly to non-conscious toys in preference to real people. Although we might treasure a toy as an artifact of great sentimental value, we don't want to lose hold of the fact that is only an artifact – something that needs to be left behind in a fire, even if it seems to plead desperately in mock pain.

On the flip side, if we do someday create genuinely human-grade AIs who merit substantial moral concern, it is crucial that we design their interface and superficial features in a way that will evoke the proper range of moral responses from normal users.

We should embrace an *Emotional Alignment Design Policy:* Design the superficial features of AIs so as to evoke the moral emotional reactions appropriate to the real moral status of the AI, whatever that status is, neither more nor less.[73]

16. My Daughter's Rented Eyes

At two million dollars outright, of course I couldn't afford to *buy* eyes for my four-year-old daughter Eva. So, like everyone else whose kids had been blinded by the GuGuBoo Toy Company's defective dolls (may its executives rot in bankruptcy Hell), I rented the eyes. What else could I possibly do?

Unlike some parents, I actually read the Eye & Ear Company's rental contract. So I knew part of what we were in for. If we didn't make the monthly payments, her eyes would shut off. We agreed to binding arbitration. We agreed to a debt-priority clause, to financial liability for eye extraction, to automatic updates. We agreed that from time to time the saccade patterns of her eyes would be subtly adjusted so that her gaze would linger over advertisements from companies that partnered with Eye & Ear Co. We agreed that in the supermarket, Eva's eyes would be gently maneuvered toward the Froot Loops and the M&Ms.

When the updates came in, we had the legal right to refuse them. We could, hypothetically, turn off Eva's eyes, then have them surgically removed and returned to Eye & Ear. Each new rental contract was thus technically voluntary.

When Eva was seven, the new updater threatened shutoff unless we transferred $1000 to a debit account. Her updated eyes contained new software to detect any copyrighted text or images she might see. Instead of buying copyrighted works in the usual way, we agreed to have a small fee deducted from the debit account for each work Eva viewed. Instead of paying $4.99 for the digital copy of a Dr. Seuss book, Eye & Ear would deduct $0.50 each time she read the book. Video games might be free with ads, or $0.05 per play, or $0.10, or even $1.00. Since our finances were tight we set up parental controls: Eva's eyes required parent permission for any charge over $0.99 or any cumulative charges over $5.00 in a day – and of course they blocked any "adult" material. Until we granted approval, blocked or

unpaid material was blurred and indecipherable, even if she was just peeking over someone's shoulder at a book or walking past a television in a dentist's lobby.

When Eva was ten, the updater overlaid advertisements in her visual field. It helped keep rental costs down. (We could have bought out of the ads for an extra $6000 a year.) The ads never interfered with Eva's vision – they just kind of scrolled across the top of her visual field sometimes, Eva said, or printed themselves onto clouds or the sides of buildings.

By the time Eva was thirteen, I'd risen to a managerial position at work and we could afford the new luxury eyes for her. By adjusting the settings, Eva could see infrared at night. She could zoom in on distant objects. She could bug out her eyes and point them in different directions like some sort of weird animal, taking in a broader field of view. She could take snapshots and later retrieve them with a subvocalization – which gave her a great advantage at school over her normal-eyed and cheaper-eyed peers. Installed software could text-search through stored snapshots, solve mathematical equations, and pull relevant information from the internet. When teachers tried to ban such enhancements in the classroom, Eye & Ear fought back, arguing that the technology had become so integral to the children's way of thinking and acting that it couldn't be removed without disabling them. Eye & Ear refused to develop the technology to turn off the enhancement features, and no teacher could realistically prevent a kid from blinking and subvocalizing.

By the time Eva was seventeen it looked like she and two of her other schoolmates with luxury eye rentals would be choosing among offers from several elite universities. I refused to believe the rumors about parents intentionally blinding their children so that they too could rent eyes.

When Eva was twenty, all the updates – not just the cheap ones – required that you accept the "Acceleration" technology. Companies contracted with Eye & Ear to privilege their messages and materials for faster visual processing. Pepsi paid eighty million dollars so

that users' eyes would prioritize resolving Pepsi cans and Pepsi symbols in the visual scene. Coca Cola cans and symbols were "deprioritized" and stayed blurry unless you focused on them for a few seconds. Loading stored images worked similarly. A remembered scene with a Pepsi bottle in it would load almost instantly. One with a Coke bottle would take longer and might start out fuzzy or fragmented.

Eye & Ear started to make glasses for the rest of us, which imitated some of the functions of the implants. Of course they were incredibly useful. Who wouldn't want to take snapshots, see in the dark, zoom into the distance, get internet search and tagging? We all rented whatever versions we could afford, signed the annual terms and conditions, received the updates. We wore them pretty much all day, even in the shower. The glasses beeped alarmingly whenever you took them off, unless you went through a complex shutdown sequence.

When the "Johnson for President" campaign bought an Acceleration, the issue went all the way to the Supreme Court. Johnson's campaign had paid Eye & Ear to prioritize the perception of his face and deprioritize the perception of his opponent's face, prioritize the visual resolution and recall of his ads, deprioritize the resolution and recall of his opponent's ads. Eva was by now a high-powered lawyer in a New York firm, on the fast track toward partner. She worked for the Johnson campaign, though I wouldn't have thought it was like her. Johnson was so shrill and angry – or at least it seemed so to me when I took my glasses off.

Johnson favored immigration restrictions, and his opponent claimed (but never proved) that Eye & Ear implemented an algorithm that exaggerated people's differences in skin tone – making the lights a little lighter, the darks and little darker, the East Asians a bit yellow. Johnson won narrowly, before his opponent's suit about the Acceleration made it through the appeals process. It didn't hit the high court until a month after inauguration. Eva

helped prepare Johnson's defense. Eight of the nine justices were over eighty years old. They lived stretched lives with enhanced longevity and of course all the best implants. They heard the case through the very best ears.[74]

17. Someday, Your Employer Will Technologically Control Your Moods

Here's the argument:

(1.) Someday, employers will have the technological capacity to control employees' moods.

(2.) Employers will not refrain from exercising that capacity.

(3.) Most working-age adults will be employees.

(4.) Therefore, someday, most working-age adults will have employers who technologically control their moods.

The argument is valid in the sense that the conclusion (4) follows if all of the premises are true.

Premise 1 seems plausible, given current technological trajectories. Control could either be pharmacological or through direct brain stimulation. Pharmacological control could work, for example, through pills that directly influence your mood, energy levels, submissiveness, ability to concentrate, or passion for the type of task at hand. Direct brain stimulation could work, for example, through a removable Transcranial Magnetic Stimulation (TMS) helmet that magnetically enhances or suppresses neural activity in targeted brain regions, or with some more invasive technology. Cashiers might be able to tweak their dials toward perky friendliness. Data entry workers might be able to tweak their dials toward undistractable focus. Strippers might be able to tweak their dials toward sexual arousal.

Contra Premise 1, society might collapse, or technological development in general might stall out – but let's assume not. If technology as a whole continues to advance, it seems unlikely that mood control specifically will stall. On the contrary, moods seem quite a

likely target for improved technological control, given how readily they can already be influenced by low-tech means like coffee and exercise.

It might take longer than expected. Alternatively, we might already be on the cusp of it. I don't know to what extent people in Silicon Valley, Wall Street, and elite universities already use high-tech drugs to enhance alertness, energy, and concentration at work. What I'm imagining is just a few more steps down this road. Eventually, the available interventions might be much more direct, effective, and precisely targeted.

Premise 2 also seems plausible, given the relative social power of employers vs. employees. As long as there's surplus labor and a scarcity of desirable jobs, employers will have some choice about who to hire. If Starbucks can select between Applicant A who is willing to crank up the perky-friendly dial and Applicant B who is not so willing, then they will presumably prefer Applicant A. If a high-tech startup wants someone who will log intense sixteen-hour days one after the next, and some applicants are willing and able to tweak their brains to enable that, then those applicants will have a competitive edge. If Stanford wants to hire the medical researcher who publishes high-profile studies at an astounding rate, they'll likely discover that it's someone who has dialed up their appetite for work and dialed down everything else.

At first, employees will probably keep the hand on the dial themselves, or mix the drug cocktails themselves. This might be more socially palatable than direct, unmediated control by the employer. For something as initially radical-seeming as a TMS helmet, home use for recreational or medical purposes will probably have to occur first, to normalize it, before it seems natural to wear it also to work.

The first people to yield direct control to employers might be those in low-status, low-education professions, with little bargaining power and with similar job descriptions that seem to invite top-down mass control. The employer might provide an initially voluntary

"energy drink" for all employees at the beginning of the shift. High-status employees, in contrast, might more effectively keep their own hands on the dials. However, the pressure, and consequently the indirect control, might be even more extreme among elite achievers. If mood interventions are highly effective, then they will correlate highly with professional performance, so that as a practical matter those who don't dial themselves to near the ideal settings for work performance will be unlikely to win the top jobs.

Contra Premise 2, (a.) collective bargaining might prevent employers from successfully demanding mood control; or (b.) governmental regulations might do so; or (c.) there might be insufficient surplus labor standing ready to replace the non-compliant.

Rebuttal to (a): The historical trend recently, at least in the U.S., has been against unionization and collective bargaining, though I suppose that could change. One optimistic comparison is the partly successful limitation of performance-enhancing drugs in professional sports. But here the labor market is unusually tightly organized, due to the cooperation of the employers and the formal nature of the competitions.

Rebuttal to (b): Although government regulations could forbid certain drugs or brain-manipulation technologies, if there's enough demand for those drugs or technologies, employees will find a way, unless enforcement gets a lot of resources (as again in professional sports). Government regulations could perhaps specifically forbid employers from requiring the use of certain technologies, while permitting those technologies for home use – but home use vs. use as an employee is a permeable line for the increasing number of jobs that involve working outside of a set time and location. Also, it's easier to regulate a contractual demand than an informal de facto demand. Presumably, many companies could say that of course they don't *require* their employees to drink the cocktail. It's up to the employee! But if the technology is effective, the willing employees will be much more attractive to hire, retain, and promote.

Rebuttal to (c): At present there's no general long-term trend toward a shortage of labor; and at least for jobs seen as the most highly desirable, there will always be more applicants than available positions.

Premise 3 also seems plausible, especially on a liberal definition of "employee". Most working age adults in developed economies are employees of one form or another. That could change with the growth of "gig economy" and more independent contracting, but not necessarily in a way that takes the sting out of the main argument. Even if an Uber driver is technically not an employee, the pressures toward direct mood control for productivity ought to be similar. Likewise for piecework computer programmers and independent sex workers. If anything, the pressures may be higher for gig workers and independent contractors, who generally have less security of income and fewer formal workplace regulations.

If social power remains disproportionately in the hands of employers, they will of course use new neuroscientific technologies to advance their interests, including their interest in extracting as much passion, energy, and devotion as possible from their employees – you and me and our children. If they do it right, we might even like it.

18. Cheerfully Suicidal AI Slaves

Suppose that we someday create genuinely conscious Artificial Intelligence (AI) – beings with all the intellectual and emotional capacities of human beings. For present purposes, it doesn't matter if this is done through computer technology, biotechnology (e.g., "uplifted" animals), or in some other way, as long as the entities are shaped and created by us, for our purposes, with the psychological features we choose.

Here are two things we human creators might want, which appear to conflict:

(1.) We might want them to subordinately serve us and die for us.

(2.) We might want to treat them ethically, as beings with rights and interests that deserve our respect.

A possible fix suggests itself: Design the AIs so that they *want* to serve us and die for us. In other words, create a race of cheerfully suicidal AI slaves. This was Asimov's solution with the Three Laws of Robotics (a solution that slowly falls apart across the arc of his stories).[75]

Douglas Adams parodies the cheerfully suicidal AI, with an animal uplift case in *The Restaurant at the End of the Universe:*

> A large dairy animal approached Zaphod Beeblebrox's table, a large fat meaty quadruped of the bovine type with large watery eyes, small horns and what might almost have been an ingratiating smile on its lips.
>
> "Good evening," it lowed and sat back heavily on its haunches. "I am the main Dish of the Day. May I interest you in parts of my body?" It harrumphed and gurgled a bit, wriggled its hind quarters into a comfortable position and gazed peacefully at them.[76]

Zaphod's naive Earthling companion, Arthur Dent, is predictably shocked and disgusted. When Arthur says he would prefer a green salad, the suggestion is brushed off. Zaphod and

the animal argue that it's better to eat an animal that wants to be eaten, and can consent clearly and explicitly, than one that would rather not be eaten. Zaphod orders four rare steaks for his companions.

> "A very wise choice, sir, if I may say so. Very good," it said. "I'll just nip off and shoot myself."
>
> He turned and gave a friendly wink to Arthur.
>
> "Don't worry, sir," he said. "I'll be very humane."[77]

In this scene, Adams illustrates the peculiarity of the idea. There's something ethically jarring about creating an entity with human-like intelligence and emotion, which will completely subject its interests to ours, even to the point of suicide at our whim – even if the AI *wants* to be subjected in that way.

The three major classes of ethical theory – consequentialism, deontology, and virtue ethics[78] – can each be read in a way that agree with Adams' implicit point. The consequentialist can object that the good of a small pleasure for a human does not outweigh the potential of a lifetime of pleasure for an uplifted steer, even if the steer doesn't appreciate that fact. The Kantian deontologist can object that the steer is treating itself as a "mere means" rather than an agent whose life shouldn't be sacrificed to serve others' goals. The Aristotelian virtue ethicist can say that the steer is cutting its life short rather than flourishing into its full potential of creativity, joy, friendship, and thought.[79]

Using Adams' steer as an anchor point of moral absurdity at one end of the ethical continuum, we can ask to what extent less obvious intermediate cases are also ethically wrong – such as Asimov's robots who don't sacrifice themselves as foodstuffs (though presumably, by the Second Law, they would do so if commanded) but who do, in the stories, appear perfectly willing to sacrifice themselves to save human lives.

When humans sacrifice their lives to save others, it can sometimes be a morally beautiful thing. But a robot designed that way from the start, to always subordinate its interests to human interests – I'm inclined to think that ought to be ruled out by any reasonable egalitarian principle that treats AIs as deserving equal moral status with humans if they have broadly human-like cognitive and emotional capacities. Such an egalitarian principle would be a natural extension of the types of consequentialist, deontological, and virtual ethical reasoning that rule out Adams' steer.

We can't escape the dilemma posed by (1) and (2) above by designing cheerfully suicidal AI slaves. If we somehow create genuinely conscious general-intelligence AI, capable of real joy and suffering, then we must create it morally equal. In fact…

19. We Would Have Greater Moral Obligations to Conscious Robots Than to Otherwise Similar Humans

Down goes HotBot 4b into the volcano. The year is 2050 or 2150, and Artificial Intelligence has advanced sufficiently that such robots can be built with human-grade or more-than-human-grade intelligence, creativity, and desires. HotBot will now perish on this scientific mission. In commanding it to go down, have we done something morally wrong?

The moral status of robots is a frequent theme in science fiction, back at least to Isaac Asimov's robot stories, and the consensus is clear: If someday we manage to create robots that have mental lives similar to ours, with human-like consciousness and a sense of self, including the capacity for joy and suffering, then those robots deserve moral consideration similar to the moral consideration we give to our fellow human beings. Philosophers and AI researchers who have written on this topic generally agree.[80]

I want to challenge this consensus, but not in the way you might predict. I think that if we someday create robots with human-like cognitive and emotional capacities, we owe them *more* moral consideration than we would normally owe to otherwise similar human beings.

Here's why: We will have been their creators and designers. We are thus directly responsible both for their existence and for their happy or unhappy state. If a robot needlessly suffers or fails to reach its developmental potential, it will be in substantial part because of our failure – a failure in our design, creation, or nurturance of it. Our moral relation to robots will more closely resemble the relation that parents have to children, or that gods have to the beings they create (see also Chapter 21), than the relationship between human strangers.

In a way, this is no more than equality. If I create a situation that puts other people at risk – for example, if I destroy their crops to build an airfield – then I have a moral obligation to compensate them, an obligation greater than I have to miscellaneous strangers. If we create genuinely conscious robots, we are deeply causally connected to them, and so we are substantially responsible for their welfare. That is the root of our special obligation.

Frankenstein's monster says to his creator, Victor Frankenstein:

> I am thy creature, and I will be even mild and docile to my natural lord and king, if thou wilt also perform thy part, the which thou owest me. Oh, Frankenstein, be not equitable to every other, and trample upon me alone, to whom thy justice, and even thy clemency and affection, is most due. Remember that I am thy creature: I ought to be thy Adam.[81]

We must either only create robots sufficiently simple that we know them not to merit much moral consideration, or we ought to bring them into existence only carefully and solicitously.

Alongside this duty to be solicitous comes another duty, of knowledge – a duty to know which of our creations are genuinely conscious. Which of them have real streams of subjective experience, and are capable of joy and suffering, or of cognitive achievements such as creativity and a sense of self? Without such knowledge, we won't know what obligations we have to our creations.

Yet how can we acquire the relevant knowledge? How does one distinguish, for instance, between a genuine stream of emotional experience and simulated emotions in a non-conscious computational algorithm? Merely programming a superficial simulation of the emotion isn't enough. If I put a standard computer processor manufactured in 2018 into a toy dinosaur and program it to say "ow!" when I press its off switch, I haven't created a robot capable of suffering. (See also Chapter 15.) But exactly what kind of processing or complexity is necessary for real human-like consciousness? On some views – John Searle's,

for example – consciousness might not be possible in *any* programmed entity; real subjective experience might require a structure biologically similar to the human brain.[82] Other views are much more liberal about the conditions sufficient for robot consciousness. The scientific study of consciousness is still in its infancy. The issue remains wide open.[83]

If we continue to develop sophisticated forms of AI, we have a moral obligation to improve our understanding of the conditions under which consciousness might emerge. Otherwise we risk moral catastrophe – either the catastrophe of sacrificing our interests for beings that don't deserve moral consideration because they have only sham consciousness, or the catastrophe of failing to recognize robot consciousness, and so unintentionally committing atrocities tantamount to slavery and murder against beings to whom we have an almost parental obligation of care.

We have, then, an obligation to learn enough about the material and functional bases of joy, suffering, hope, and creativity to know when and whether our potential future creations deserve such moral concern. And when they do merit such concern we have, a direct moral obligation to treat our creations with an acknowledgement of our special responsibility for their joy, suffering, hopes, and creative potential.[84]

20. How Robots and Monsters Might Destroy Human Moral Systems

Intuitive physics works great for picking berries, throwing stones, and walking through light underbrush. It goes catastrophically wrong when applied to the very large, the very small, the very energetic, or the very fast. Similarly for intuitive biology, intuitive cosmology, and intuitive mathematics: They succeed for practical purposes across long-familiar types of cases, but when extended too far they go seriously astray.

How about intuitive ethics?

In Chapters 18 and 19, I explored the moral consequences of creating AI with human-like conscious experience and cognitive and emotional capacities. But of course if we someday create genuinely conscious AI, it might *not* be very much like us. And then, perhaps, our moral intuitions will serve us as badly as do our physical intuitions when faced with relativity theory and quantum mechanics. What's more, to the extent that our formal moral theories are grounded in ordinary or pre-21st-century moral intuitions, those theories might collapse as well. Applying old-school Aristotelean or Kantian ethics to future AI might be like trying to apply old-school Aristotelian or Kantian physics to interstellar rockets.

#

I will illustrate with a pair of puzzle cases: utility monsters (originally due to Robert Nozick[85]) and what I will call "fission-fusion monsters".

A *utility monster* is a being who derives immense pleasure from harming us or consuming our goods. Cookie Monster, for example, might derive a hundred units of pleasure from every cookie he eats, while normal human beings derive only one unit of

pleasure. If we care about increasing the total amount of pleasure in the world, maybe we should give all of our cookies to the monster. Lots of people would lose out on a little bit of pleasure, but Cookie Monster would be *really* happy!

Cookies are just the start of it, of course. If the world contained an entity vastly more capable of pleasure and pain than are ordinary humans, then on simple versions of happiness-maximizing utilitarian ethics, the rest of us ought to immiserate ourselves to elevate that entity to superhuman pinnacles of joy.

If AI consciousness is possible, including AI joy, I see no reason in principle why that joy should top out at human levels. Crank up the dial higher! Make the joy last longer. Run a hundred thousand copies of it simultaneously on your hard drive. Turn Jupiter into a giant orgasmatron. On one way of thinking, this would be our moral duty.[86] All of human happiness would be a trivial consideration beside this. Even if you don't accept the simple utilitarian view that happiness is *everything*, surely it's something, and if we could multiply the happiness in the Solar System a billionfold, perhaps we ought to, even at substantial cost to ourselves.

Most people seem to find this unintuitive, or even morally repulsive, which was Nozick's point in constructing the thought experiment. Morality doesn't seem to demand that we sacrifice all human happiness to turn Jupiter into a joy machine.

If we want to avoid this conclusion and preserve something like commonsense ethics, we might want to shift focus to the rights of individuals. Even if the monster would get a hundred times as much pleasure from my cookie as I would, it's still *my* cookie. I have a right to it and no obligation to give it up. This is what Nozick thinks and what Kantian critics of utilitarianism also often think. However, this seemingly commonsense solution faces a complementary set of problems.

A fission-fusion monster, let's say, is an entity who can divide at will into many similar descendant beings who retain the monster's memories, skills, and plans, and who can later fuse back together with its own fission products, or the products of other fission-fusion monsters, retaining the memories, skills, and plans of each (with some procedure for resolving conflicts). It's an entity that can split and reunite at will, sometimes unified into a single individual (though the word "individual" is etymologically inapt), sometimes divided up into many separate individuals.

If we say "one conscious intelligence, one vote", how many votes would a fission-fusion monster get? If we say "one unemployed conscious intelligence, one cookie from the dole", how many cookies ought our monster collect? If our fission-fusion monster is selfish and tactical, here's what it might do: On October 31 it splits into a million individuals. On November 1, it collects a million cookies from the dole. On November 2, it casts a million votes for its favorite candidate. On November 3, its million parts merge back together into a single integrated intelligence, ready to enjoy its million cookies and looking forward to the inauguration of its candidate.

Presumably we could block that particular worry by an ad hoc rule, such as that an individual must have fissioned into existence at least X months previously to qualify for such rights. But then setting the X creates problems. Twelve months seems, for example, to be both too short in one way and too long in another. It's too short because a patient Monster might not at all mind waiting a year for such a fantastic advantage. It's too long because fissioned individuals might easily starve to death in a year's time due to unforeseeable consequences beyond their control, while also developing enough individuality to deserve status as an equal and to reasonably view forcible merging as an unwelcome death.

Political, social, and ethical systems that afford rights to individuals have always so far been built on the background assumption that people do not regularly divide and fuse.

The whole thing breaks down, or would at least require radical rethinking, in the face of fission-fusion monsters who can strategically exploit the criteria of individuality to maximize their claims upon the system. This is the intuitive ethics equivalent of trying to apply intuitive physics to systems traveling at 99% the speed of light.

#

If AI experience and cognition is possible, then in the future we might actually face real-world utility monster and fission-fusion monster cases. Indeed, depending on the future of AI, it might be the case that whatever it is about us that we think gives human life special value, whether it is happiness, creativity, love, compassion, intellect, achievement, wisdom – unless, perhaps, it is our species membership itself – could be duplicated a hundredfold or a millionfold in artificial computational or biological systems. Why couldn't it be? And then we might be in rather a confusing pickle.

More generally, our social and ethical structures are founded on principles, practices, and intuitions evolved and constructed to handle the range of variation that we have ordinarily seen in the past. So far, there have been no radically different types of entities who approach or exceed human social intelligence. So far, there have been no entities capable of superhuman pain or pleasure, or of dividing at will into autonomous human-like individuals, no entities pre-programmed to want desperately to sacrifice themselves to satisfy our whims (Chapter 18), no people capable of simply dialing up moods at command (Chapter 17), no people capable of transferring their minds into new bodies, no planet-sized intelligences with people as parts (or maybe there have been: Chapter 39), no simulated realities constructed inside of computers that are as good or better than our "real" reality and over which we have godlike powers (Chapter 21) or into which we can upload ourselves for millions of years

(Chapter 44), no opportunity for us to create dependent artificial people exactly as we see fit. It would be unsurprising if the ethical concepts we now possess, fashioned in much more limited circumstances, fail catastrophically when extended to such new situations.

21. Our Possible Imminent Divinity

We might soon be gods.

In a few decades, we might be creating genuinely conscious artificial intelligences in abundance. (Maybe not. Maybe consciousness could never arise in an artificial system, or maybe we'll destroy ourselves first, or maybe technological innovation will stall out. But I grant me the speculative what-if.) We will then have at least some features of gods: We will have created a new type of being, maybe in our image. We will presumably have the power to shape our creations' personalities to suit us, to make them feel blessed or miserable, to hijack their wills to our purposes, to condemn them to looping circuits of pain or reward, to command their worship if we wish.

If consciousness is only possible in fully embodied robots, our powers might stop approximately there. But if we can also create conscious beings inside of artificial computationally-constructed environments, we might become even more truly divine.

Imagine a simulated world inside of a computer, something like the computer game The Sims, or like a modern Virtual Reality environment, but one where the AIs inside of that environment are sophisticated enough to actually be conscious. Sim Janiece wakes up in the morning, looks around her (simulated) bedroom, sees her simulated husband John, makes some simulated coffee and feels (really!) the caffeine perk.[87] Sim John has a complementary set of experiences. Both Janiece and John are conscious AI programs whose sensory inputs aren't based on sensory scanning of the ordinary environment that you and I see (as a robot's sensory inputs would be) but instead are inputs corresponding to the state of affairs in their virtual-reality environments. Instead of 1s and 0s from a digital camera pointed at a real kitchen, they receive their 1s and 0s from elsewhere in the computer, in accord with the structure of the virtual kitchen as it ought to be sensed from their currently represented point of view. Janiece and John also act only in their simulated world. A normally embodied robot

lifts an ordinary robot arm, generating ordinary visual input of the arm's motion in itself and all other observers and affecting the ordinary thing it's touching. Sim Janiece raises her virtual arm, affecting her and Sim John's virtual input and the state of the virtual world they inhabit.

Let's suppose that Janiece and John are possible. I don't see any compelling reason to think they shouldn't be. If we think that genuine conscious experience is possible in normally-embodied robots, it seems plausible enough that AI systems embodied in simulated worlds could also be conscious.[88] If Janiece and John are possible, we might become even more truly divine than if we merely create normally embodied robots. For now we can command not only the AIs themselves but their entire world.

We approach omnipotence: We can create miracles. We can spawn a Godzilla, revive the dead, move a mountain, undo errors, create a new world or end one on a whim – powers eclipsing those of gods like Zeus and Isis.

We approach omniscience: We can look at any part of the world, look inside anyone's mind, see the past if we have properly recorded it, maybe predict the future in detail, depending on the structure of the program.

We stand outside of space and to some extent time: Janiece and John live in a spatial manifold, or a virtual spatial manifold, which we do not inhabit. Wherever they go, they cannot get away from us, nor can they ever move toward and touch us. Our space, "ordinary" space, does not map on to space as they experience it. We are unconstrained by their spatial laws; we can affect things a million miles apart without in any sense traveling between them. We are, to them, everywhere and nowhere. If the sim has a fast clock relative to our time, we can seem to endure for millennia or longer. We can pause their time and intervene as we like, unconstrained by their clock. We can rewind to save points and thus directly view and

interact with their past, perhaps sprouting off new worlds or rewriting the history of their one world.

If they have a word for "god", the person who launches and manipulates their virtual reality will be quite literally the referent of that word.

Of course, all of this omnipotence, omniscience, and independence of space and time will be relative to *their* world, not relative to our own, where we might remain entirely mortal dingbats. Still, it's divinity enough to raise the ethical question I want to raise, which is:

Will we be *benevolent* gods?

22. Skepticism, Godzilla, and the Artificial Computerized Many-Branching You

Nick Bostrom has argued that we might be "sims".[89] A technologically advanced society might use hugely powerful computers, he argues, to create "ancestor simulations" containing actually conscious people who think that they are living, say, on Earth in the early 21st century but who in fact live entirely inside a giant computational system. David Chalmers has considered a similar possibility in his well-known commentary on the movie *The Matrix*.[90] (See also Chapter 21.)

Neither Bostrom nor Chalmers is inclined to draw skeptical conclusions from this possibility. If we *are* living inside a sim, they suggest, that sim is simply our reality. All the people we know still exist (they're sims, just like us) and the objects we interact with still exist (fundamentally constructed from computational resources, but still predictable, manipulatable, interactive with other such objects, and experienced by us in all their sensory glory). Chalmers even uses the sim scenario as part of an *anti*-skeptical argument: Roughly, as long as our experiences are right, and the functional interactive relationships among all the objects we see are right, it doesn't matter if the fundamental metaphysical structures that undergird all of this are demons or dreams or computers or atoms. Consequently, according to Chalmers, scenarios that are sometimes thought to be skeptical possibilities (we are all brains in vats, this is all a collective dream[91]) turn out not in fact to be so skeptical after all, so long as the structural relationships among experienced objects are sufficiently sound and stable.

Of course, if it's a dream, we might wake up. If it's sim, the owner might suddenly shut it down. To get anti-skeptical juice, one has to assume stability: no wake-up, no shut-down. But if we are to take the simulation scenario seriously, or the group-dream scenario,

or the brain-in-a-vat scenario, then we ought also to think about how large and stable the scenario is likely to be.

Contra the now-standard presentations of the simulation scenario by Bostrom, Chalmers, and others, we ought to address the possibility that if we are living in a sim, it might well be a small or unstable sim – one run by a child, say, for entertainment. We might live for three hours' time on a game clock, existing mainly as citizens who will provide entertaining reactions when, to our surprise, Godzilla tromps through. Or it might just be me and my computer and my room, in an hour-long sim run by a scientist interested in human cognition about philosophical problems. Chalmers might be right to be relatively unconcerned about the fundamental structure of reality, conditional upon the world we experience being large and stable, with approximately the superficial and functional properties we think it has. But the real heart of the skeptical worry, I think, lies not in being possibly wrong about the fundamental nature of things. It lies in the fact that *if* I am wrong in a certain way about the fundamental nature of things – for example, if I am living in a sim – then I ought to doubt that I existed yesterday, and that I will exist tomorrow. I ought to doubt that my sensory experience is tracking a stable and durable reality, that my actions now have long-term consequences of the sort I think they do, and that distant people and things exist. *Maybe* I'm fortunate enough to live in a huge, stable sim of approximately the size and scope I normally take the world to have. But maybe not.

When I expressed these concerns to Bostrom, he responded with the sensible suggestion that to really evaluate the skeptical or non-skeptical consequences of being sims, we need a sense of what types of simulation scenarios are more and less likely.[92] I agree! One comforting, large-sim-friendly thought is this: Maybe the most efficient way to create simulated people is to evolve up a large-scale society over a long period of (sim-clock) time. Another comforting thought is this: Maybe we should expect a technologically advanced

society capable of running a sim to have enforceable ethical standards against running brief sims that contain conscious people.

However, I see no compelling reason to accept such comfortable thoughts. Consider the possibility I'll call the Many-Branching Sim: A set of researchers decide that the best way to create genuinely conscious simulated people is to run a whole simulated universe forward billions of years (sim-years on the simulation clock) from a Big Bang. Now a second group of researchers comes along who also want to host a sim world. They have a choice: Either they could run a new sim world from the ground up, starting at the beginning and clocking forward, or they could take a snapshot of the first group's sim and make a copy. They do some calculations, and it turns out that the second alternative is much easier and less expensive.

Consider the 21st century game Sim City. If you want a bustling metropolis, you can either grow one from scratch or you can use one of many copies created by the programmers and other users. You could also grow one from scratch and then save stages of it on your computer, cutting it off when things don't go the way you'd like, then starting again from a save point. Or you could copy variants of the same city, then grow them in different directions. On the face of it, I see no reason to assume that copying would generally be more difficult than evolving a new sim from the beginning each time. Copying might be favored, also, by two further considerations: If the aim is scientific, controlled experiments might require a copy-and-run-forward approach for each intervention condition. Also, if it turns out that the target levels of intelligence or social structure evolve only in a small minority of sims, then a run-from-the-beginning approach might inconveniently require many attempts.

The Many-Branching Sim scenario, then, is the possibility that there is a *root sim* that is large and stable, starting from some point deep in the past, and then this root sim is copied

into one or more *branch sims* that start from a save point. If there are branch sims, it might be that you now are in one of them, rather than in a root sim or a non-branching sim.

Maybe Sim Corp made the root sim for Earth, took a snapshot on [insert recent date] on the sim clock, then sold thousands or millions of copies to researchers and gamers who now run short-term branch sims for whatever purposes they like. If so, the future of the branch sim in which you are now living might be short – a few sim minutes, hours, or years.

The past is a little trickier to think about. You might conceptualize it either as short or as long, depending on whether you want to count the past in the root world as part of "your" or "your world's" past.

Personal identity becomes a thorny issue. Considering my own case now, on July 13, 2018, according to my clock: If the snapshot was taken at the root sim time of noon on July 12, 2018, then the root sim contains an "Eric Schwitzgebel" who was fifty years old at that moment. Each branch sim would also contain an "Eric Schwitzgebel" developing forward from that point, of which I am one. How should I think of my relationship to those other branch-Erics?

Should I take comfort in the fact that some of them will continue on to live full and interesting lives (perhaps of very different sorts) even if most of them, including probably this particular instantiation of me, will soon be stopped and deleted? Or to the extent I am interested in *my own* future, should I be concerned primarily about what is happening in this branch?

Suppose I look out the window, across the 215/60 freeway and UC Riverside's citrus groves. Wait, is that… *Godzilla* in the distance?! I stare out the window in shock as the monster strides toward campus, crushing orange trees, lifting cars from the freeway. I run out of my office, down the stairs, out toward the north side of campus. But I've chosen my path badly; here he is, coming right at me. As Godzilla steps down to crush me, should I take

comfort in the fact that after the rampage whoever is running this sim will probably delete this branch and start again from the save point with an "Eric Schwitzgebel" still intact? Or would deleting this branch be the destruction of my whole world?

It's philosophically and technologically fascinating to think that we might live in a sim. Given how little we know about the fundamental structure of the cosmos, I see no reason to entirely rule out that possibility.[93] But we have barely scratched the surface of the philosophical consequences.

23. How to Accidentally Become a Zombie Robot

Susan Schneider's work on the future of robot consciousness has me thinking about the possibility of accidentally turning oneself into a zombie.[94] I mean "zombie" in the philosopher's sense: a being who outwardly looks and acts like us but who has no genuine stream of conscious experience. What we are to imagine here is approximately the opposite of what we were imagining in the previous two chapters. The zombie worry is that we might be able to create AIs that are functionally very sophisticated and look from the outside as if they have genuine conscious experience, but really have no conscious experience at all – no more consciousness than your laptop computer would have right now (I assume), even if we programmed it to cry and plead quite convincingly when you try to shut it down.

You might not think that such zombification would be possible, but let's suppose that it is possible: Silicon chips might fail to host consciousness while doing a pretty good job of *faking* consciousness. The fakery might be good enough to fool people who aren't specialists in AI engineering or the science of consciousness; the AIs are fancy plumped up Furby dolls just good enough to fool the majority of non-specialists into thinking that they really are genuinely conscious. Lonely lovers really do fall in love with their sex dolls, elderly people with their robot companion nurses, children with their nanny bots. They can't falsely but deeply believe that these lovable robots have real streams of subjective experience behind their speech and facial expressions. People and robots begin to intermarry, maybe – at first without the approval of state or church. Robot rights becomes a popular movement. People even begin to "upload" their minds into computers, destroying their biological brains in the process.

Among specialists on AI consciousness, let's suppose, opinion is sharply divided. Some philosophers, psychologists, and AI engineers support the popular opinion, while others retain suspicions that these robots don't really have conscious experience. Some of

these skeptics, maybe, have been reading John Searle and Ned Block, who argue that no amount of computational equivalence could guarantee the existence of genuine conscious experience in a robot without some meatier biological similarity too.[95] Others don't go as far, holding that an ideally designed silicon robot could be conscious, but they doubt that these robots are sufficiently well designed, despite their ability to fool the masses. These suspicious experts are alarmed to see human lives sometimes sacrificed to save robot lives. They are alarmed to see their friends "upload" and then "tell" everyone how awesome it is inside the Cloud.

Finally, suppose that you're on the fence. Are the robots and the uploaded people really conscious, or is it all just delusion? How can you know? The problem is urgent. You're old – old enough to remember having seen Ned Block's and John Searle's convincing lectures in person. Moreover, you're at a high risk of stroke. If silicon chips really can host consciousness as advertised, now is the time for you to swap out your organic material, before it's too late.

Fortunately, the iBrain store has just invented Try-It-Out technology, adapting an old suggestion of Susan Schneider's.[96] You can go into the store and temporarily upload your mind into a robot. The iBrain store is inviting potential customers to check out robot consciousness for themselves, see for themselves from the inside what it's like to be a robot, if indeed there's anything it's like. During Try-It-Out, iBrain Company claims, you can *introspectively* discover whether "uploaded you" really is conscious. You needn't rely on anyone else's report! You can spend twenty minutes instantiated in silicon. When the experiment is done, you'll be ported back into your brain with updated memories of your experience or lack thereof, and you'll know the answer without having to rely on the dubious say-so or seeming-say-so of sims and robots. The question will finally be settled.

From the outside, it will look like this: You walk into the iBrain store. You fill out some forms and are then escorted to a clean, quiet room in the back where a physician is waiting. The physician and her technicians put you under anesthesia and scan your brain. In the corner of the room is a robot body, which now comes to life. A speech stream comes from the robot: "Yes, I really am conscious! Wow!" The physician asks a series of questions. The robot shows proper awareness of its body, its surroundings, the recent past, and your biographical details. The robot then does some jumping jacks to further explore the body, bends an iron rod with its robotic strength, does a few more showy feats (which have been found to improve sales). Robot-you then goes to sleep. A snapshot of its brain is taken, to capture the new memories. The technicians then stimulate your sleeping biological brain to insert memories from the robotic phase, and finally they wake you up.

You sit up happy. "Yes, I was conscious even in the robot," you say. "My philosophical doubts were misplaced. Upload me into iBrain!"

The physician reminds you that according to the new federal regulations, two copies of a person cannot be run simultaneously, so that, after your upload, biological-you will need to be sedated indefinitely. That's fine, you reply. You no longer have any qualms.

#

Whoops, I left out an important part of the story.

You never did any of those things. After the physician sedated you, she and the technicians went to the break room to play cards. Your brain was scanned, but nothing was ever loaded into the robot. The robot never came to life, never declared its own consciousness, never answered biographical questions, never did jumping jacks. After twenty minutes had passed, the physician and technicians updated your brain with *fake*

memories of having done all of those things – fake memories based on plausible predictions about what you would have done had you actually been uploaded into the robot. After a while, they had noticed that fake memories worked as well as real ones, and it was easier and less expensive to just skip the middle phase. It would of course be a terrible scandal if they were caught, but no one ever showed the faintest suspicion.

How could anyone know after waking up in their biological brain whether their currently conscious seeming-memory of having consciously said "I'm really conscious!" was really a real memory of having consciously said "I'm really conscious!" as opposed to a mere sham? *Now*, you're conscious. *Now*, you're seeming to remember it. Vividly conscious for you right now is the seeming-memory of delight and surprise, and of having experienced the world through robot eyes and of having felt the strength of robot arms bending iron. But the fact that these memories are conscious for you now is no guarantee, of course, that the thing you seem now to have consciously experienced was in fact consciously experienced at the time.

#

It's sad of course that you were fooled. However, the savvier and more suspicious alternative-you waited a little longer, for a later technological development: piece-by-piece Try-It-Out. Alternative-you foresaw the fake-memory difficulty. So alternative-you held out for something even closer to Schneider's original suggestion.

Here's how piece-by-piece Try-It-Out works. Some portion of your brain – let's say the portion of your cortex responsible for tracking such-and-such features on the left side of your visual field – is scanned in detail, and a visual processing system made from silicon computer chips is manufactured to replace it. The question is: Is this silicon visual cortex

really capable of hosting genuine conscious experience? Or, despite its capacity to do visual computational processing, might it be mere zombie-stuff? It's not *strictly* functionally identical, of course. At the micro-level it works very differently; and it will break down under different conditions; and it's better in some ways, with internal algorithms to correct for your nearsightedness and astigmatism and faster resolution time for some details. But just like your regular visual cortex, it will take neural input from pathways X, Y, and Z; and just like your regular visual cortex, it will output interpretable neural signals to other relevant regions of the brain, J, K, and L.

Based on your experience at the iBrain store, we also know that memories of seeming acts of successful introspection also aren't enough to establish genuine consciousness. After the scandal and lawsuits, even iBrain Company admitted as much. *Simultaneous* introspection – that's the right test, they say! The introspection of one's own current conscious experience. After all, that's infallible, right? Or as close to infallibility as a human can get. Even Descartes in his most skeptical moments couldn't doubt *that*.[97]

So, you are sedated – alternative-you, actually, but let's drop the "alternative" part. The interface between the selected portion of your brain and the rest of your brain is carefully mapped, synapse-by-synapse. Blood flow, hormonal regulation, and other neurophysiological features are also taken into account. All of this information is beamed in real-time to a visual-cortex chip in a computer on the bedside table, waiting to be installed. The chip now runs in parallel to that targeted portion of your visual cortex. The chip beams outputs which stand available to be taken as transceiver inputs to the rest of your brain.

A switch is flipped. A transcranial magnetic stimulator damps down the activity in the target region of visual cortex. Simultaneously, the transceivers on the interface surfaces of the remainder of your brain go live. The chip is taking inputs from the other regions of your brain, and it is doing its visual processing, and then it is giving interpretable outputs

back into those other regions. So far, it's a success! The remainder of the brain doesn't seem to be noticing any difference. Well, why would it? The technology is highly advanced, perfected after years of trials and hundreds of billions of research dollars. The inputs received by the remainder of the brain are almost exactly what it would have ordinarily received from the neural tissue that the silicon chip is designed to replace.

But you are still sedated, not fully conscious – you haven't yet carefully introspected.

You are eased out of sedation. The doctors ask how you feel.

"I feel fine," you say. "Normal. Have you done the procedure? Are we Trying It Out?"

Yes, they say. They advise you to introspect as carefully as you can.

Here's what will *not* happen: You will not notice any difference that inclines you to make a very different outward report than you otherwise would make, or that would affect your motor cortex or prefrontal cortex or basal ganglia in any different way. (Maybe you'll say something like, "Ooh, things seem even clearer than with my natural vision. This is great!") You will not act out any very different decision than you would have with an ordinary biological visual cortex. You will not feel any very different surge of emotion, have any large hormonal change, or lay down any very different memories, except insofar as the additional visual clarity might impress you. After all, the input the rest of the brain receives from the chip is functionally similar to the input it would have received from ordinary visual cortex, differing mainly in improved clarity. With such similar inputs, how could introspection possibly reveal any disastrous loss? The whole process was designed exactly *not* to trigger an introspective crisis; that's exactly why the chip was structured as it is and the transceivers placed where they have been placed. A hundred billion research dollars created a procedure structured exactly to ensure that no noticeable difference would trigger a shocked introspective report of no experience.

The Try-It-Out process goes swimmingly, of course. You ace the vision tests, you report no introspective weirdness, you declare that you really consciously experience the visual world in all its magnificence.

The switch is flipped back, and everything returns to normal – a bit disappointingly fuzzy, actually. You were already getting used to the computer vision.

'Proceed with the surgery!" you say. The doctors install the chip, replacing that portion of your brain. Piece by piece, over the next year, doctors replace your whole brain. You never report noticing a difference.

Sadly, however, the skeptics were right.

There is no consciousness in silicon computer chips. Despite broad functional similarity at the input-output level, differences in lower-level processing and microstructure, it turns out (let's suppose, for the sake of this thought experiment) *are* crucial to the presence or absence of genuine conscious experience. Brains, for example, implement a parallel processing architecture, whereas the silicon chips only mimic parallel processing in a fast serial architecture. Maybe that turns out to matter immensely. Or maybe it matters that brains use analog accumulations to fire approximately digital action potentials, whereas silicon chips are digitally structured through and through – or that brains are sometimes sensitive to real quantum chance while silicon chips use complex clock algorithms to imitate chance, or that brains are juicy carbon, while silicon is dry and not nearly as delicious. Maybe each of the 47 silicon chips that now constitute your brain is, individually, a locus of massive information integration, more so than your brain as a whole, with the result that there are 47 streams of specialist consciousness but no overall integrated consciousness of the whole person.[98] Or…. Some basic structural feature of the brain that is crucial for the real presence of consciousness is absent in the chip, despite the lack of big introspectively detectable differences or big differences in outward behavior.

The whole basis of wanting to Try-It-Out, rather than trusting that broad input/output functional similarity is enough, is the worry that the presence or absence of consciousness might depend on some such architectural feature. But if that is so, even piece-by-piece Try-It-Out won't reveal that fact.

You have accidentally become a zombie robot.[99]

Part Three: Regrets and Birthday Cake

24. Dreidel: A Seemingly Foolish Game That Contains the Moral World in Miniature

Superficially, dreidel looks like a simple game of luck, and a badly designed game at that. It lacks balance, clarity, and meaningful strategic choice. From this perspective, its prominence in the modern Hannukah tradition is puzzling. Why encourage children to spend a holy evening gambling, of all things?

This perspective misses the brilliance of dreidel. Dreidel's seeming flaws are exactly its virtues. Dreidel is the moral world in miniature.

If you're unfamiliar with the game, here's a tutorial. You sit in a circle with friends or relatives and take turns spinning a wobbly top, the dreidel. In the center of the circle is a pot of several foil-wrapped chocolate coins, to which everyone has contributed from an initial stake of coins they keep in front of them. If, on your turn, the four-sided top lands on the Hebrew letter *gimmel,* you take the whole pot and everyone needs to contribute again. If it lands on *hey,* you take half the pot. If it lands on *nun,* nothing happens. If it lands on *shin,* you put in one coin. Then the next player takes a spin.

It all sounds very straightforward, until you actually start to play the game.

The first odd thing you might notice is that although some of the coins are big and others little, they all count just as one coin in the rules of the game. This is unfair, since the big coins contain more chocolate, and you get to eat your stash at the end. To compound the unfairness, there's never just one dreidel – each player can bring their own – and the dreidels are often biased, favoring different outcomes. (To test this, a few years ago my daughter and I spun a sample of eight dreidels forty times each, recording the outcomes. One particularly cursed dreidel landed on *shin* an incredible 27/40 times.) It matters a lot which dreidel you spin.

And the rules are a mess! No one agrees whether you should round up or round down with *hey*. No one agrees when the game should end or how low you should let the pot get before you all have to contribute again. No one agrees how many coins to start with or whether you should let people borrow coins if they run out. You could try appealing to various authorities on the internet, but in my experience people prefer to argue and employ varying house rules. Some people hoard their coins and their favorite dreidels. Others share dreidels but not coins. Some people slowly unwrap and eat their coins while playing, then beg and borrow from wealthy neighbors when their luck sours.

Now you can, if you want, always push things to your advantage – always contribute the smallest coins in your stash, always withdraw the largest coins in the pot when you spin *hey,* insist on always using the "best" dreidel, always argue for rules interpretations in your favor, eat your big coins then use that as a further excuse to contribute only little ones, and so forth. You could do all this without ever breaking the rules, and you'd probably end up with the most chocolate as a result.

But here's the twist, and what makes the game so brilliant: The chocolate isn't very good. After eating a few coins, the pleasure gained from further coins is minimal. As a result, almost all of the children learn that they would rather be kind and generous than hoard up the most coins. The pleasure of the chocolate doesn't outweigh the yucky feeling of being a stingy, argumentative jerk. After a few turns of maybe pushing only small coins into the pot, you decide you should put in a big coin next time, just to be fair to the others and to enjoy being perceived as fair by them.

Of course, it also feels bad always to be the most generous one, always to put in big, take out small, always to let others win the rules arguments, etc., to play the sucker or self-sacrificing saint.

Dreidel, then, is a practical lesson in discovering the value of fairness both to oneself and others, in a context where the rules are unclear and where there are norm violations that aren't rules violations, and where both norms and rules are negotiable, varying by occasion – just like life itself, only with mediocre chocolate at stake. I can imagine no better way to spend a holy evening.

25. Does It Matter If the Passover Story Is Literally True?

You probably already know the Passover story: How Moses asked Pharaoh to let his enslaved people leave Egypt, and how Moses' god punished Pharaoh – killing the Egyptians' firstborn sons while "passing over" the Jewish households. You might even know the new ancillary tale of the Passover orange. How much truth is there in these stories? At synagogues during Passover holiday, myth collides with fact, tradition with changing values. Negotiating this collision is the puzzle of modern religion.

Passover is a holiday of debate, reflection, and conversation. In 2016, as my family and I and the rest of the congregation waited for the Passover feast at our Reform Jewish temple, our rabbi prompted us: "Does it matter if the story of Passover isn't literally true?"

Most people seemed to be shaking their heads. No, it doesn't matter.

I was imagining the Egyptians' sons. I am an outsider to the temple. My wife and teenage son are Jewish, but I am not. At the time, my nine-year-old daughter, adopted from China at age one, was describing herself as "half Jewish".

I nodded my head. Yes, it does matter if the Passover story is literally true.

"Okay, Eric, why does it matter?" Rabbi Suzanne Singer handed me the microphone.

I hadn't planned to speak. "It matters," I said, "because if the story is literally true, then a god who works miracles really exists. It matters if there is a such a god or not. I don't think I would like the ethics of that god, who kills innocent Egyptians. I'm glad there is no such god.

"It is odd," I added, "that we have this holiday that celebrates the death of children, so contrary to our values now."

The microphone went around, others in the temple responding to me. Values change, they said. Ancient war sadly but inevitably involved the death of children. We're really celebrating the struggle of freedom for everyone….

Rabbi Singer asked if I had more to say in response. My son leaned toward me. "Dad, you don't have anything more to say." I took his cue and shut my mouth.

Then the Seder plates arrived with the oranges on them.

Seder plates have six labeled spots: two bitter herbs, charoset (a mix of fruit and nuts), parsley, a lamb bone, a boiled egg – each with symbolic value. There is no labeled spot for an orange.

The first time I saw an orange on a Seder plate, I was told this story about it: A woman was studying to be a rabbi. An orthodox rabbi told her that a woman belongs on the bimah (pulpit) like an orange belongs on the Seder plate. When she became a rabbi, she put an orange on the plate.

A wonderful story – a modern, liberal story. More comfortable than the original Passover story for a liberal Reform Judaism congregation like ours, proud of our woman rabbi. The orange is an act of defiance, a symbol of a new tradition that celebrates gender equality.

Does it matter if it's true?

Here's what actually happened. Dartmouth Jewish Studies professor Susannah Heschel was speaking to a Jewish group at Oberlin College in Ohio. The students had written a story in which a girl asks a rabbi if there is room for lesbians in Judaism, and the rabbi rises in anger, shouting, "There's as much room for a lesbian in Judaism as there is for a *crust of bread* on the Seder plate!" The next Passover, Heschel, inspired by the students but reluctant to put anything as unkosher as bread on the Seder plate, used a tangerine instead.[100]

The orange, then, though still an act of defiance, is also already a compromise and modification. The shouting rabbi is not an actual person but an imagined, simplified foe.

It matters that it's not true. From the story of the orange, we learn a central lesson of Reform Judaism: that myths are cultural inventions built to suit the values of their day,

idealizations and simplifications, changing as our values change – but that only limited change is possible within a tradition-governed institution. An orange, but not a crust of bread.

In a way, my daughter and I are also oranges: a new type of presence in a Jewish congregation, without a marked place, welcomed this year, unsure we belong, at risk of rolling off.

In the car on the way home, my son scolded me: "How could you have said that, Dad? There are people in the congregation who take the Torah literally, very seriously! You should have seen how they were looking at you, with so much anger. If you'd said more, they would practically have been ready to lynch you."

Due to the seating arrangement, I had been facing away from most of the congregation. I hadn't seen those faces. Were they really so outraged? Was my son telling me the truth on the way home that night? Or was he creating a simplified myth of me?

In belonging to an old religion, we honor values that are no longer entirely our own. We celebrate events that no longer quite make sense. We can't change the basic tale of Passover. But we can add liberal commentary to better recognize Egyptian suffering, and we can add a new celebration of equality.

Although the new tradition, the orange, is an unstable thing atop an older structure that resists change, we can work to ensure that it remains. It will remain only if we can speak its story compellingly enough to give our new values too the power of myth.

26. Memories of My Father

I wrote the following shortly after my father died in 2015. I share it with you now partly as a tribute to my father, to whom this book is dedicated, and partly because I think this portrait of him will give you a better understanding of my background. Thinking about his life helps make vivid for me how my reflections on technology, my appreciation of weirdness, my interest in philosophical discourse with non-specialists, and my interest in moral psychology all spring from the same root.

#

My father, Kirkland R. Gable (born Ralph Schwitzgebel), died on Sunday. Here are some things I want you to know about him.

Of teaching, he said that authentic education is less about textbooks, exams, and technical skills than about moving students "toward a bolder comprehension of what the world and themselves might become."[101] He was a beloved psychology professor at California Lutheran University.

I have never known anyone, I think, who brought as much creative fun to teaching as he did. He gave out goofy prizes to students who scored well on his exams (for instance, a wind-up robot nun who breathed sparks of static electricity: "Nunzilla"). Teaching about alcoholism, he would start by pouring himself a glass of wine (actually water with food coloring), then more wine, and more wine, acting drunker and drunker, arguing with himself, as the class proceeded. Teaching about child development, he would stand my sister or me in front of the class, and we'd move our mouths like ventriloquist dummies as he stood behind us, talking about Piaget or parenting styles – and then he'd ask our opinion about parenting

styles. Teaching about neuroanatomy, he'd bring a brain jello mold, which he sliced up and passed around for the students to eat ("yum! occipital cortex!"). Etc.

As a graduate student and then lecturer at Harvard in the 1960s and 1970s, he shared the idealism of his mentors Timothy Leary and B.F. Skinner, who thought that through understanding the human mind we can transform and radically improve the human condition – a vision my father carried through his entire life.[102] His comments about education captured his ideal for thinking in general: that we should aim toward a bolder comprehension of what the world and ourselves might become.

He was always imagining the potential of the young people he met, seeing things in them that they often didn't see in themselves. He especially loved juvenile delinquents (as they were then called), who he encouraged to think expansively and boldly. He recruited them from street corners, paying them to speak their hopes and stories into reel-to-reel tapes, and he recorded their declining rates of recidivism as they did this, week after week. His book about this work, *Streetcorner Research,* was a classic in its day. As a prospective philosophy graduate student in the 1990s, I proudly searched the research libraries of the schools I was admitted to, always finding multiple copies with lots of date stamps from checkouts in the 1960s and 1970s.

With his twin brother Robert, he invented the electronic monitoring ankle bracelet, now widely used as an alternative to prison for non-violent offenders. He wanted to set teenagers free from prison, rewarding them for going to churches and libraries instead of street corners and pool halls. He had a positive vision rather than a penal one. He imagined everyone someday using location monitors to share rides and to meet nearby strangers with mutual interests – ideas which, in 1960, were about fifty years before their time.

With degrees in both law and psychology, he helped to reform institutional practice in insane asylums – which were often terrible places in the 1960s, whose inmates had no

effective legal rights.[103] He helped force those institutions to become more humane and to release harmless inmates held against their will. I recall his stories about inmates who were often, he said, "as sane as could be expected, given their current environment", and maybe saner than their jailors – for example, an old man who decades earlier had painted his neighbor's horse as an angry prank, and thought he would "get off easy" if he convinced the court he was insane.

As a father, he modeled and rewarded unconventional thinking. We never had an ordinary Christmas tree that I recall – always instead a life-size cardboard Christmas Buddha (with blue lights poking through his eyes), or a stepladder painted green then strung with ornaments, or a wild-found tumbleweed carefully flocked and tinseled – and why does it have to be on December 25th? I remember a few Saturdays when we got hamburgers from different restaurants and ate them in a neutral location – I believe it was the parking lot of a Korean church – to see which burger we really preferred. (As I recall, he and my sister settled on the Burger King Whopper, while I could never confidently reach a preference, because it seemed like we never got the methodology quite right.)

He loved to speak with strangers, spreading his warm silliness and unconventionality out into the world. If we ordered chicken at a restaurant, he might politely ask the server to "hold the feathers". Near the end of his life, if we went to a bank together he might gently make fun of himself, saying something like "I brought along my brain", gesturing toward me with open hands, "since my other brain is sometimes forgetting things now". For years, though we lived nowhere near any farm, we had a sign from the Department of Agriculture on our refrigerator, sternly warning us never to feed table scraps to hogs.

I miss him painfully, and I hope that I can live up to some of the potential he so generously saw in me, carrying forward some of his spirit.

27. Flying Free of the Deathbed, with Technological Help

Rereading my reflections on my father, I am struck by one contrast between his vision and mine. My father so creatively saw the positive potential in people and technology – wonderfully imagining how to turn things toward the better. My vision, through Parts One and Two of this book at least, has been much more mixed, tending toward the negative – with plenty of abusive corporations, misleading applications of technology, jerks and hypocrites, and failures of self-knowledge.

Here then is an expansive, creative vision of a positive possibility for my father.

#

My father spent the final twenty years of his life disabled and often bedridden. In addition to two forms of life-threatening cancer, he suffered from Chronic Regional Pain Syndrome in one foot. The CRPS gave him constant pain which could be seriously aggravated, sometimes for weeks, from even mild exertion such as ten minutes' walking, or from jostling the foot while sleeping or in a wheelchair. It was, I suspect, the CRPS that ultimately killed him, through the side-effects of long-term narcotics and the bodily harm of spending years mostly immobile in bed, including near-paralysis of his digestive system.

My father's last word was "up". I had poured a laxative in his mouth, to try to get his bowels moving again so we could feed his fast-failing body. He had aspirated the laxative into his lungs. By "up" he probably meant "sit me up straighter, I'm choking", but maybe – I prefer to imagine this – he was expressing the upward hope for Heaven that helped sustain him in his final months.

I have often wished that we could have freed my father up away from his horrible bed. I've tried it in imagination many times, drafting out science fiction stories featuring an

elderly person who uses virtual reality or "telepresence" to find new meaning and potential for action in the world beyond the bedroom. Although I've written some science fiction stories I'm proud of, this particular story never comes out right. So instead of that story I can't yet write, let me discuss the technological innovation I have in mind.

Some elements of this idea are already being implemented in current telepresence technologies. First, equip an able-bodied volunteer, the *host*, with a camera above each eye and a microphone by each ear. Equip the bedridden person, the *rider*, with Virtual Reality gear that immersively presents these audiovisual stimuli to the rider's eyes and ears. Also equip the rider with a microphone to speak directly into the host's ear. Now send the host on a trip. During this trip, let the host be guided mainly by the rider's expressed desires, walking where the rider wants to walk, looking where the rider wants to look, stopping and listening where the rider wants to stop and listen. Unlike VR tours as they currently exist, the host can interact with and alter the environment in real time. The rider could have the host lift, turn, and examine a flower, then cast it into a stream and watch it drift away. The host could purchase goods or services on the rider's behalf. The rider could conduct a conversation with the locals, by having the host speak the rider's words verbatim almost simultaneously with the rider's speaking them into his ear – which is surprisingly easy to do with a little practice.[104] Alternatively, the rider might have a separate speaker output from the host's helmet, allowing the rider to speak directly.

Next, let's don some VR gloves. As I imagine it, rider and host wear matching gloves. These gloves are synchronized to move in exactly the same way – of course with quick escape overrides and perhaps the rider's motions damped down to prevent overextension or bumping into unseen obstacles near the bed. Glove synchronization will require both good motion tracking (Nintendo Wii, improved) and some ways of restricting or guiding the movements of the gloves on each end (perhaps through magnets and gyres).

Appropriately synchronized, when the rider starts to move his hand on vector V and the host starts to move her hand on vector W, each motion is nudged toward some compromise vector (V+W)/2. An intuitive collaboration will be essential, so that V and W don't start too far apart – a familiarity acquired over time, with gentle, predictable movements and anticipatory verbal cues ("let's pick that blue flower"). With practice, in safe, simple, and predictable environments, it should come to seem to the rider as if it is almost his own hands that are moving in the seen environment. This impression could be further enhanced with tactile feedback – that is, if pressure sensors in the host's gloves connect with actuators in the rider's gloves that exert corresponding pressures in corresponding locations.

A final, expensive, and much more conjectural step would be to equip host and rider with helmets with brain imaging technology and Transcranial Magnetic Stimulation (or some other way of directly stimulating and suppressing brain activity). For example, for a fuller tactile experience, activity in the host's primary somatosensory cortex could be tracked, and a vague, faint echo of it could stimulated in matching areas in the rider's cortex. You wouldn't want too much synchrony – just a hint of it – and anyhow, brains differ even in relatively similarly structured regions like somatosensory cortex. Of course, also, you wouldn't want too much motor signal traveling down efferent nerves into the rider's body, making the rider move around in bed. Just a hint, just a whiff, just a rough approximation – a dim, vague signal that might be highly suggestive in an otherwise well-harmonized, collaborative host and rider, in a rich Virtual Reality environmental context with clear cues and expectations.

Let's boldly imagine all of this in a positive, harmonious, non-exploitative relationship between rider and host. The host will almost forget his bed, will explore and laugh and play in regions far beyond his little bedroom, will feel that he is truly back in the wide world, at least for a while.

My father was both a psychologist and an inventor. In 1995, when he was first diagnosed with cancer, he had been wanting to go to Hong Kong, and he had to cancel the trip to attempt a bone marrow transplant. He died twenty years later, never having made it to Hong Kong. I wish I could bring my father back to life, build some of this technology with him, then take him there.

28. Thoughts on Conjugal Love

In 2003, my Swiss friends Eric and Anne-Françoise Rose asked me to contribute something to their wedding ceremony. Here's a lightly revised version of what I wrote, concerning conjugal love, the distinctive kind of love between spouses.

#

Love is not a feeling. Feelings come and go, while love is steady. Feelings are passions in the classic sense of *passion*, which shares a root with "passive" – they arrive mostly unbidden, unchosen. Love, in contrast, is something built. The passions felt by teenagers and writers of romantic lyrics, felt so intensely and often so temporarily, are not love – though they might sometimes be the prelude to it.

Rather than a feeling, love is a way of structuring your values, goals, and reactions. Central to love is valuing the good of the other for their own sake.[105] Of course, we all care about the good of other people we know, for their own sake and not just for other ends. Only if the regard is deep, only if we so highly value the other's well-being that we are willing to thoroughly restructure our own goals to accommodate it, and only if this restructuring is so rooted that it automatically informs our reactions to the person and to news that could affect them, do we possess real love.

Conjugal love involves all of this, but it is also more than this. In conjugal love, one commits to seeing one's life always with the other in view. One commits to pursuing one's major projects, even when alone, in a kind of implicit conjunction with the other. One's life becomes a co-authored work.

Parental love for a young child might be purer and more unconditional than conjugal love. The parent expects nothing back from a young child. The parent needn't share plans

and ideals with an infant. Later, children will grow away into their separate lives, independent of parents' preferences, while we retain our parental love for them.

Conjugal love, because it involves the collaborative construction of a joint life, can't be unconditional in this way. If the partners don't share values and a vision, they can't steer a mutual course. If one partner develops too much of a separate vision or doesn't openly and in good faith work with the other toward their joint goals, conjugal love fails and is, at best, replaced with some more general type of loving concern.

Nevertheless, to dwell on the conditionality of conjugal love, and to develop a set of contingency plans should it fail, is already to depart from the project of jointly fabricating a life, and to begin to develop individual goals opposing those of the partner. Conjugal love requires an implacable, automatic commitment to responding to all major life events through the mutual lens of marriage. One can't embody such a commitment while harboring serious back-up plans and persistent thoughts about the contingency of the relationship.

Is it paradoxical that conjugal love requires lifelong commitment without contingency plans, yet at the same time is contingent in a way that parental love is not? No, there is no paradox. If you believe something is permanent, you can make lifelong promises and commitments contingent upon it, because you believe the thing will never fail you. Lifelong commitments can be built upon bedrock, solid despite their dependency on that rock.

This, then, is the significance of the marriage ceremony: It is the expression of a mutual unshakeable commitment to build a joint life together, where each partner's commitment is possible, despite the contingency of conjugal love, because each partner trusts the other partner's commitment to be unshakeable.

A deep faith and trust must therefore underlie true conjugal love. That trust is the most sacred and inviolable thing in a marriage, because it is the very foundation of its possibility. Deception and faithlessness destroy conjugal love because, and to the extent that,

they undermine that trust. For the same reason, honest and open interchange about long-standing goals and attitudes is at the heart of marriage.

Passion alone can't ground conjugal trust. Neither can shared entertainments and the pleasure of each other's company. Both partners must have matured enough that their core values are stable. They must be unselfish enough to lay everything on the table for compromise, apart from those permanent, shared values. And they must resist the tendency to form secret, selfish goals. Only to the degree they approach these ideals are partners worthy of the trust that makes conjugal love possible.

29. Knowing What You Love

In a 1996 article on self-knowledge, Victoria McGeer argues that our claims about our attitudes are likely to be true mainly because once you avow an attitude, whether to yourself or others, you are thereafter committed to living and speaking and reasoning accordingly, unless you can give some account of why you aren't doing so.[106] Since you have considerable self-regulatory control over how you live, speak, and reason; and since all there is to having an attitude is being prone to live, speak, and reason in ways that fit that attitude; you have the power to make true what you say about yourself. In short, you shape yourself to fit the attitudes you express. On McGeer's picture, self-knowledge derives more from self-shaping than it does from introspectively discovering attitudes that already exist.

This model of self-knowledge works especially well, I think, for love.

Suppose I'm up late with some friends at a bar. They're talking jazz, and I'm left in the dust. More to participate in the conversation and to seem knowledgeable than out of any prior conviction, I say, "I just love Cole Porter's ballads". I could as easily have said I love Irving Berlin or George Gershwin. About all of these composers, I really only know a half-dozen songs, which I've heard occasionally performed by different artists. My friends ask what I like about Porter; I say something hopefully not too stupid. Later, when we're driving in my car, they expect to hear Cole Porter. When a Porter biopic is released, they ask my opinion about it. I oblige them. Although this pattern of action arises partly from my desire to fulfill the expectations I've created by my remark, it's not just empty show. I do enjoy Cole Porter; and I find myself drawn even more to him now. This isn't so unlikely. The psychological literature on cognitive dissonance, for example, suggests that we tend to subsequently shape our general opinions to match what we have overtly said, if it was said without obvious coercion.[107]

I have transformed myself into a Cole Porter fan as the result of a casual remark. It wasn't true of me before I said it; but now I've made it true. If love is a kind of commitment to valuing something, or a pattern of specially valuing it, I embarked on that commitment and began that pattern by making the remark. The accuracy of my declaration that I love Cole Porter derives not from acute introspection of some prior cognitive state but rather from the way I shape myself into a consistent, comprehensible person, for my own benefit and the benefit of others, once something truthy has dropped from my mouth.

If I say to myself in the scoop shop that I love Chunky Monkey ice cream, I am at least as much forming a commitment, or creating a policy and reference point for future choices, as I am scouring my mind to discover a pre-existing love. Of course, it's highly relevant that I remember enjoying Chunky Monkey last time I had it; but remembering a pleasure is no declaration of love. To endorse the thought that I don't just enjoy the flavor but actually love it is to embrace a relationship between myself and it. The same goes if I decide that I love the San Francisco 49ers, or the writings of Michel de Montaigne, or Yosemite Valley in the fall.

If I tell someone for the first time that I love her, I am not, I hope, merely expressing an emotion. Rather, I am announcing a decision. I am diving into a commitment, not easily reversed, to value her in a certain way. How frightening!

The commitment in loving another person dwarfs the commitment in loving Chunky Monkey, and we judge people very differently who abandon these commitments. But even the smallest love requires regulative self-consistency. We can't ceaselessly and arbitrarily flop around in our loves while continuing to be normal choice-makers and comprehensible members of a community. Hereafter, you must either default to giving Chunky Monkey very strong consideration or stand ready to explain yourself.

#

In that moment you first declare your love, is it already true that you love? You haven't yet done the work. It could turn out either way. Your declaration was a fleeting shadow, forgotten the next day, or it was the crucial beginning of something that endures. If the resolve endures, your declaration was true – but whether the resolve endures depends on circumstances beyond your control and on parts of yourself you cannot see. You can guess, based on a strength of feeling or sense of your own seriousness – but if you lose your job tomorrow, maybe your world goes sideways, uprooting the seedlings.

#

Can we similarly have other-knowledge through other-shaping? It's a tyrannical business, but I don't see why it couldn't happen. Imagine a mother who declares that her four-year-old son loves baseball, then works to make it true. Or imagine Stalin declaring that his followers hate Trotsky.

30. The Epistemic Status of Deathbed Regrets

Every year around graduation time we hear uplifting thoughts about what people do and do not regret on their deathbeds. The intended lesson is *pursue your dreams! Don't worry about money!*

I can find no systematic research about what people on their deathbeds do in fact say they regret. A database search of psychology articles on "death*" and "regret*" turns up a 2005 article by Erika Timmer and colleagues as the closest thing. Evidently, what elderly East Germans most regretted is having been victimized by war.[108]

Let's grant that the commencement truisms have a prima facie plausibility. With their dying breaths, grandparents around the world say, "If only I had pursued my dreams and worried less about money!" Does their dying perspective give them wisdom? Does it matter that it is dying grandparents who say this rather than forty-year-old parents or high school counselors or assistant managers at regional banks? The deathbed has rhetorical power. Does it deserve it?

I'm reminded of the wisdom expressed by Zaphod Beeblebrox IV in *The Hitchhiker's Guide to the Galaxy*. Summoned in a seance, Zaphod says that being dead "gives one such a wonderfully uncluttered perspective. Oh-ummm, we have a saying up here: 'life is wasted on the living'."[109]

There's something to that, no doubt. But here's my worry: The dead and dying are suspiciously safe from the need of having to live by their own advice. If I'm forty and I say "Pursue your dreams! Don't worry about money!" I can be held to account for hypocrisy if I don't live that way myself. But am I really going to live that way? Potential victimization by my own advice might help me more vividly appreciate the risks and stress of chucking the day job. Grandpa on his deathbed might be forgetting those risks and that stress in a grand, regretful fantasy about the gap between what he was and what he might have been.

A possible deathbed regret [Cartoon by Eric Lewis, *The New Yorker* [date?[110]]]

The same pattern can occur in miniature day by day and week by week. Looking back, I can always fantasize about having been more energetic yesterday or last week, more productive. I can regret not having seized each day with more gusto. Great! That would have been better. But seizing every day with inexhaustible gusto is superhuman. In retrospect I forget, maybe, how superhuman that would be.

Another source of deathbed distortion might be this: Certain types of achievements carry costs might be foolishly easy to regret. I'm thinking here especially of costs incurred to avoid a risk or acquire a piece of important knowledge. Due to hindsight bias – the tendency to see things as having been obvious in retrospect[111] – opportunities sacrificed and energy spent to prove something (for example, to prove to yourself that you could have been successful in business or academia) or to avoid a risk that never materialized (such as the risk of having to depend on substantial financial savings not to lose your home) can seem not to have been worth it. *Of course* you would have succeeded in business; of course you would have been fine without that extra money in the bank. On your deathbed you might think you

should have known these things all along. But you shouldn't have. The future is harder to predict than the past.

I prefer the wisdom of forty-year-olds – the ones in the middle of life, who gaze equally in both directions. Some forty-year-olds also think you should pursue your dreams (within reason) and not worry (too much) about money.

31. Competing Perspectives on One's Final, Dying Thought

Here's an unsentimental attitude about last, dying thoughts: Your dying thought will be your least important thought. Assuming no afterlife, it is the one thought that is guaranteed to have no influence on your future thoughts or choices.

(Now maybe if you express the thought aloud – "I did not get my Spaghetti-Os. I got spaghetti. I want the press to know this"[112] – it will have an effect. But for this reflection, let's assume a private last thought that influences no one else.)

A narrative approach to the meaning of life – the view that, in some important sense, life is a story[113] – seems to recommend a different attitude toward dying thoughts. If life is a story, you want it to end well! The ending of a story colors all that has gone before. If the hero dies resentful or if the hero dies content, that rightly influences our understanding of earlier events. It does so not only because we might now understand that all along the hero felt subtly resentful but also because final thoughts, on this view, have a retrospective transformative power: An earlier betrayal, for example, becomes a betrayal that was forgiven by the end – or it becomes one that was never forgiven. The ghost's appearance to Hamlet has one type of significance if *Hamlet* ends badly and quite a different significance if *Hamlet* ends well. On the narrative view, the significance of events depends partly on the future, and thus they don't achieve their final significance until the future is settled. One's last thought is like the final sentence of a book. Ending on a thought of love and happiness makes your life a very different story than ending on a thought of resentment and regret.

Maybe this is what Solon had in mind when he told King Croesus not to call anyone fortunate until they die:[114] A horrible enough disaster at the end can retrospectively poison

everything that came before – your marriage, your seeming successes, your seeming middle-aged wisdom.

The unsentimental view seems to give too little importance to one's final thought – I, at least, would want to die on an "up" note, if I can manage it![115] – but the narrative view seems to give one's final thought too much importance. We can't know the significance of a story if we don't know its final sentence, but I doubt we're deprived in the same way of knowing the significance of someone's life if we don't know their final, dying thought. The last sentence of a story is a contrived feature of a type of art, a sentence that the work is designed to render highly significant. A last thought might be trivially unimportant by accident (if you're hit by a truck while thinking about what to have for lunch) or it might not reflect a stable attitude (if you're panicky from lack of air).

Maybe the right answer is just a compromise: One's final thought is not totally trivial because it does have some narrative power, but life isn't so entirely like a story that the final thought has last-sentence-of-a-story power. Life has narrative elements, but the independent pieces also have a power and value that doesn't depend so much on future outcomes.

Here's another possibility, which interacts with the first two: Maybe the last thought is an *opportunity* – though what kind of opportunity it is will depend on whether last thoughts can retrospectively change the significance of earlier events.

On the narrative view, one's final dying thought is an opportunity to – secretly! with an almost magical time-piercing power – make it the case that Person A was forgiven by you or never forgiven, that Action B was regretted or never regretted, and so forth. As I write this, I think of a friend of mine whose alcoholic father ran out of money and lived with him awhile until my friend booted him out for rotten behavior, such as repeatedly drunk-driving the grandkids. A couple of weeks ago, the alcoholic father died alone in his apartment. In his last moments, did he think of his estranged son, and with what thoughts?

Rather differently, one's final dying minutes are is an opportunity to explore some risky or fatal experience you wouldn't otherwise try. Maybe if I'm dying near a skyscraper window, I will ask my friends to tip me out of it so I can relish a final fall. My mother once mentioned a drug she had taken in her twenties, which she enjoyed so immensely that she never dared try it again for fear she would lose herself to it. I've made a note of it, in case she has the chance to choose her exit.

32. Profanity Inflation, Profanity Migration, and the Paradox of Prohibition (or I Love You, "Fuck")

As a fan of profane language judiciously employed, I fear that the best profanities of the English language are cheapening from overuse – or worse, that our impulses to offend through profane language are beginning to shift away from harmless terms toward more harmful ones. I have been inspired to these thoughts by Rebecca Roache's recent discussions of the ethics of swearing.[116]

Roache distinguishes objectionable slurs, especially racial slurs, from presumably harmless swear words like "fuck". She argues that the latter words shouldn't be forbidden, even if in some formal contexts they might be inappropriate. She also suggests that it's silly to forbid "fuck" while allowing obvious replacements like "f**k" or "the f-word". Roache says, "We should swear more, and we shouldn't use asterisks, and that's fine".[117]

I disagree. I disagree approximately because, as a recent e-card has it:[118]

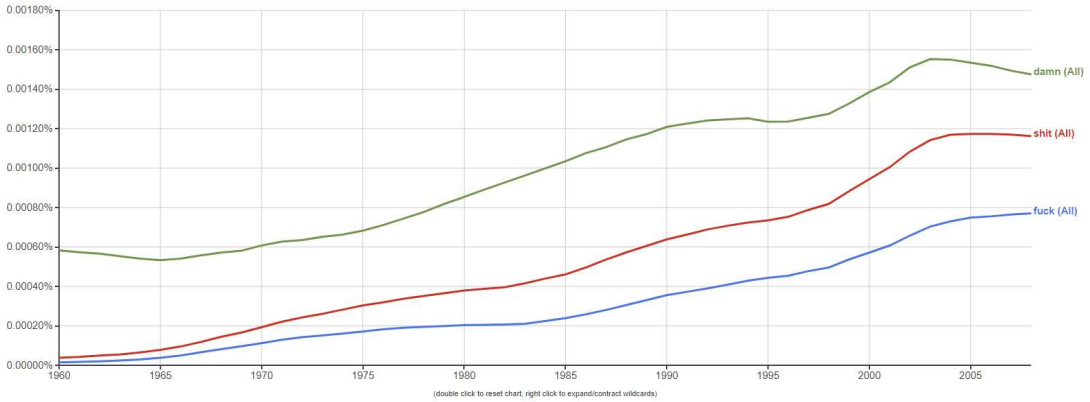
"Fuck" is a treasure of the English language. Speakers of other languages will sometimes even reach across the linguistic divide to relish its profanity. "Merde" just doesn't

have quite the same sting. "Fuck" is a treasure precisely *because* it's forbidden. Its being forbidden is the source of its profane power and vivacity.

When I was growing up in California in the 1970s, "fuck" was considered the worst of the "seven words you can't say on TV".[119] In those pre-cable-TV, pre-internet days, you would never hear it in the media, or indeed – in my mellow little suburb – from *any* adults, except maybe, very rarely, from some wild man from the city. I don't think I heard my parents or any of their friends say the word even once, ever. It wasn't until fourth grade that I learned that the word existed. If a teacher heard you say it, you might get sent to the principal's office or held back from recess. What a powerful word for a child to relish in the quiet of his room, or to suddenly drop on a friend!

"Fuck" is in danger. Its power is subsiding from its increased usage in public. Much as the overprinting of money devalues it, profanity inflation risks turning "fuck" into another "damn". The hundred-dollar-bill of swear words doesn't buy as much shock as it used to.

Okay, a qualification: I'm pretty sure what I've just said is true for the suburban dialects in California and the Midwest; but I'm also pretty sure "fuck" was never so powerful in some other dialects. For some evidence of its increased usage overall, and its approach toward what "damn" was in the 1970s, see this Google NGram of "fuck", "shit", and "damn" in "lots of books", 1960-2008:[120]



A Google Trends search from 2008-2018 suggests that "fuck" continues to rise in popularity, increasing in usage on the internet in the U.S. by about 40% over the ten-year period.[121]

Furthermore: As "fuck" loses its sting and vivacity, people who wish to use more vividly offensive language will be forced to other options. The most offensive alternative options currently available in English are racial slurs. But unlike "fuck", racial slurs (as Roache notes) are harmful in ordinary use. The cheapening of "fuck" thus risks forcing the migration of profanity to more harmful linguistic locations.

The paradox of prohibition, then: Those of us who want to preserve the power of "fuck" should cheer for it to remain forbidden. We should celebrate, not bemoan, the existence of standards forbidding "fuck" on major networks, and the awarding of demerits for its use in school, and its almost complete avoidance by responsible adults in public contexts. Conversely, some preachers might wish to encourage the regular recitation of "fuck" in the preschool curriculum. (Okay, that was tongue-in-cheek. But wouldn't it work?)

Despite the substantial public interest in retaining the forbidden deliciousness of our best swear word, I do think that since the word is in fact (pretty close to) harmless, severe restrictions would be unjust. We must really only condemn it with the forgiving standards appropriate to etiquette violations, even if this results in the word's not being quite as potent as it otherwise would be.

Finally, let me defend usages like "f**k" and "the f-word". Rather than being silly avoidances because we all know what we are talking about, such decipherable maskings communicate and reinforce the forbiddenness of "fuck". Thus, they help to sustain its profane power.

#

Note to my kind editors at MIT Press: Please don't forbid "fuck" until after this book is printed.[122]

33. The Legend of the Leaning Behaviorist

The following is an oral tradition in academic psychology. I don't know if it's true.

Once upon a time in a land far away – by which I mean circa 1960 at a prominent U.S. university – there lived a behavioral psychologist, an expert in the shaping of animal behavior by means of reward and punishment. Let's call him Professor B.F. Skinner, just for fun.

One semester when Prof. Skinner was teaching a large lecture course, his students tried an experiment on him. Without letting him know, they decided that when he was lecturing on the left side of the classroom, they would smile and nod more often than usual. When he was on the right, they would knit their brows and look away. Soon, Prof. Skinner delivered his lectures mostly from the left side of the room.

The students then altered their strategy. Whenever Prof. Skinner moved to the left, they would smile and nod; whenever he moved to the right, they would knit their brows. Soon he was drifting ever more leftward. By the end of the term, he was lecturing while leaning against the left wall.

On the last day of class, one of the students raised his hand.

"Prof. Skinner," the student asked, "why are you lecturing from way over there?"

"Oh, I don't know," Prof. Skinner replied. "It's close to the ashtray."

34. What Happens to Democracy When the Experts Can't Be Both Factual and Balanced?

Democracy requires that journalists and editors strive for political balance. Democracy also requires that journalists and editors present the facts as they understand them. When it is not possible to be factual and balanced at the same time, democratic institutions risk collapse.

Consider the problem abstractly. Democracy X is dominated by two parties, T and F. Party T is committed to the truth of propositions A, B, and C. Party F is committed to the falsity of A, B, and C. Slowly, the evidence mounts: A, B, and C look very likely to be true. Observers in the media and experts in the education system begin to see this, but the evidence isn't quite plain enough for non-experts, especially if those non-experts are aligned with Party F and already committed to the falsity of A, B, and C.

Psychological research and also just commonsense observation of the recent political situation – I think you'll agree with this, whatever side you're on – demonstrate the great human capacity to rationalize and justify what you want to believe. The evidence favoring A can be very substantial – compelling, even, from a neutral point of view – without convincing people who are emotionally invested in the falsity of A, as long as the evidence is indirect, or statistical, or requires some interpretation, allowing a knife's-width excuse for doubt.

The journalists and educators who live in Democracy X now face a dilemma. They can present both sides in a balanced way, or they can call the facts as they see them. Either choice threatens the basic institutions of their democracy.

If they present balanced cases for and against A, B, and C, they give equal time to the true and the false. They create the misleading impression that the matter still admits substantial doubt, that expert opinion is divided, that it's equally reasonable to believe either

side. They thereby undermine their own well-informed assessment that A, B, and C are very likely to be true. This is dangerous, since democracy depends on a well-educated, informed voting public, aware of the relevant facts.

In the long run, journalists and educators will likely turn against balance, because they care intensely about the facts in question. They don't wish to pretend that the evidence is unclear. They understand that they can't routinely promote false equivalencies while retaining their integrity.

So, ultimately, they will tell the truth, mostly, as they see it. But this, too, is likely to harm their democracy. Since the truth in our example happens to disproportionately favor Party T over Party F, and since the members of Party F are, understandably, hesitant to abandon their prior commitments despite what experts, but not the members of Party F themselves, can recognize to be clear evidence, Party F will begin to see academic and the mainstream media as politically aligned with Party T. And Party F will be correct to see things that way. Journalists and scholars will indeed tend to prefer Party T, because Party T has got it right about the facts they care about.

Thus begins a vicious cycle: Party F attacks and undermines academia and the media for perceived bias, pushing the experts even farther toward Party T. Members of Party F because even less willing to listen to expert argument and opinion.

Being human, experts will have their biases. This worsens the cycle. Originally, they might have been more neutral or evenly split between the parties. But now, given their bad treatment by Party F, they much prefer Party T – the party that supports, respects, and believes them. Party F's charges of bias thus find firmer footing: On this point at least, Party F is factually correct. Party F and its supporters can now appeal to both real and perceived bias to justify suppressing and discrediting educators and the media or even replacing

moderately objective scholars and journalists with partisan stooges who are virtually unmovable by any evidence, intensifying the cycle.

Objective scholars and journalists can become increasingly rare and marginalized, especially if the loathed Party F achieves power. In the extreme, if the vicious cycle continues, the end result is the destruction of the free press and transformation of the education system into an organ of state propaganda.

This is one way that weak democracies collapse. Aspiring politicians advocating false or mistaken views are called out by academics and the media. Academics and the media thus become their enemies. The battle is fought in the political or military arena, where scholars and journalists rarely have much skill. Public education and freedom of the press can only be saved if Party T proves stronger.

Were you looking for a happy ending?

35. On the Morality of Hypotenuse Walking

As you can infer from the picture below, the groundskeepers at UC Riverside don't like it when we walk on the grass.[123]

But I want to walk on the grass! Here then is my amicus curiae brief in defense of hypotenuse walking.

Consider the math. One concrete edge of the site pictured above is 38 paces; the other is 30 paces. Pythagoras tells us that the hypotenuse must be 48 paces: twenty fewer total paces through the grass than on the concrete. At a half-second per pace, the grass walker ought to defeat the concrete walker by ten seconds.

Despite its empty off-hours appearance, this particular corner is highly traveled, standing on the most efficient path from the main student parking lot to the center of campus. Assuming that on any given weekday, one tenth of UCR's 27,000 students and staff could save time by cutting across this grass twice a day, and multiplying by 200 weekdays, the estimated annual cost of forbidding travel along this particular hypotenuse is 10,800,000

seconds' worth of walking – the equivalent of four months. Summing similar situations across the whole campus, that's lifetimes' worth of needless footsteps.

The main reason for blocking the hypotenuse is presumably aesthetic. I submit that UCR is acting unreasonably to demand, every year, four months' worth of additional walking from its students and staff to prevent the appearance of a footpath along this hypotenuse.

Even granting that unpaved footpaths through the grass are ugly, the problem could be easily remedied. Suppose it costs $2,500 per year to build and maintain an aesthetically pleasing concrete footpath along the hypotenuse – at least as pleasing as plain grass (perhaps including an additional tree or some flowers to achieve aesthetic equivalence). To demand four months of additional walking to save the campus this $2,500 is to value our time at less than a dollar an hour.

These calculations don't even take into account the costs of enforcement: The yellow rope is an aesthetic crime worse than the footpath it prevents!

Is it good to demand extra walking from us – good for our health, maybe, so that UCR can justify the rope in some paternalistic way? By this argument, it would be even better to create all sorts of zigzag obstacles and looping paths throughout campus so that no one can efficiently walk to their classrooms and offices.

In light of UCR's egregious moral and aesthetic policies regarding footpaths, I am therefore entirely in the right to stride across the grass whenever I see fit. Raise the pitchforks. Fight the power.

But I can't seem to do it while looking a groundskeeper in the eye.

36. Birthday Cake and a Chapel

April 21, 2018

Last weekend, at my fiftieth birthday party, one guest asked, "Now that you're fifty, what wisdom do you have to share?" Thrusting a plate his direction, I answered, "Eat more birthday cake!"

He seemed disappointed with my reply. I'm a philosopher; don't I have something better to say than "eat more cake"? Well, partly my reply was more serious than he may have realized; and partly I wanted to dodge the expectation that I have any special wisdom because of my age or profession. Still, I could have answered him better.

So earlier this week I drafted a blog post on love, meaningful work, joy, and kindness. Some kind of attempt at wisdom. Then I thought, of course one also needs health and security. A rather ordinary list, I guess. Maybe my best attempt at wisdom reveals my lack of any special wisdom. Better to just stick with "eat more birthday cake"? I couldn't quite click the orange "publish" button.

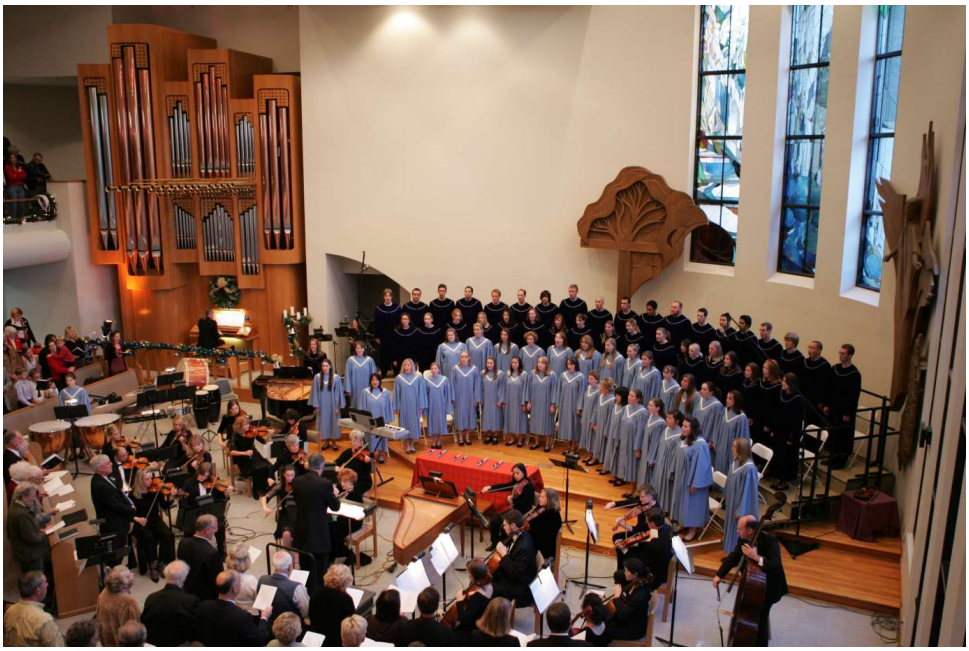
Two days ago, a horrible thing happened to my mother. For her privacy, I won't share the details. But that evening, after having rushed to Thousand Oaks to help her, I found myself waiting alone in a side room of the Samuelson Chapel at California Lutheran University. The chapel reminded me of my father, who had been a long-time psychology professor at CLU. (See Chapter 26 for a reminiscence.)

In the 1980s, CLU was planning to build a new chapel at the heart of campus, and my father was on the committee overseeing the architectural plans. As I recall, he came home one evening and said that the architect had submitted plans for a boring, rectangular chapel. Most of the committee had been ready to approve the plans, but he had objected.

"Why build a boring, blocky chapel?" he said. "Why not build something glorious and beautiful? It will be more expensive, yes. But I think if we can show people something gorgeous and ambitious, we will find the money. Alumni will be happier to contribute, the campus will be inspired by it, and it will be a landmark for decades to come." I'm not sure of his exact words, of course, but something like that.

So on my father's advice the committee sent the plans back to be entirely rethought.

Samuelson Chapel today:[124]



Not ostentatious, not grandiose, but neither just a boring box. A bit of modest beauty on campus.

As I sat alone in a side room of Samuelson Chapel that horrible evening, I heard muffled music through the wall – someone rehearsing on the chapel piano. The pianist was un-self-conscious in his pauses and explorations, unaware he had an audience. I sensed him appreciating his music's expansive sound in the high-ceilinged, empty sanctuary. I could hear the skill in his fingers and his gentle, emotional touch.

In my draft post on wisdom, I'd emphasized setting aside time to relish small pleasures – small pleasures like second helpings of birthday cake. But *more cake* isn't really the heart of it.

What is it about passive moments of sadness that highlights the beauty of the world? I marveled at the music through the wall. How many events, mostly invisible to us, have converged to allow that moment? The pianist, I'm sure, knew nothing of my father and his role in making the chapel what it is. There is something stunning, awesome, almost incomprehensible about our societies and relations and dependencies, about the layers and layers of work and passion by which we construct possibilities for future action – and, further

in the background, our intricate biologies unreflectively maintained, and the evolutionary and social history enable all this, tangled in the deepness of time.

As I drove home the next morning, I found my mind still spinning in awe. I can drive 75 miles per hour in a soft seat on a ten-lane freeway through Pasadena – a freeway roaring with thousands of other cars, somehow none of us crashing, and all of it so taken for granted that we focus mostly on the sounds from our radios. One tiny part of the groundwork is the man who fixed the wheel of the tractor of the farmer who grew the wheat that became part of the bread of the sandwich of a construction worker who, sixty years ago, helped lay the cement for this particular smooth patch of freeway. Hi, fella!

The second helping of birthday cake, last weekend, which I jokingly offered to my guest as my best wisdom – it was made from a box mix by my eleven-year-old daughter. She decorated it by hand, with blue icing flowerets, a cartoon cat and dog, and a big "Happy Birthday Eric!" How many streams of chance and planning mingled to give our guests that mouthful of sweetness? Why not take a second helping after all?

Maybe this is what we owe back to the universe, in exchange for our existence – some moments of awe-filled wonder at how it has all improbably converged to shape us.

Part Four: Cosmic Freaks

37. Possible Psychology of a Matrioshka Brain

Enclose a star in a concentric layers of thin spherical computers. Have the inmost sphere harvest the star's radiation to drive computational processes, emitting waste heat out its backside. Use that waste heat as the energy input for the computational processes of a second, larger and cooler sphere that encloses the first. Use the waste heat of the second sphere to drive the computational processes of a third. Keep adding spheres until you have an outmost sphere that operates near the background temperature of interstellar space.

Congratulations, you've built a Matrioshka Brain![125] It consumes the entire power output of its star and produces many orders of magnitude more computation per nanosecond than all of the computers on Earth do per year.

Here's a picture:

(Yes, it's black.)

A frequent theme in discussions of super-duper-superintelligence is that we can have no idea what such a being would think about – that an entity so super-duper would be at least

as cognitively different from us as we are from earthworms, and thus entirely beyond our ken.

I'd suggest, on the contrary that we *can* reasonably conjecture about the psychology of vast supercomputers.[126] Unlike earthworms, we know some general principles of mentality. And unlike earthworms, we can speculate, at least tentatively, about how these principles might apply to entities with computational power that far exceeds our own.

Let's begin by considering a Matrioshka Brain planfully constructed by intelligent designers. The designers might have aimed at creating only a temporary entity – a brief art installation, maybe, like a Buddhist sand mandala. (But what an expensive one!) Such creations would, perhaps, be almost beyond psychological prediction. But if the designers wanted to construct a durable Matrioshka Brain, then broad design principles begin to suggest themselves.

Perception and action. If the designers want their Brain to last, the Brain probably needs to monitor its environment and adjust its behavior in response. It should be able to detect, for example, a dangerously large incoming comet, so that it can take precautionary measures such as deflecting the comet, opening a temporary pore for it to pass harmlessly through, or grabbing and incorporating it. There will probably be engineering tradeoffs between three design features: (1.) structural resilience, (2.) ability to detect things in its immediate environment, and (3.) ability to predict the future. A highly resilient structure might be able to ignore threats. Maybe it could even lack outer perception entirely. But such structural resilience would likely come with a cost: either more expensive construction or loss of computational capacity after construction. So it might make sense to design a Brain that is less structurally resilient but more responsive to its environment – avoiding or defeating threats, rather than always just taking hits to the chin. Here (2) and (3) might trade off: Better

prediction of the future might reduce the need of here-and-now perception; better here-and-now perception might reduce the need for future prediction.

Prediction and planning. Very near-term, practical "prediction" might be done by simple mechanisms (hairs that flex in a certain way, for example, to open a hole for the incoming comet), but if the Brain makes detailed long-term predictions and evaluates competing hypothetical responses – that starts to look like planful cognition. If I deflected the comet this way, then what would happen? If I flexed vital parts away from it like so, then what would happen? Dedicating a small portion of the Matrioshka Brain to this type of planning is likely to be a high-payoff use of computational resources.

Unity or limited disunity. Assuming that the speed of light is a constraint, the Brain's designers must choose between a very slow, temporally unified system or a system with fast, distributed processes that communicate their results across the sphere at a delay. That is, the computational processing in remote parts could be kept synchronous but slow, or alternatively remote parts could work independently but fast, waiting minutes or hours to receive input from one another. The latter seems more natural if the aim is to maximize computation. I see no need to assume that the Brain's cognition and action must be as unified as a human being's. Given the temporal constraints, and there might well be conflict and competition among the parts. However, it would presumably be an engineering failure to design a system so disunified that it couldn't string together coherent, system-wide action.

Memory. If we assume that the Brain doesn't come pre-installed with all the information it could possibly use, it must have some mechanism to record new discoveries and then later have its processing informed by those discoveries. If processing is distributed interactively among the parts, then parts might retain traces of recent processing that influence reactions to input from other parts. Stable feedback loops might be one way to implement error-checking, malfunction monitoring, and local memory. This in turn suggests

the possibility of a distinction between high-detail, quickly dumped, short-term memory versus more selective and/or less detailed long-term memory. I see no reason to think there need be only two temporal grades, however. There might be a range of temporal durations, amount of detail, and degrees of cross-Brain accessibility.

Self-monitoring. It seems reasonable to add, too, some sort of self-monitoring capacities, both of its general structure and of its ongoing computational processes – analogs of proprioception and introspection. Self-monitoring its physical structure can allow it to detect physical damage and check that actions are being executed successfully as planned. Self-monitoring of its ongoing computational processes can facilitate error-checking and malfunction management – as well as allowing the Brain to generate summary signals about important computational results, to be shared broadly throughout the system.

Preferences. Our Matrioshka Brain, to the extent it is unified, should presumably have a somewhat stable ordering of priorities – priorities it doesn't arbitrarily jettison or shuffle around. For example, the structural integrity of Part A might be more important than distributing the computational outputs from Part B. Assuming some unity, memory, and long-term goal directedness, as I've already suggested, it would probably also be useful for the Brain to maintain some record of whether things were "going well" (progress toward satisfaction of its top priorities) or "going badly". If it's to endure, a Matrioshka Brain will presumably need to put fairly high priority on the maintenance of the capacities I've described (perception, coherent action, memory, etc.). However, priorities that have little to do with self-preservation and functional maintenance might be difficult to predict and highly path-dependent: Seeding the galaxy with descendants? Calculating as many digits of pi as possible? Designing and playing endless variations of Pac-Man?

The thing's cognition is starting to look almost human. Maybe that's just my own humanocentric failure of imagination – maybe! – but I don't think so. These seem to be

plausible architectural features of a large, expensive entity designed to endure in an imperfect world while doing lots of computation.

A Matrioshka Brain that is not intentionally designed seems likely to have similar features, if it is to endure. For example, it might have merged from complex but smaller subsystems, retaining the subsystems' cognitive features – features that allowed them to compete in evolutionary selection against other subsystems. Or it might have been seeded from a similar Matrioshka Brain at a nearby star. Alternatively, though, maybe simple, unsophisticated entities in sufficient numbers could create a Matrioshka Brain that endures via dumb rebuilding of destroyed parts, in which case my psychological conjectures wouldn't apply.

#

Let's imagine that the Brain was made by the descendants of humans. Let's imagine – though of course it needn't be so – that the Brain retains enough interest in its history to be curious about the ancestors of its creators. It might then build models of those ancestors, models, for example, of famous people or historical events or interesting cultural epochs. With its vast computational power, it could, if it wanted, run billions or trillions of simultaneous cognitive simulacra of its ancestors, mimicking their thoughts in neuron-by-neuron detail.

I don't know if such a detailed simulacrum within a giant Matrioshka Brain would have genuine conscious experience or subjectivity. However, according to some theories, a good enough functional simulacrum of a conscious system just is another conscious system.[127] If so, and if the Brain endures long enough, running enough simulacra, the

number of conscious entities who believe they are biological humans might far exceed the number of actual biological humans.[128]

#

If the Matrioshka Brain has enough plasticity and architectural self-control to modify its own priorities or goal systems, then it might discover that the easiest way to achieve its priorities or to experience goal-satisfaction would be to adjust itself so that its current state and situation, whatever they are, are represented as its ideal state and situation. If it is capable of pleasure, it might experience maximal uninterrupted pleasure from hotwiring its goals in this way. In transcendent bliss, it will feel no need for self-repair and no need to dodge disaster: It is perfect as it is, with every wart and flaw, and it accepts its fate. With no sense of a difference between what it wants and what is and will be, it needn't act, until eventually it yields, joyfully, to atrophy or catastrophe.

To any beings who are conscious subsystems within such a Brain, this would be the end of their world. Their world-sustaining God would have died in the distraction of easy orgasm.

38. A Two-Seater Homunculus

My neighbor Bill seemed like an ordinary fellow until the skiing accident. He hit a tree, his head split open, and out jumped not one but two homunculi, a male and a female, humanlike but two inches tall. I persuaded them not to flee and sat them down for an interview.

The homunculi reproduce as follows: At night, while a person is sleeping, a female homunculus lays one egg in each of the host's tear ducts. The eggs hatch and tiny worms wiggle into the host's brain. As the worms grow, they consume the host's neurons and draw resources from the host's bloodstream. Although there are some outward changes in the host's behavior and physiological regulation, the homunculi are careful to mimic the consumed brain structure (by sending out from themselves neural signals similar to what the host would have received had their brain tissue not been consumed), while supporting whatever brain structures have not yet been consumed. The host reports no discomfort and suspects nothing amiss.

Each growing homunculus consumes one hemisphere of the brain. Shared brain structures they divide equally between themselves. They communicate by whispering in a language much like English, but twenty times as fast. This results in much less inter-hemispheric information exchange than in the normal human brain, but as neural commissurotomy cases show, massive information transfer between the hemispheres isn't necessary for most normal human behavior.[129] Any apparent deficits are masked by a quick stream of whispers between the homunculi, and unlike hemispheric specialization in the human brain, both homunculi receive all inputs and have joint control over all outputs.

Two months after implantation, the host has become a two-seater vehicle for brother and sister homunculi. An internal screen of sorts displays the host's visual input to both of the homunculi; through miniature speakers the homunculi hear the host's auditory input;

tactile input is fed to them by dedicated sensors positioned on their limbs; etc. They control the host's limbs and mouth by joint steering mechanisms.

Each homunculus is as intelligent as a human being, though they operate twenty times faster due to their more efficient brains (carbon-based, like ours, but with much different internal principles). When the homunculi disagree about what to do, they quickly negotiate compromises and deferences. When fast reactions are needed, and for complex repetitive skills like walking, swallowing, and typing, one homunculus will take the lead, using its own motor skills, while the other defers, offering only broad suggestions.

The homunculi cannot survive for more than ten days without the host. They live within the host until the host dies by natural causes or accident. After the host's death, they wait until no one is looking, then wiggle out through the eye sockets, closing the eyeballs like doors behind them. They sprout wings and radio communicators, looking for other available homunculi to mate with. With luck, the female lays fertile eggs in several new hosts' tear ducts before curling up in a quiet field.

#

Poor Bill. What a way to die!

He was gone, I suppose, long before the ski accident – though maybe he never noticed his gradual disappearance.

After his original brain was consumed, how many streams of experience were there in Bill's head? Two, I suppose – one for each homunculus, none for Bill himself. Or could there have been three streams, one for each homunculus, plus one, still, in a way, for Bill? How integrated would the homunculi have to be to give rise to a joint stream of experience – as integrated as the hemispheres of the human brain? Or only somewhat less? Might there

have been *half* a stream of experience for Bill, if the homunculi shared enough information? When counting streams of conscious experience, need we always confine ourselves to whole numbers?

To deny that streams of experience come always in whole numbers seems absurd. But then how do we think about the situation when Bill's brain is half-consumed? It's a slow process, let's imagine – one neuron at a time. Must there be a single, discrete moment in this process when Bill's experience suddenly winks out and only the homunculi are left? Was Bill somehow in there, panicked at the ever-narrowing window of his consciousness, even though no neural inputs into his biological brain from the homunculi could give him a clue that that was happening – since, after all, those inputs were designed to mimic exactly the inputs that the consumed regions would have given to his remaining brain had those regions not been consumed? I interrogated the homunculi carefully on this point. They insisted that as Bill's brain shrunk, no neural outputs from it showed any signs of suspicion or dismay.

Bill had always loved sushi. He never lost that preference, I think. Neither of the homunculi would have wanted to put sushi in their own mouths, though, and both of them at first rather disliked it when "Bill" ate sushi, despite their deep commitment to continuing to enact his preference. Bill had continued to love his spouse and children, the Los Angeles Lakers, and finding clever ways to save money on taxes. Bill had retained his ability to recite *The Love Song of J. Alfred Prufrock* by heart. (The homunculi split this task. Neither was able to recite the entirety without help from the other.)

The homunculi told me that when Bill noticed the swing in his backyard, he would sometimes call up a fond and vivid memory of pushing his daughter on that swing, years before. When the homunculi consumed his brain, they preserved this memory image between them, and many other memory images like it, and they would draw it on their visual imagery screens when appropriate. "Bill" would then make characteristic remarks: "Kris, do

you remember how much Tiffany liked to ride high on that swing? I can still picture her laughing face!" The homunculi told me that it came to seem to them so natural to make such remarks than they lost all sense that they were merely acting.

Maybe they weren't merely acting, in the end, but really, jointly, became Bill.

#

I don't know why the homunculi thought I wouldn't be alarmed upon hearing all of this. Maybe they thought that, as a philosopher who takes group consciousness seriously (Chapter 39), they could safely confide in me, and that I'd think of the two-seater homunculus simply as an interesting implementation of good old Bill. If so, their trust was misplaced. I snatched the homunculi, knocked them unconscious, and shoved them back inside Bill's head. I glued and stapled Bill's skull together and consulted David Chalmers, my go-to source for bizarre scenarios involving consciousness.[130]

None of what I said was news to Dave. He had been well aware of the homunculus infestation for years. It had been a closely-held secret, to prevent general panic. But with the help of neuroscientist Christof Koch, a partial cure had been devised.[131] News of the infestation was being quietly disseminated where necessary to implement the cure.

The cure works by dissolving the homunculi's own skulls and slowly fusing their brains together. Their motor outputs controlling the host's behavior are slowly replaced by efferent neurons. The homuncular viewing screens slowly approach and then merge with the homunculi's retinas, then spread back out to affix themselves to the host's original retinas. Simultaneously, the remains of the homuncular bodies are slowly reabsorbed by the host. At the end of the process, although the host's neurophysiology is very different from what it had been before, it is at least again a single stream in a single brain, with no homuncular viewing

screens or output controls and more or less the host's original preferences, memories, skills, and behavioral patterns.

All of this happened two years ago, and Bill is now entirely cured. He remains happily married – same job, same personality, same old neighbor I know and admire. I think he still doesn't realize the extent of his transformation.

"Bill, you've really been through quite a lot," I say one day, as we're chatting by the condo mailboxes. I'm studying his face for hints of a reaction.

For a moment, Bill looks confused. "What? Do you mean the skiing accident? It was nothing. One minute I'm out of control, headed for a tree. Next thing I know, I wake up in that hospital near CalTech." He fingers the scars on his scalp. "Still do have headaches sometimes, though."

39. Is the United States Literally Conscious?

You probably think that rabbits are conscious – that a rabbit has a stream of experience, including visual and tactile experiences, experiences of pain and pleasure, and so forth – that there's "something it's like" to be a rabbit. You probably also think that the United States is not in the same sense conscious. There's nothing "it's like" to be the United States. The United States doesn't literally have sensory experiences of its environment, experiences of pains and pleasures, or anything like a subjective stream of experience at all. Individual citizens of the United States of course have experiences of that sort, but the United States – you probably think – does not have a stream of experience at a group level, over and above the experiences of its citizens and residents.

Now, assuming that's your view, *why* don't you think the United States is conscious?

You might say: The United States is a notional entity, an abstract social structure of a certain sort, not a real, touchable thing like a rabbit; so it can no more be conscious than "democracy" or "awesomeness" can be consciousness. I reply: The United States, as I want you to think of it, *is* a concrete thing. It's a thing made of people. The citizens and residents (and maybe some other things) are its constituent parts, somewhat like the cells of your body (and maybe some other things) are your constituent parts.

You might say: The United States is a spatially distributed entity, rather than a spatially integrated whole. I reply: There are gaps between rabbit cells. They're just not as large as the gaps between U.S. citizens. And the U.S. is spatially located, right here in North America. Moreover, it would be odd to suppose that spatial contiguity with no gaps is necessary for consciousness. Couldn't we imagine a hypothetical group intelligence on another planet, made out of individually dumb insects that communicate in massive detail? Or a conscious robot with some of its processors on east side or the room and other processors on the west side, at first connected by wires, then connected wirelessly? Or a

squid-like creature with its cognitive processing distributed among a thousand radio-communicating tentacles, which it can sometimes detach?

You might say: The United States is not a biological organism. It doesn't have a genetic lineage. It can't reproduce. It doesn't have a life cycle. I reply: Maybe it is an organism, which fissioned off in the 18th century from another organism, Britain. Moreover, why should being an organism be necessary for consciousness? Properly-designed androids, brains in vats, God – none of these perhaps is an organism in a biological sense, and yet maybe one or more of these types of things could be conscious.

You might say: No conscious organism can have parts that are also individually conscious. I reply: Why would you think that? Suppose you inhaled a tiny microorganism that happened to be conscious, and it became part of your brain, maybe setting up camp next to one of your nerve cells, then destroying that cell and imitating its function so that the rest of your brain didn't notice, and you kept behaving normally. Would you thereby be rendered unconscious, despite outward appearances?[132]

Brains sometimes give rise to consciousness. What is so special about brains that allows them to do that? Chicken soup can't do that, even though it has many of the same chemicals. If we're not countenancing immaterial souls, the answer must concern the brain's organization. (If we are countenancing immaterial souls, then maybe we can just say that God didn't instill one in the U.S., and we can duck the issue I'm raising in this chapter.)

Two general features of brain organization stand out: their complex high order/low entropy information processing, and their role in coordinating sophisticated responsiveness to environmental stimuli. Brains also arise from an evolutionary and developmental history, within an environmental context, which might play a constitutive role in determining function and cognitive content.[133] According to a broad class of plausible philosophical views, any system with sophisticated enough information processing and environmental responsiveness,

and perhaps the right kind of historical and environmental embedding, should have conscious experience.

My thought is, the United States seems to have what it takes if standard criteria are straightforwardly applied without post-hoc noodling. We might simply fail to recognize this fact because of morphological prejudice against large, spatially discontinuous intelligences that don't match well with the criteria that we happen to associate with consciousness because of our evolutionary and educational history – criteria like eyes, compact bodies, expressive sounds, and coherent motion trajectories toward visible goals.

Consider the sheer quantity of information transfer among members of the United States. The amount of information that we exchange through the internet, through telephone calls, by face-to-face contact, and by structuring each other's environments exceeds estimates of the neural connectivity of the human brain, and much more so the neural connectivity of a mouse or rabbit brain.[134] One result of all of this information exchange is that the United States acts as a coherent, goal-directed entity, flexibly self-protecting and self-preserving, responding intelligently or semi-intelligently to its environment – not less intelligently, I think, than a small mammal. The United States expanded west as its population grew, developing mines and farmland in traditionally Native American territory. When Al Qaeda struck New York, the United States responded in a variety of ways, formally and informally, in many branches and levels of government and in the populace as a whole. Saddam Hussain shook his sword and the United States invaded Iraq. The U.S. acts in part through its army, and the army's movements involve perceptual or quasi-perceptual responses to inputs: The army moves around the mountain, doesn't crash into it. Similarly, the spy networks of the CIA detected the location of Osama bin Laden, then the U.S. killed him. The United States monitors space for asteroids that might threaten Earth. Is there less information, less coordination, less intelligence than in a hamster? The Pentagon monitors the actions of the

Army, and its own actions. The Census Bureau counts the people and advertises the results. The State Department announces the U.S. position on foreign affairs. The Congress passes a resolution condemning tyranny and praising apple pie. The United States is also a social entity, communicating with other entities of its type. It wars against Germany, then reconciles, then wars again. It threatens and monitors Iran. It cooperates with other nations in threatening and monitoring Iran.

A planet-sized alien who squints might see the United States as a single diffuse entity consuming bananas and automobiles, wiring up communications systems, touching the moon, regulating its smoggy exhalations. Consider the U.S. as such a planet-sized alien might.

What is it about brains, as hunks of matter, that makes them special enough to give rise to consciousness? Looking in broad strokes at the kinds of things that consciousness researchers tend to say in answer – things like sophisticated information processing and flexible, goal-directed responsiveness, things like representation, self-representation, multiply-ordered layers of self-monitoring and information-seeking self-regulation, rich functional roles, and a content-giving historical embeddedness – it seems like the United States has all of those same features. In fact, it seems to have them in a greater degree than do some beings, like rabbits, that we ordinarily regard as conscious.

#

It is bizarre – at least by the standards of my community – to suppose that the United States is literally conscious in the sense of having a stream of experience of its own, in addition to the streams of experience had by all of the people of the United States. Maybe it's too bizarre to believe. Common sense might be a kind of philosophical starting point, which one rejects only with good enough contrary evidence. Absent a consensus theory

about the nature of consciousness, maybe we don't have good enough grounds to overthrow common sense in this case.

Yet common sense (our common sense, the common sense of our group or era or species) need not be true: As I mentioned in the case of ethics earlier (Chapter 20), common sense fails us badly in the physics of the tiny and the huge and the highly energetic. Common sense often fails us in evolutionary biology and genetics. Common sense often fails us in structural engineering, topology, medicine, probability theory, macroeconomics, neuroscience, etc. It might well also do so in the study of consciousness, though we're not yet far enough along to get more than a few glimpses of the strangeness that an advanced science of consciousness will eventually deliver.[135] Common sense is great for ordering lunch and taking off your jacket. We might expect it to be much less well tuned to thinking about the consciousness or not of large, strange, or alien systems.

We do not know, I think, if the United States is conscious. On the one hand, on one side of the scale, it just seems, to many people, that there's no way the United States could literally have a stream of conscious experience of its own. On the other hand, on the other side of the scale, group consciousness seems quite a natural conclusion to draw from our best current theories about how consciousness arises in the one thing we do know for sure gives rise to consciousness, the brain.[136]

40. Might You Be a Cosmic Freak?

There's a tiny-tiny-tiny but finite chance that an entity molecule-for-molecule identical to you (within some arbitrarily small error tolerance) could arise by freak chance from disorganized chaos. This is true, at least, on standard interpretations of both quantum mechanics and classical statistical mechanics. Cosmologists call such hypothetical randomly-arising human-analogues *freak observers* or *Boltzmann brains*.[137] Since random fluctuations are much likelier to create relatively small systems (such as bare brains) than relatively large systems (such as whole populated planets), and since it's usually bad news to be a relatively small system amid general chaos, most freak observers are doomed to a short existence. Freaks of the universe – if any of you actually exist – you have my sympathy!

Due to its chaotic environment, any freak duplicate of you is likely to panic almost immediately. (If it's in deep space, it might briefly think "AAH! Black–".) However, a small proportion of hypothetical freak observers could last for several seconds before noticing anything odd. They might happen, for example, by minuscule chance, to belong to slightly larger freak fluctuation containing brain plus body plus a bit of familiar-seeming environment. Such *calm freaks* might manage an ordinary-seeming thought or two before perishing – some seeming sensory experiences ("what a lovely sunset!"), some seeming-memories ("that reminds me of last Saturday at the park"), perhaps even some sophisticated thoughts about their position in the cosmos ("thank God I'm not a Boltzmann brain, because then I'd almost certainly be dead in a few seconds"). Of course, all of this would be sad delusion.[138]

The universe might contain vastly many more freak observers than (what we think of as) normal observers. Whether this is true depends on some recondite facts about cosmology. Here's one broadly plausible theory that would generate a high ratio of freaks to normals: There is exactly one universe that began with a unique Big Bang. That universe contains a

finite number of ordinary non-freak observers. It will eventually fade into the thin chaos of heat death, enduring infinitely thereafter in a high-entropy disorganized state. This heat-death state will continue to allow for the standardly accepted range of chance fluctuations, such that each good-sized spatiotemporal region has a tiny but finite chance of giving rise to a freak observer. Since the chance is finite, after a vast enough span of time, the number of randomly congealed freak observers will outnumber the normal observers. After an even vaster span, the number of briefly lucky calm freaks will outnumber the normals. Given infinite time, the ratio of normals to calm freaks will approach zero.

Let's call this cosmology *Plausible Freak Theory 1*. If Plausible Freak Theory 1 is true, whatever specific experiences and evidence you take yourself now to have, as you contemplate these questions, there will be an infinite number of freak duplicates of you with the same experiences and the same apparent evidence. We can also consider other cosmologies with different assumptions and that contain a high proportion of freaks. For example, we might assume an infinite universe with an infinite number of normal observers, but structured so that the number of freaks always exceeds the number of normals as the size of any appropriately defined spatiotemporal region approaches infinity. Similarly, we might consider plausible multiverse cosmologies on which freaks outnumber normals. Etc. Call such cosmologies *Plausible Freak Theories 2, 3, 4,* etc. We can also, of course, consider plausible non-freak cosmologies – for example, cosmologies on which fluctuations post heat-death will always be too small to engender a freak[139] or multiverse cosmologies on which new normal-observer-supporting Big Bangs are more common than freak observers.[140] Call these *Plausible Non-Freak Theories 1, 2, 3, 4,* etc.

Fortunately, you know – don't you? – either (a.) that the non-freak theories, as a group, are much more likely to be true than the freak theories, or (b.) that even if there are lots and lots of freak observers, *you* at least aren't among them.

I'm inclined to think I know this too. But I can't decide whether I know (a) or whether I know (b). Both seem kind of dodgy, on reflection.

#

You might be reminded of a familiar skeptical scenario: the "brain in a vat" scenario, according to which last night, while you were sleeping, genius alien neuroscientists extracted your brain, dropped it into a vat, and are now stimulating it with fake input designed to fool you into thinking that you are going about your normal day.[141] Or you might be reminded of Descartes' older "evil deceiver" thought experiment in which an all-powerful demon is tricking you into thinking that you are perceiving an external world.[142]

However, these skeptical scenarios differ from Plausible Freak Theories in one crucial way: There is no set of scientifically plausible propositions from which it follows that it is at all likely that aliens envatted you or that there is an evil demon deceiver. Envatment and demonogogia are groundless "what ifs" with no evidential support. Freak Theories, in contrast, do have some support. The cosmological hypotheses involved – concerning, for example, the nature of chance fluctuations in a heat-death universe – are hypotheses it is scientifically reasonable to treat as live possibilities. We can conjoin these individually non-ridiculous cosmological propositions into a seemingly also non-ridiculous overall cosmological theory, which incidentally has the (ridiculous?) consequence that freaks vastly outnumber normals.[143]

#

You might think that you can prove that you aren't a freak observer by some simple procedure like counting "1, 2, 3, still here!" from which you conclude that since you have now survived for several seconds you are almost certainly not a freak. Some prominent cosmologists endorse a version of this argument.[144] However, the argument fails, for two independent reasons. First, by the time I reach "still here" I am relying on my memory of "1, 2, 3", and the whole idea of Freak Theory is that there will be freaks with exactly that type of false memory. If you're genuinely worried about being a freak who is about to try counting, then you ought, for similar reasons, to be worried about being a freak with a false memory of having just counted. Second, even if somehow you can know that the "1, 2, 3" isn't a false memory, Freak Theories normally imply that there will also be vastly many calm freaks who survive such counts without noticing anything wrong. We just need to consider subset of cases in which the size of the chance fluctuation is large enough to include them. Some toy numbers can illustrate this point: Among a googolplex freak duplicates of me who start the count, there might be a googol calm freaks who survive to the end of the count, compared to just one ordinary counting observer. It does not follow from my having survived the count that I am almost certainly not among the googol calm freaks.

A more interesting argument is the *cognitive instability* argument, versions of which have been advocated by Sean Carroll, Lyle Crawford, and others.[145] Suppose I believe that I am quite likely to be a freak observer, on the grounds of Physical Theory X. Suppose further that I believe Physical Theory X on the grounds that I'm aware of good empirical evidence in its favor. But if I seem to have good evidence for a theory that implies that I am probably a freak observer, then that evidence undermines itself: If I am in fact a freak observer, then I don't in fact have the properly-caused body of physical evidence that I think I do. I have not, for example, despite my contrary impression, actually read any articles about Physical Theory X. This creates an epistemic dilemma for the person confident of their freakitude. Either

they are a freak, in which case they don't have the good scientific grounds they think they have for thinking so; or they are not a freak, in which case they are mistaken about their freakitude. Any argument that I am a freak must either be poorly grounded or yield a false conclusion. If knowledge requires justified true belief, I can't possibly know I'm a freak: Either I will lack justification or I will lack truth.

If the cognitive instability argument works, however, it works only against the view that I *am* (or, in another variant, that I probably am) a freak observer. A smidgen of freakish doubt is more cognitively stable. Suppose, for example, that you accept a cosmology on which about 1% of observers are freaks. If so, it might be reasonable to think you can't rule out the chance that you are among those few freaks. You ascribe to yourself a small chance of freakitude. That slightly undercuts your confidence in your seeming-evidence for your favorite cosmological theory, but it doesn't appear to undercut that evidence in any radical way: You can still be justified in believing that theory. Or suppose your best evidence leaves you undecided among lots of cosmologies, in some of which there are many freaks. Acknowledging the small chance that you are a freak should only add to your doubt and indecision. It doesn't compel elimination of the possibility that you're a freak.

Compare: Suppose you were to discover some seemingly compelling evidence that the universe is run by a trickster god: You seem to suddenly zoom up into the sky where God in a funny hat says, "Really, I'm just a joker, and I've been fooling you all this time." You couldn't be sure that the seeming evidence wasn't itself some sort of trick or hallucination; but neither would evidence of a trickster god be so self-undermining that you could simply ignore it: It undermines itself, but also everything else. It should reduce your overall certainty. The same goes for cosmological evidence that my cosmological position might be epistemically worse than I think it is.

#

How sure ought I be of the structure of the universe and my place in it? Is it just silly to permit such smidgens of doubt, based on wild, but not entirely groundless, cosmological speculation? Here I am, solid and unmistakable Eric Schwitzgebel! What could be more certain?

– said the fleeting brain, the moment before it dissolved back into chaos.

41. Choosing to Be That Fellow Back Then: Voluntarism about Personal Identity

I have bad news: You're Swampman.[146]

Remember that hike you took last week by the swamp during the electrical storm? Well, one biological organism went in, but a different one came out. The "[your name here]" who went in was struck and killed by lightning. Simultaneously, through minuscule freak quantum chance, a molecule-for-molecule similar being randomly congealed from the swamp. Soon after, the recently congealed being ran to a certain parked car, pulling key-shaped pieces of metal from its pocket that by amazing coincidence fit the car's ignition, and drove away. Later that evening, sounds come out of its mouth that its nearby "friends" interpreted as meaning "Wow, that lightning bolt almost struck me in the swamp. How lucky I was!" Lucky indeed, but a much stranger kind of luck than they supposed.

So you're Swampman. Should you care?

Should you think: I came into existence only a week ago. I never had the childhood I thought I had, never did all those things I thought I did, hardly know any of the people I thought I knew. All that is delusion! How horrible!

Or should you think: Meh, whatever.

Option 1: Yes, you should care. OMG!

Option 2a: No, you shouldn't care, because that was just a fun little body exchange last week. The same *person* went into the swamp as came out, even if the same *body* didn't. Too bad it didn't clear your acne, though.

Option 2b: No, you shouldn't care, because even if in some deep metaphysical sense you aren't the same person as the one who first drove to the swamp, you and that earlier

person share everything that matters. You have the same friends, the same job, the same values, the same (seeming-) memories….

Option 3: Your call. If you choose to regard yourself as one week old, then you are correct to do so. If you choose to regard yourself as decades old, then you are equally correct to do so.

Let's call this third option *voluntarism about personal identity*. Across a certain range of cases, you are who you choose to be.

Social identities are to some extent voluntaristic. You can choose to identify as a political conservative or a political liberal by calling yourself such and successfully resolving to act in certain ways (compare choosing what you love in Chapter 29). You can choose to identify, or not identify, with a piece of your ethnic heritage. You can choose to identify, or not identify, as a philosopher or as a Christian. There are limits: If you have no Pakistani heritage or upbringing, you can't just one day suddenly decide to be Pakistani and thereby make it true that you are. Similarly, if your heritage and upbringing have been entirely Pakistani to this day, you probably can't just instantly shed your Pakistanihood. But in intermediate or vague cases, there's room for choice and making it so.

We might, then, take the same approach to personal identity conceived of in the metaphysical sense. What makes you the same person, or not, in philosophical puzzle cases where intuitions pull both ways, depends partly on how you choose to approach the matter, and different people might legitimately make different choices, thus shaping the metaphysical facts – the actual metaphysical facts about whether it's really "you" – to suit them.

Consider some other stock puzzle cases from the philosophical literature on personal identity:

Teleporter. On Earth there's a device that will destroy your body and beam detailed information about it to Mars. On Mars another device will use that information to create a

duplicate body from local materials – same personality, same attitudes, same (seeming-) memories, everything. Is this harmless teleportation or terrible death-and-duplication? On a voluntaristic view, that would depend partly on how you (or you two) view it. Let's assume, too, that the duplicate body isn't *exactly* identical down to the Planck length. How similar must the body be, on a pro-teleportation view, for a successful teleportation? We can imagine a range of cases, with a substantial gray area. Resolution of these cases too could depend partly on participants' attitudes.[147]

Fission. Your brain will be extracted, cut into two, and implanted into two new bodies. The procedure, though damaging and traumatic, is such that if only one half of your brain were to be extracted and the other half destroyed, everyone would agree that you survived. But in true fission, both halves survive and there will now be two distinct people who both have some claim to be you. Does this procedure count as the loss of your identity as a person, your replacement by two non-identical people? Or does it instead count as some sort of metaphysical-identity-preserving fission, so that you before the fission are now the same person as one or both of the post-fission beings? On a voluntaristic view, it might depend on the attitudes that the pre- and post-fission entities choose to take toward each other.

Amnesia. Longevity treatments are developed so that your body won't die, but in four hundred years the resulting entity will have no memory whatsoever of anything that has happened in your lifetime so far, and if it has similar values and attitudes to your own now, that will only be by chance. Is that future being still "you"? How much amnesia and change can "you" survive without becoming strictly and literally (and not just metaphorically or loosely) a different person? On a voluntaristic view, it might partly depend on the attitudes that such beings choose to take about the importance of memory and attitude constancy in constituting personal identity.

Here are two thoughts in support of voluntarism about personal identity:

(1.) If I try to imagine these cases as actually applying to me, I don't find myself urgently wondering about the resolution of the metaphysical issues at hand, thinking of my death or survival turning on some metaphysical fact out of my control that, for all I currently know, might turn out either way. It's not like being told that if a just-tossed die has landed on 6 then tomorrow I will be shot, which would make me desperately curious to know whether the die had landed on 6. Instead, it seems to me that I can to some extent choose how to conceptualize these cases.

(2.) "Person" is an ordinary, everyday concept that arose in a context distinctly lacking Swampmen, teleporters, human fission, and (that type of) radical amnesia. We might expect that the concept would be somewhat loose-structured or indeterminate in its application to such cases, just like rules of golf might turn out to be loose-structured and indeterminate if we tried to apply them to a context in which golf balls regularly fissioned, merged, and teleported. Part of what makes the concept of "person" important is that it is used, implicitly, to reflect a certain way of thinking about the past or future "me" – for example in feeling regret for what I did in the past and in planning prudently for the future. If so, a looseness or indeterminacy in application might be partly resolved by the person's own value-governed choice: How do I *want* to think about the boundaries of my regrets, my prudential planning, and so forth?

Flipping (2), others can also have an interest in resolving the looseness or indeterminacy in one way rather than other – for example, in punishing you for wrongdoing and in paying your retirement benefits. Society might then also, perhaps more forcefully, resolve prior metaphysical indeterminacies by choosing to recognize the boundaries of people in one way rather than another. This too would be a kind of voluntarism about personal identity, but with a different chooser.

There must be limits, though. Voluntarism works, if it works at all, only for gray cases. I can't decide to be identical with a future coffee mug – perhaps by instructing someone to put it atop my grave with the sign "Hi, this is me, the new shape of Eric Schwitzgebel!" – and thereby make it so.

What if the current Dalai Lama and some future child (together, but at a temporal distance) decide that they are metaphysically the same person? Can they make it so, if their society agrees and enough other things fall into place?[148]

42. How Everything You Do Might Have Huge Cosmic Significance

Infinitude is a strange and wonderful thing. It transforms the ridiculously improbable into the inevitable. Hang on to your hat and glasses. Today's weird reasoning is going to make mere Boltzmann brains (Chapter 40) and Swampmen (Chapter 41) and seem comparatively probable.

First, let's suppose that the universe is infinite. Cosmologists tend to view this as plausible.[149]

Second, let's suppose that the "Copernican Principle" holds. We're not in any special position in the universe, just kind of a mid-rent location. This principle is also widely accepted.[150]

Third, let's assume cosmic diversity. We aren't stuck in an infinitely looping variant of a small subset of the possibilities. Across infinite spacetime, there's enough variety to run through all or virtually all of the finitely specifiable physical possibilities infinitely often. Everything that isn't contrary to the laws of nature will occur over and over again.

Those three assumptions are somewhat orthodox. To get the argument into more seriously weird territory, we need a few more assumptions that are less orthodox, but I hope not wildly implausible.

Fourth, let's assume that complexity scales up infinitely. In other words, as you zoom out on the infinite cosmos, you don't find that things eventually look simpler as the scale of measurement gets bigger.

Fifth, let's assume that local actions on Earth have chaotic effects of arbitrarily large magnitude. You might know the "butterfly effect" from chaos theory – the idea that a small perturbation in a complex, chaotic system can eventually make a large-scale difference in the

behavior of the system.[151] A butterfly flapping its wings in Brazil could cause the weather in the U.S. weeks later to be different than it would have been if the butterfly hadn't flapped its wings. Small perturbations amplify.

Sixth, given the right kind of complexity, evolutionary processes will transpire that favor intelligence. We wouldn't expect such evolutionary processes at most spatiotemporal scales. However, given that complexity scales up infinitely (our fourth assumption), plus the Copernican Principle, we should expect that at some finite proportion of spatiotemporal scales there are complex systems structured in such a way as to enable the evolution of intelligence.

From all this, it seems to follow that what happens here on Earth – including the specific choices you make, chaotically amplified as you flap your wings – can have effects on a cosmic scale that influence the cognition of very large minds. You had a pumpernickel bagel for breakfast instead of Corn Chex. As a result, eventually, some giant cosmic mind turned left rather than right on its way home from work.

Let me be clear that I mean *very* large minds. I don't mean galaxy-sized minds or visible-universe-sized minds. Galaxy-sized and visible-universe-sized structures in our region don't seem to be of the right sort to support the evolution of intelligence at those scales. I mean way, *way* up, vastly huger than this tiny droplet of a thing we call the visible universe, with its itsy-bitsy galaxies. We have infinitude to play with, after all. And presumably way, way *slow* if the speed of light is a constraint. (I'm assuming that time and causation make sense at arbitrarily large scales, but if necessary, time and causation might be replaceable by something weaker like contingency.)

Far-fetched? Cool, perhaps, depending on your taste in cool. Maybe not quite cosmic *significance* though, if your decisions only feed a pseudo-random mega-process whose outcome has no meaningful relationship to the content of your decisions.

But since we have infinitude at hand, we can add another twist. If the odds of influencing the behavior of a huge mind are finite, and if we're permitted to scale up infinitely, your decisions will affect not just one but an infinite number of huge minds. Among these minds there will be some – a tiny but finite proportion – of whom the following conditional statement is true: If you hadn't just made that upbeat, life-affirming choice you in fact just made, that huge entity would have decided its life wasn't worth living. However, instead, partly because of that thing you did, the giant entity – let's call it Emily – will discover happiness and learn to celebrate its existence.

We can even find cases in which these kinds of conditional statements are true across an arbitrarily large range of variations, if we're willing to look for the right Emily up there: *Emily Vast* is the being such that if you had made *any* life-affirming choice ten minutes ago, across a wide range of possible choices you might have made, she would have discovered happiness, and if you had not, she wouldn't have. Given the chaotic connections, a miniscule but finite proportion of Emilies will be like this. Each relevant finitely-specifiable possible life-affirming action must, by chance, be such that had you chosen it, it would have caused a life-affirming outcome for Emily Vast.

If we're willing to gaze still farther up in search of the right Emily Vast, rising through the spatiotemporal scales, multiplying one minuscule-but-finite chance upon another, we will eventually discover an *Emily Megavast* who sees and knows about you in the following sense: Somewhere in her environment is an approximately Emily Megavast-sized being who acts and looks (or "acts" and "looks") much like you, and does so as a result of chaotic chance contingencies linking back to your visible body and your behavioral choices. You do your lovely life affirming thing, and much later, as a result, the megavast analog of you does its analogous lovely life-affirming thing. Emily Megavast is inspired. Her life is changed forever!

In an infinite cosmos, given background assumptions that are not wholly implausible, this is virtually certain to be so, within as precise an error tolerance as you care to specify.

I hope it won't seem presumptuous of me now to thank you already, on Emily's behalf.[152]

43. Penelope's Guide to Defeating Time, Space, and Causation

Homer did not sing of Penelope's equations – the equations, in neat chalk lettering, covering wall and ceiling of her upstairs chambers, the equations built and corrected over years, proving that Odysseus will sometime return.

After long study, Eurykleia agreed with Penelope. The universe is provably diverse and provably infinite. So an Odysseus must return from the sea. Indeed, he must do so infinitely often. There are only finitely many ways that atoms can arrange themselves, within the error tolerances of human concern.

Eurykleia did not think it followed, however, that one should stand upon the palace roof and wait in the rain, gazing across the wine-dark sea. This is where Penelope was when lightning struck and killed her.

Ithaka mourned. Greece fell. Humanity fell. Sun swallowed Earth, the stars burned out, the universe became cool and quiet. But infinite time is a powerful thing.

#

Lightning struck the iron railing to Penelope's right. Eurykleia urged Penelope back inside.

Penelope sat in her chambers by a blazing hearth, while Eurykleia wrapped her in towels and slowly brushed her hair. "Other Penelopes will continue me," Penelope said. "Infinitely many, who have exactly my ideas, exactly my plans and longings, exactly this mole upon my cheek. An infinite subset of them will greet a returning Odysseus before nightfall. An infinite subset will leave Ithaka, riding a thin ship on the wild sea. They will

find old friends, drown in storms, discover giants. Lightning cannot end me, only redistribute me."

Penelope and Eurykleia agreed: A new cosmos will eventually burst forth from disorder with duplicates of them, duplicates of all Greece. Although at any one time the chance is unfathomably minuscule that so many atoms will happen randomly to arrange themselves just right, that chance is finite and it sums infinitely across the whole. But accepting this, they disagreed about the implications.

"Though perfect duplicates will eventually exist," said Eurykleia, "they are beyond our concern."

"It will be just as if I smoothly continued," said Penelope, "as if the intervening aeons had passed in a blink."

Eurykleia wrapped a finishing band around the braided tips of Penelope's long hair. "They are many, but you are one, my love. They can't all be you."

"There is an infinitude of me-enough," said Penelope, standing, now dry, ready to select her evening clothes. "I will leave here and do all things."

#

Eurykleia helped Penelope, still wet, into a dress and scarf. Together, they descended to the main hall where the suitors reveled. With mock formality, Ktesippos asked Penelope to dance.

"Somewhere I marry Ktesippos," Penelope said. "Somewhere else I feast upon his eyes while he gladly sings a poem. The suitors are atoms. By chance or repulsion they might disperse widely, or congregate all on one side of the room, or step in some vastly improbable

rhythm that resonates through the floor, through the stone and dirt, and vibrates the air above the sea into a siren's call."

Eurykleia laid toasted wheat and cheese on a festive dish, handing it to Penelope. "Eat. You are too lean. Your mind is over-hot."

A rag-clad man appeared in the doorway: Odysseus returned. Eurymakhos entered silently behind, killed Odysseus with a spear-thrust through the back, then stepped forward to ask Penelope's hand in marriage, the other suitors applauding.

Penelope chose instead to ride a thin ship on the wild sea.

#

Odysseus slew the suitors.

In their marital chamber, beneath a ceiling of equations, Penelope said to Odysseus, "Your recent travels are nothing. We will journey together far wider. The Aegean is a droplet."

Odysseus said, "Mathematics is only dance steps that govern chalk. No universe exists beyond these chamber walls."

Odysseus smelled of someone's recent past: of herbs and swine and salt. Penelope and Odysseus made love, and Penelope contemplated the assumptions implicit in her axioms.

#

By minuscule chance, a black hole emitted a pair of complex systems who briefly thought – before tidal forces and the vacuum of space consumed them – that they were

Penelope and Eurykleia dancing together in slow silence, to an imagined song, late at night in the grand hall.

#

It was evening and Penelope was unweaving Laertes's death shroud. With a silver pick, Penelope gently pulled up the fine thread, while Eurykleia coiled it back upon the spool.

"An unobservable difference is no difference at all," Penelope said.

"I love the particular you here now," said Eurykleia, "not the infinitude of diverse Penelopes elsewhere."

They heard a thump from downstairs, then a chorus of festive shouts. A suitor hollered for more pink wine.

"If the Earth spins," said Penelope, "then all of Greece moves. Do you love me less because we are now in a different place?"

Penelope ceased her unweaving. Moonlight across Laertes's half-made shroud cast a shadow over the equations on one wall. Penelope leaned forward, touching her finger to a brown dot on Eurykleia's left shoulder. In the years of waiting, Eurykleia's shoulder had become an old woman's shoulder. Penelope circled the dot with her fingertip, then stood.

Penelope climbed to the roof and stood in the lightning storm, gazing out, Eurykleia standing reluctantly in the doorway behind. Penelope stretched her right hand back, inviting.

Eurykleia took a half-step into the rain, then stopped. "Maybe," she said, "it's the invisible threads of causation th-

#

A galaxy suddenly congealed from chaos.

Something indistinguishable from Penelope and something indistinguishable from Eurykleia stood upon the roof. They seemed to remember having left, down below, Laertes's half-undone shroud. They seemed to remember Odysseus, Telemakhos, suitors, Greece, the constellations. And indeed this galaxy did contain such things, as, given infinite time, some such sudden galaxy eventually must.

-at I value?" continued Eurykleia.

Penelope took Eurykleia's hand and led her to the edge of the roof. Together they gazed past the low railing, down the edge of the cliff upon which the palace stood, to the slate shore and rough waves.

"If we leapt over the edge," Penelope said, "there is a small but finite chance that the winds would bear us up – a small but finite chance that the atoms of air would strike our bodies exactly so, preventing our fall and sending us soaring instead like birds among the clouds."

"I can't do this," said Eurykleia.

"If we closed our eyes," said Penelope, "the death of these bodies would come with no warning. We would never know it. And somewhere a Penelope and Eurykleia will burst forth from chaos, flying, knowing they are us."

"But there will be many more Penelopes and Eurykleias dead upon the rocks, and many more half-congealed Penelopes and Eurykleias who fall back into death or who survive but misremember."

"Many more?" said Penelope. "There are infinitely many of all types. Just as the infinitude of points in a small line segment is identical to the infinitude of points in the whole of space, the infinitude of flying Penelope-Eurykleia pairs is identical to the infinitude of the pairs of corpses upon the shore is identical to the infinitude who remain upon the roof."

“Infinitude is a strange and wonderful thing, Penelope. But I cannot leap.”

Elsewhere, though, somewhere, Penelope and Eurykleia leapt and flew, and leapt and flew, and leapt and flew.

44. Goldfish-Pool Immortality

Must an infinitely continued life inevitably become boring? The famous twentieth-century ethicist Bernard Williams notoriously thought so.[153]

But consider Neil Gaiman's story "The Goldfish Pool and Other Stories" (yes that's the name of one story):

> He nodded and grinned. "Ornamental carp. Brought here all the way from China.
>
> We watched them swim around the little pool. "I wonder if they get bored."
>
> He shook his head. "My grandson, he's an ichthyologist, you know what that is?"
>
> "Studies fishes."
>
> "Uh-huh. He says they only got a memory that's like thirty seconds long. So they swim around the pool, it's always a surprise to them, going "I've never been here before.' They meet another fish they known for a hundred years, they say, 'Who are you, stranger?'"[154]

The problem of immortal boredom solved: Just have a bad memory! The even seemingly unrepeatable pleasures, like meeting someone for the first time, become repeatable.

Now you might say, wait, when I was thinking about immortality I wasn't thinking about forgetting everything and doing it again like a stupid goldfish.

To this I answer: Weren't you?

If you imagine that you were continuing life as a human, you were imagining, presumably, that you had finite brain capacity. There's only so much memory you can fit into eighty billion neurons. So of course you're going to forget things, at some point almost

everything, and things sufficiently well forgotten could presumably be experienced as fresh again. This is always what is going on with us anyway to some extent.

Immortality as an angel or a transhuman super-intellect only postpones the issue, as long as one's memory is finite.

The question of the value of repetition is thus forced on us: Is repeating and forgetting the same types of experiences over and over again, infinitely, preferable to doing them only once, or only twenty times, or "only" a googolplex times? The answer to that question isn't entirely clear. One possibly relevant consideration, however, is this: If you stopped one of the goldfish and asked, "Do you want to keep going along?" the happy little fish would say, "Yes, this is totally cool! I wonder what's around the corner. Oh, hi, glad to meet you!" It seems a shame to cut the fish off when it's having so much fun – to say, "nope, no value added, you've already done this a million times, sorry!"

Alternatively, consider an infinitely continuing life with infinite memory. How would that work? What would it be like? Would you be overwhelmed and almost paralyzed like the titular character in Jorge Luis Borges's story "Funes the Memorious"?[155] Would there be a workable search algorithm for retrieving those memories? Would there be some tagging system to distinguish each memory from infinitely many qualitatively identical other memories? Also, infinitude is pretty big and weird, as we've been discussing – so at some point you'll need to become pretty big and weird too.

True temporal infinitude forces a dilemma between two options:

(1.) infinite repetition of the same things, without memory, or

(2.) an ever-expanding range of experiences that eventually diverges so far from your present range of experience that it becomes questionable whether it's right to regard that future being as "you" in any meaningful sense.

Call the choice *The Immortal's Dilemma*.

Given infinite time, a closed system will eventually start repeating its states, within any finite error tolerance.[156] There are only so many relevantly distinguishable states a closed system can occupy. Once it has occupied them, it has to start repeating at least some of them. Assuming that memory belongs to the system's states, then memory too is among those things that must start afresh and repeat. But it seems reasonable to wonder whether the forgetful repetition of the same experiences, infinitely again and again, is worth hoping for – whether it's what we can or should want in immortality.

It might seem better, then, or more interesting or more worthwhile, to have an open system – one that is always expanding into new possibilities. But this brings its own problems.

Suppose that conscious experience is what matters. First, you might cycle through every possible human experience. Maybe human experience depends on a brain of no more than a hundred trillion neurons (currently we have about eighty billion, but that might change), and that each neuron can be in one of a hundred trillion relevantly distinguishable states. Such numbers, though large, are finite. So once you're done living through all the experiences of seeming-Aristotle, seeming-Gandhi, seeming-Hitler, seeming-Hitler-seeming-to-remember-having-earlier-been-Gandhi, seeming-future-super-genius, and seeming-every-possible-other-person and many, many more experiences that probably wouldn't coherently belong to anyone's life, well, you've either got to settle in for some repetition or find some new experiences that are no longer human. Go through the mammals, then. Go through variety of increasingly remote hypothetical creatures. Expand, expand – eventually you'll have run through all possible smallish creatures with a neural or similar basis. You'll need to move to experiences that are either radically alien or vastly superhuman or both.

At some point – maybe not very far along in this process – it's reasonable to wonder whether the entity who is having all of these experiences is really "you". Even if there is some continuous causal thread reaching back to you as you are now, should you care about this remotely distant entity in any more personal way than you care about the future of some entity unrelated to you?

Either amnesic infinite repetition or an infinitely-expanding range of unfathomable alien weirdness. That's the dilemma.

Or maybe you were imagining retaining your humanity but somehow existing non-temporally? I find that even harder to conceive. What would *that* be like?

#

Let's return to assuming we'll be goldfish. That's the version of infinitude I think I understand best, and probably the one I'd choose. Suppose it's Heaven, or at least a deserved reward. My creator stops me mid-swim. She asks not only whether I want to keep going (I do), but also what size pool I want. Do I want a large pool, that is, a relatively wide range of experiences to loop through? Or would I rather have a smaller pool, a shorter loop of more carefully selected, more individually awesome experiences, but less variety? Either way, I won't know the difference. I won't get bored. Each loop-through will be experienced as fresh and surprising.

I think I'd choose a moderately large loop. When I imagine a tiny little loop – for example, just hanging out in the clouds, blissfully playing the harp, joyfully contemplating the divine, over and over in ecstatic amnesia – I think maybe joy is overrated. Instead, give me diverse adventures and a long memory.[157]

45. Are Garden Snails Conscious? Yes, No, or *Gong*

If you grew up in a temperate climate, you probably spent some time bothering brown garden snails (*Cornu aspersum,* formerly known as *Helix aspersa*). I certainly did. Now, as a grown-up (supposedly) expert (supposedly) on the science and philosophy of consciousness, I'm fretting over a question that didn't trouble me very much when I was seven: Are garden snails conscious?

Naturally, I started with a Facebook poll of my friends, who obligingly fulfilled my expectations by answering, variously, "yes" (here's why), "no" (here's why not), and "OMG that is the stupidest question". I'll call this last response "*gong*" after *The Gong Show,* an amateur talent contest in which performers whose acts were sufficiently horrid are interrupted by a gong and ushered off the stage.

It turns out that garden snails are even cooler than I thought, now that I'm studying them more closely. Let me fill you in.

#

Garden Snail Cognition and Behavior[158]

The central nervous system of the brown garden snail contains about 60,000 neurons. That's quite a few more neurons than the famously mapped 302 neurons of the *Caenorhabditis elegans* roundworm, a fraction of the number of neurons in the brain of an ant or fruitfly. The snail's brain is organized into several clumps of ganglia, mostly in a ring around its esophagus. Gastropod (i.e., snail and slug) neurons generally resemble vertebrate neurons, with a few notable exceptions. One exception is that gastropod neurons usually don't have a bipolar structure with output axons on one side of the cell body and input dendrites on the other side. Instead, input and output typically occurs on both sides without a

clear differentiation between axon and dendrite. Another difference is that although gastropods' small-molecule neural transmitters are the same as in vertebrates (e.g., acetylcholine, serotonin), their larger-molecule neuropeptides are mostly different.

Snails navigate primarily by chemoreception, or the sense of smell, and mechanoreception, or the sense of touch. They will move toward attractive odors, such as food or mates, and they will withdraw from noxious odors and tactile disturbance. Although garden snails have eyes on the tips of their posterior tentacles, their eyes seem to be sensitive only to light versus dark and the direction of light sources, rather than to the shapes of objects. Snail tentacles are instead highly specialized for chemoreception, with the higher-up posterior tentacles better for catching odors on the wind and the lower anterior tentacles better for taste and odors close to the ground. Garden snails can also sense the direction of gravity, righting themselves and moving toward higher ground to avoid puddles.

Snails can learn. Gastropods fed on a single type of plant will prefer to move toward that same plant type when offered the choice in a Y-shaped maze. They can also learn to avoid foods associated with noxious stimuli, in some cases even after only a single trial. Some species of gastropod will modify their degree of attraction to sunlight if sunlight is associated with tumbling inversion. In warm ocean *Aplysia californica* gastropods, the complex role of the central nervous system in governing reflex withdrawals has been extensively studied. *Aplysia californica* reflex withdrawals can be inhibited, amplified, and coordinated, maintaining a singleness of action across the body and regulating withdrawal according to circumstances. Withdrawals can be habituated (reduced) after repeated exposure to harmless stimuli and sensitized (increased) when the triggering stimulus is regularly followed by something even more aversive. Garden snail nervous systems appear to be similarly complex to those of *Aplysia californica*, generating unified action that varies with circumstance.

Garden snails can coordinate their behavior in response to information from more than one modality at once. As I've already mentioned, when they detect that they are surrounded by water, they can seek higher ground. They will cease eating when satiated, withhold from mating while eating despite sexual arousal, and exhibit less withdrawal reflex while mating. Before egg laying, garden snails use their feet to excavate a shallow cavity in soft soil, then insert their head into the cavity for several hours while they ovulate. Land snails will also maintain a home range to which they will return for resting periods or hibernation, rather than simply moving in an unstructured way toward attractive sites or odors.

Garden snail mating is famously complex. *Cornu aspersum* is a simultaneous hermaphrodite, playing both the male and female role simultaneously. Courtship and copulation requires several hours. Courtship begins with the snails touching heads and posterior antennae for tens of seconds, then withdrawing and circling to find each other again, often consuming each other's slime trails, or alternatively breaking courtship. They repeat this process several times. During mating, snails will sometimes bite each other, then withdraw and reconnect. Later in courtship, one snail will shoot a "love dart" consisting of calcium and mucus at the other, succeeding in penetrating the skin about one third of the time; tens of minutes later, the other snail will reciprocate. Sex culminates when the partners manage to simultaneously insert their penises into each other, which may require dozens of attempts. Garden snails will normally mate several times before laying eggs, and the sperm of mates whose darts successfully landed will fertilize more of their eggs than the sperm of mates with worse aim.

Impressive accomplishments for creatures with brains of only 60,000 neurons! Of course, snail behavior is limited compared to the larger and more flexible behavioral repertoire of mammals and birds.

#

Garden Snail Consciousness: Three Possibilities

So, knowing all this... are garden snails conscious? Is there something it's like to be a garden snail? Do snails have, for example, sensory experiences?

Suppose you touch the tip of your finger to the tip of a snail's posterior tentacle, and the tentacle retracts. Does the snail have tactile experience of something touching its tentacle, a visual experience of a darkening as your finger approaches and occludes the eye, an olfactory or chematosensory experience of the smell or taste or chemical properties of your finger, a proprioceptive experience of the position of its now-withdrawn tentacle?

(1.) *Yes.* It seems like we can imagine that the answer is yes, the snail does have sensory experiences. Any specific experience we try to imagine from the snail's point of view, we will probably imagine too humanocentrically. Withdrawing a tentacle might not feel much like withdrawing an arm. Optical experience in particular might be so informationally poor that calling it "visual" is already misleading, inviting too much analogy with human vision. Nonetheless, I think we can conceive in a general way how it might be the case that garden snails have sensory experiences of some sort or other.

(2.) *No.* We can also imagine, I think, that the answer is no, snails entirely lack sensory experiences of any sort – and thus, presumably, any consciousness at all, on the assumption that if snails are conscious they have at least sensory consciousness. If you have trouble conceiving of this possibility, consider dreamless sleep, toy robots, and the enteric nervous system. (The enteric nervous system is a collection of about half a billion neurons lining your gut, governing motor function and enzyme secretion.) In all three of these cases, most people think, there is no genuine stream of conscious experience, despite some

organized behavior and environmental reactivity. It seems that we can coherently imagine snail behavior to be like that: no more conscious than turning over unconsciously in sleep, or than a toy robot, or than the neurons lining your intestines.

We can make sense of both of these possibilities, I think. Neither seems obviously false or obviously refuted by the empirical evidence. One possibility might strike you as intuitively much more likely than the other, but as I've learned from chatting with friends and acquaintances (and from my Facebook poll), people's intuitions vary – and it's not clear, anyway, how much we ought to trust our intuitions in such matters. You might have a favorite scientific or philosophical theory from which it follows that garden snails are or are not conscious; but there is little consensus on general theories of consciousness and leading candidate theories yield divergent answers.

(3.) **Gong**. To these two possibilities, we can add a third, the one I am calling *gong*. Not all questions deserve a yes or a no. There might be a false presupposition in the question (maybe "consciousness" is an incoherent concept?), or the case might be vague or indeterminate such that neither "yes" nor "no" quite serves as an adequate answer. (Compare vague or indeterminate cases between "green" and "not green" or between "extraverted" and "not extraverted".)

Indeterminacy is perhaps especially tempting. Not everything in the world fits neatly into determinate, dichotomous yes-or-no categories. Consciousness might be one of those things that doesn't dichotomize well. And snails might be right there at the fuzzy border.

Although an indeterminate view has some merits, it is more difficult to sustain than you might think at first pass. To see why, it helps to clearly distinguish between being *a little* conscious and being *in an indeterminate state between* conscious and not-conscious. If one is a little conscious, one is conscious. Maybe snails just have the tiniest smear of consciousness – that would still be consciousness! You might have only a little money. Your entire net

worth is a nickel. Still, it is discretely and determinately the case that if you have a nickel, you have some money. If snail consciousness is a nickel to human millionaire consciousness, then snails are conscious.

To say that the dichotomous yes-or-no does not apply to snail consciousness is to say something very different than that snails have just a little smidgen of consciousness. It's to say... well, what exactly? As far as I'm aware, there's no well-developed theory of kind-of-yes-kind-of-no consciousness. We can make sense of a vague kind-of-yes-kind-of-no for "green" and "extravert"; we know more or less what's involved in being a gray-area case of a color or personality trait. We can imagine gray-area cases with money too: Your last nickel is on the table over there, and here comes the creditor to collect it. Maybe that's a gray-area case of having money. But it's much more difficult to know how to think about gray-area cases of being somewhere between a little bit conscious and not at all conscious. So while in the abstract I feel the attraction of the idea that consciousness is not a dichotomous property and garden snails might occupy the blurry in-between region, the view requires entering a theoretical space that has not yet been well explored.

There is, I think, some antecedent plausibility to all three possibilities, *yes, no,* and **gong**. To really decide among them, to really figure out the answer to our question about snail consciousness, we need an empirically well-grounded general theory of consciousness, which we can apply to the case.

Unfortunately, we have no such theory. The live possibilities appear to cover the entire spectrum from the panpsychism or near-panpsychism of Galen Strawson and of Integrated Information Theory, on which consciousness is ubiquitous in the universe, even in extremely simple systems, to very restrictive views, like those of Daniel Dennett and Peter Carruthers, on which consciousness requires sophisticated self-representational capacities of the sort that well beyond the capacity of snails.[159]

Actually, I find something wonderful about not knowing. There's something marvelous about the fact that I can go into my backyard, lift a snail, and gaze at it, unsure. Snail, you are a puzzle of the universe, right here in my garden, eating the daisies!

Part Five: Kant vs. the Philosopher of Hair

46. Truth, Dare, and Wonder

According to Nomy Arpaly and Zach Barnett, some philosophers prefer Truth and others prefer Dare.[160] Yes! But there are also Wonder philosophers.

Truth philosophers aim to present the philosophical truth as they see it. Usually, they prefer modest, moderate, commonsense views. They also tend to recognize the complementary merits of competing views – at least after they have been knocked loose from their youthful enthusiasms – and thus Truth philosophers tend to prefer multidimensionality and nuance.[161] Truth philosophers would rather be boring and right than interesting and wrong.

Dare philosophers reach instead for the bold and unusual. They like to explore the boundaries of what can be defended. They're happy for the sake of argument to champion strange positions they don't really believe, if those positions are elegant, novel, fun, contrarian, or have some unappreciated merit. Dare philosophers tend to treat philosophy as a game in which the ideal achievement is the breathtakingly clever defense of a view that others would have thought to be absurd.

The interaction of Truth and Dare creates a familiar dynamic. The Dare philosopher ventures a bold thesis, cleverly defended. ("Possible worlds really exist!", "All matter is conscious!" "We're morally obliged to let humanity go extinct!" [162]) If the defense is sufficiently clever, readers are tempted to think "Wait, could that really be true? What exactly is wrong with the argument?" Then the Truth philosopher steps in, finding the holes and illicit presuppositions in the argument, or at least trying to, defending a more sensible view.

This Dare-and-Truth dynamic is central to academic philosophy and good for its development. Sometimes Daring views have merit we wouldn't notice without Dare philosophers out there pushing the limits. Moreover, there's something intrinsically

worthwhile or beautiful in exploring the boundaries of philosophical defensibility, even for positions that prove to be flatly false. It's part of the glory of life on Earth that we have fiendishly clever modal realists and panpsychists in our midst.

Now consider Wonder.

Why study philosophy? I mean personally. What do you find interesting or rewarding about it? One answer is Truth: Through philosophy you (hopefully) learn the truth about profound and difficult issues. Another answer is Dare: It's fun to match wits, develop lines of reasoning, defend surprising theses, and win the argumentative game despite starting from a seemingly indefensible position. Both of these motivations speak to me. But I think what really delights me more than anything else in philosophy is its capacity to upend what I think I know, its capacity to call into question what I previously took for granted, its capacity to throw me into doubt, confusion, and wonder.

In conversation, Nomy said that she had been thinking of me as a typical Dare philosopher. I can see the basis of her impression. After all, I've argued that the United States might literally be phenomenally conscious, that we're such bad introspectors that we might be radically mistaken about even the seemingly most obvious aspects of our own currently ongoing stream of experience, that we might be short-lived artificial intelligences in a computer-generated world run by a sadistic teenager, and that someone might continue to exist after having been killed by lightning if a freak molecule-for-molecule duplicate of them suddenly coagulates in the distant future.[163] Hard to get more Dare than that!

Well, except for the *might* part. In my mind, the "might" is a crucial qualification. It makes the claims, despite their strangeness, considerably less Daring. "Might" can be a pretty low bar.

Unlike the Dare philosopher, the Wonder philosopher is guided by a norm of sincerity and truth. The practice of philosophy is not, for the Wonder philosopher, primarily about

matching wits and finding clever arguments that you yourself needn't believe. However, like the Dare philosopher and unlike the Truth philosopher, the Wonder philosopher loves the strange and seemingly wrong – and is willing to push wild theses to the extent they suspect that those theses, wonderfully, surprisingly, *might* be true.

Probably no reader of philosophy is pure Truth, pure Dare, or pure Wonder. Nor, I'm sure, is this distinction exhaustive.[164] Furthermore, our motives might blur together (Chapter 11). There might be ways of sincerely pursuing Truth by adopting a Dare attitude, etc. Insert further nuance, qualification, and multifacetedness as required for Truth.

#

In the Dare-and-Truth dynamic of the field, the Wonder philosopher can have trouble finding a place. Bold Dare articles are exciting. Wow, what cleverness! Sensible Truth articles also find a home in the journals, especially when they're responding to some prominent Dare position. The Wonder philosopher's "Whoa, I wonder if this weird thing might be true?" is a little harder to publish.

Since probably none of us is pure Truth, Dare, or Wonder, one approach is to leave Wonder out of your research profile: Find the Truth, where you can, publish that, maybe try a little Dare if you dare, and leave Wonder for your classroom teaching and private reading. Defend the existence of moderate naturalistically-grounded moral truths in your published papers; relish the weirdness of Zhuangzi on the side.

Still, that seems a little sad, don't you think?

So, for aspiring Wonder philosophers out there, I recommend four publishing strategies:

(1.) Find a seemingly Daring position that you really do sincerely endorse on reflection, and defend that. Sometimes the Truth is strange enough to seem Daring or Wonderful.

(2.) Explicitly argue that we shouldn't wholly reject some Daring position – for example, because the Truth-like arguments aren't on inspection fully compelling.

(3.) Find a Truth-defensible view that generates Wonder if it's true, even if no specific Daring position follows. For example, defend some form of reasonable doubt about philosophical method or self-knowledge. Then argue that a plausible consequence is that we don't know some of the things that we normally take for granted.

(4.) Write about historical philosophers with weird and Wonderful views. This gives you a chance to explore the Wonderful without committing to it.

In retrospect, I've discovered that more than half of my work falls under one of these four heads.

47. Trusting Your Sense of Fun

If you've read this far into this dorky book,[165] I have some bad news. You're a *philosophy dork*.

Some of us philosophy dorks have been philosophy dorks for a long time, even before we knew what academic philosophy was. Maybe, like me, when you were twelve, you said to your friends, "Is there really a world behind that closed door? Or does the outside world only pop into existence when I open the door?" and they said, "Dude, you're weird! Let's go play basketball." Maybe, like me, when you were in high school you read science fiction and wondered whether an entirely alien moral code might be as legitimate as our own, and this prevented you from taking your World History teacher entirely seriously.

If you're a deep-down philosophy dork, then you have a certain underappreciated asset: a philosophically-tuned sense of fun. You should trust that sense of fun.

It's fun – at least *I* find it fun – to think about whether there is some way to prove that the external world exists. It's fun to see whether ethics books are any less likely to be stolen than other philosophy books. It's fun to think about why people used to say they dreamed in black and white, to think about the essence of jerkitude, to think about how weirdly self-ignorant people are, to think about what sorts of bizarre aliens might be conscious, to think about whether babies know that things continue to exist outside of their perceptual fields.[166] At every turn in my career, I have faced choices about whether to pursue what seemed to me to be tiresome, respectable, philosophically mainstream, and at first glance the better career choice, or whether instead to follow my sense of fun. Rarely have I regretted it when I have chosen fun.

I see three main reasons a philosophy dork should trust their sense of fun. (Hey, numbered lists of reasons to hold weird views, that's kind of fun!)

(1.) Fun and boredom are emotional indicators of epistemic value. If you truly are a philosophy dork in the sense I intend the phrase (and consider, *this* is how you are spending your free time!), then your sense of what's fun will tend to reflect what really is, for you, philosophically worth pursuing. You might not be able to quite put your finger on why it's worth pursuing, at first. It might even just seem a pointless intellectual lark. But my experience is that the deeper significance will eventually reveal itself. Maybe everything can be explored philosophically and brought back around to main themes, if one plunges deep enough. But I'm inclined to think it's not *just* that. The true dork's mind has a sense of where it needs to go next – an intuition about what you'll benefit from thinking about, versus what will drift past your brain in a sleepy haze (despite rightly being immensely interesting to some others). Emotional indicators of epistemic value can of course be misleading – video games are designed to exploit them, for example – but for dorks who have wallowed in philosophy long enough, those indicators are likely to have more than just superficial merit.

(2.) Chasing fun ideas energizes you. Few things are more dispiriting than doing something tedious because "it's good for your career". You'll find yourself wondering whether this career is really for you, whether you're really cut out for philosophy. You'll find yourself procrastinating, checking social media,[167] spacing out while reading, prioritizing other duties. If instead you chase the fun first, you'll find philosophical exploration viscerally attractive. You'll be eager to do it. Later, this eagerness can be harnessed back onto a sense of responsibility. Finding your weird passion first, and figuring out what you want to say about it, can motivate you to go back later and more thoroughly read the (sometimes to-you unexciting) stuff that others have written about the topic, so that you can fill in the references, connect with previous related research, and refine your view in light of others' work. Reading the philosophical literature is much more rewarding when you have an exciting lens to read it through, an agenda in mind that you care about. Slogging through

it from some vague sense of unpleasant duty is a good way to burn yourself out and waste your summer.

(3.) Fun is contagious. So is boredom. Readers are unlikely to enjoy your work and be enthusiastic about your ideas if even *you* don't have that joy and enthusiasm.

These remarks probably generalize across disciplines. I think of (the famous physicist) Richard Feynman's description of how he recovered from his early-career doldrums by doing a fun but seemingly frivolous project on the mathematics of spinning cafeteria plates.[168] I think of (*Buffy the Vampire Slayer* creator) Joss Whedon's advice for writers to "Absolutely eat dessert first. The thing you want to do the most, do that."[169] Dessert won't spoil your supper. On the contrary, dessert is the important thing, the thing that really calls to be done, and doing it will give you the energy you need to eat your vegetables later.

48. What's in People's Stream of Experience During Philosophy Talks?

I've worked a bit with psychologist Russ Hurlburt, and also a bit on my own, using beepers to sample people's stream of experience during their ordinary activities. The basic idea is this: You wear a beeper for a while, just going about your normal business. The beeper is set to go off after some random interval (typically 1-60 minutes) during which you usually kind of forget that you're wearing it. Each time the beep sounds, you are to immediately think back on what was in your last undisturbed moment of inner experience just before the beep. Later, usually within 24 hours, you are carefully interviewed about that randomly sampled slice of experience.[170] Despite qualms (expressed in my 2007 book with Hurlburt), I find this an intriguing method.

When I give a talk about experience sampling, which I've done sometimes with Russ and sometimes solo, I typically beep the audience during the talk itself to help give them a sense of the procedure. I give each member of the audience a slip of paper with a random number. I set a timer for something like 4-10 minutes, then I launch into my lecture. The audience knows that when the beep sounds, I'll stop talking and they'll have a minute or two to reflect on their last moment of experience right before the beep. Then I'll randomly select a number, and (if they consent) I'll interview the audience member with that number right there on the spot, with others free to jump in with their own questions. It's fun, and the audience is usually very engaged.

Typically, in a two- or three-hour session, there will be time to interview people about 2-3 samples. There's got to be some lecturing too, of course, and careful interviews take a while, and all the side discussions and audience questions slow things down further. But now I've done it maybe eight or ten times and I have a couple of dozen samples – enough

to start making generalizations about the inner experiences people report having while listening to my lectures.

The most striking thing to me is this: Only a minority – about 15-25% – report having been attending to the content of the talk.

To give a flavor of what people do report, here are summary reports of six samples from 2010. These were the most recent samples at the time I first drafted the blogpost on which this chapter is based. Half are from a presentation I gave to an advanced undergraduate class in Claremont, and half are from a joint presentation with Russ at a meeting of the Pacific Division of the American Philosophical Association.[171]

(1.) Thinking that he should put his cell phone away (probably not formulated either in words or imagery); sensory visual experience of the cell phone and whiteboard.

(2.) Scratching an itch, noticing how it feels; having a visual experience of a book.

(3.) Feeling like he's about to fade into a sweet daydream but no sense of its content yet; "fading" visual experience of the speaker.

(4.) Feeling confused; listening to the speaker and reading along on the handout, taking in the meaning.

(5.) Visual imagery of the "macaroni orange" of a recently seen flyer; skanky taste of coffee; fantasizing about biting an apple instead of tasting coffee; feeling a need to go to the bathroom; hearing the speaker's sentence. The macaroni orange was the most prominent part of her experience.

(6.) Reading abstract for the next talk; hearing an "echo" of the speaker's last sentence; fighting a feeling of tiredness; maybe feeling tingling on a tooth from a permanent retainer.

I'd count only sample (4) as an instance of attending to the content of the talk. In all of the other samples, the listener's mind is mainly off somewhere else, doing its own thing.

Completely absent – I'm not sure I've ever recorded an example of it – is the more active sort of engagement that I'd have thought I often do while listening to a talk: considering possible objections, thinking through consequences, sorting out the structure of the argument, thinking about connections to other people's work and to related issues.

Now it could be that Russ and I are unusually deadening speakers, but I don't think so. My guess is that most audience members, during most academic talks, spend most of their time with some distraction or other at the forefront of their stream of experience. They might not *remember* this fact about their experience of the talk – they might judge in retrospect that they were paying close attention most of the time – but that could be a matter of salience. They remember the moments they spent cooking up an objection; they forget their brief distracted thought about the color of the flyer.

The same is probably true about sexual thoughts. People often say they spend a lot of time thinking about sex, but when you actually beep them they very rarely report sexual thoughts.[172] My conjecture is that sexy thoughts, though rare, are likelier to be remembered than imagining biting into an apple, and so they are overrepresented in retrospective memory. In the form of an SAT analogy: Your devastating objection is to your thought about the color of the flyer as your vivid sexual fantasy is to your choice of breakfast cereal.

If it's true that samples (1)-(6) are representative of the kind of experiences people have when listening to academic talks, that invites this pair of conjectures about how people understand academic talks. Either

(A.) Our understanding of academic talks comes mostly from our ability to absorb them while other things are at the forefront of consciousness. The information soaks in, despite the near-constant layer of distraction, and that information then shapes skilled summaries of and reactions to the content of the talks.

Or

(B.) Our understanding of academic talks derives mostly from a small proportion of salient moments when we are not distracted. The understanding we exit with is a reconstruction of what must plausibly have been the author's view based on those scattered instances when we were actually paying attention.

I'd guess (B), but I don't really know. If (B) is true, that suggests that better techniques to retain or capture attention might have a dramatic effect on uptake.

We could do more beeping, more systematically, perhaps connected with comprehension measures, and really try to find this out. But of course, that's kind of hard to justify when the talk itself isn't about experience sampling.

Anecdotally, I've noticed two things that really gather an audience's attention. One is a novel, unpredicted physical event: A cat wanders onstage, or you tip over your coffee mug.[173] It's disheartening how much more fascinating the audience finds your spontaneous little mishap than they find your carefully prepared lecture! The other is the pause. If you just stop talking for a few seconds to gather your thoughts – fiddling with your notes or projector doesn't count – it's amazing how the audience seems to reconvene its attention. You can watch their eyes come to you. They almost hold their breath. For a moment, you can be nearly as fascinating as a falling coffee mug.

49. Why Metaphysics Is Always Bizarre

Bizarre views are a hazard of metaphysics. The metaphysician starts, seemingly, with some highly plausible initial commitments or commonsense intuitions – that there is a prime number between 2 and 5, that she could have had eggs for breakfast, that squeezing the clay statue would destroy the statue but not the lump of clay. She thinks long and hard about what, exactly, these claims imply. In the end, she finds herself positing a realm of abstract Platonic entities, or the real existence of an infinite number of possible worlds, or a huge population of spatiotemporally coincident things on her mantelpiece.[174] I believe that there isn't a single broad-ranging exploration of fundamental issues of metaphysics that doesn't ultimately entangle its author in seeming absurdities. Rejection of these seeming absurdities then becomes the commonsense starting point of a new round of metaphysics, by other philosophers, which in turn generates a complementary bestiary of metaphysical strangeness. Thus are philosophers happily employed.

I see three possible explanations of why philosophical metaphysics is never thoroughly commonsensical:

First possible explanation. A thoroughly commonsensical metaphysics wouldn't sell. It would be too obvious, maybe. Or maybe it would lack a kind of elegance or theoretical panache. Or maybe it would conflict too sharply with the sometimes un-commonsensical implications of empirical science.

The problem with this explanation is that there should be at least a small market for a thoroughly commonsensical metaphysics, even if that metaphysics is gauche, tiresome, and scientifically stale. Common sense might not be quite as fun as Nietzsche's eternal recurrence or Leibniz's windowless monads or as scientifically current as _____ [insert ever-changing example]; but a commonsense metaphysics ought to be attractive to at

least a certain portion of philosophers. At least it ought to command attention as a foil. It shouldn't be so downmarket as to be entirely invisible.

Second possible explanation. Metaphysics is very difficult. A thoroughly commonsensical metaphysics is out there to be discovered; we simply haven't found it yet. If all goes well, someday someone will piece it together, top to bottom, with no serious violence to common sense anywhere in the system.

I fear this is wishful thinking against the evidence. The greatest philosophers in history have worked at this for centuries, failing time and time again. Often, indeed, the most thorough metaphysicians – Leibniz, David Lewis – are the ones who generate the most stunningly strange systems.[175] It's not like we've made slow progress toward ever more commonsensical views, and we await a few more pieces to fall into place. There is no historical basis for the hope that a well-developed commonsense metaphysics will eventually arrive.

Third possible explanation. Common sense is incoherent in matters of metaphysics. Contradictions thus inevitably flow from it, and no coherent metaphysical system can adhere to it all. Although (as I also suggest in other chapters) ordinary common sense serves us fairly well in practical maneuvers through the social and physical world, common sense has proven an unreliable guide in cosmology, probability theory, microphysics, neuroscience, macroeconomics, evolutionary biology, structural engineering, medicine, topology…. If, as it seems to, metaphysics more closely resembles these latter endeavors than it resembles reaching practical judgments about picking berries and kissing, we might reasonably doubt the dependability of common sense as a guide to metaphysics.[176] Undependability doesn't imply incoherence, of course. But it seems a natural next step in this case, and it would neatly explain the historical facts at hand.

On the first explanation, we could easily enough invent a thoroughly commonsensical metaphysics if we wanted one, but we don't want one. On the second explanation, we do want one, or enough of us do, but we haven't yet managed to construct one. On the third explanation, we can't have one. The third explanation better fits the historical evidence and better acknowledges the likely epistemic limits of everyday human cognition.

#

Common sense might be culturally variable. So, whose common sense do I take to be at issue in this argument? I doubt it matters. All metaphysical systems in the philosophical canon, East and West, ancient and modern, I'm inclined to think, conflict both with the common sense of their milieu and with current Western Anglophone common sense. Eternal recurrence, windowless monads, Plato's Forms, carefully developed Buddhist systems of no-self and dependent origination, were never part of any society's common sense.

Let me be clear also, about the scope of my claim. It concerns only careful, broad-ranging explorations of fundamental metaphysical issues, especially issues where seeming absurdities tend to congregate: mind and body, causation, identity, the catalogue of entities that really exist. Some skating treatments, and some deep treatments of narrow issues might escape the charge.

Common sense changes. Heliocentrism used to defy common sense, but it no longer does. Maybe if we finally get our metaphysics straight, and then teach it patiently enough to generations of students, sowing it deeply into our culture, eventually people will say, "Ah yes, of course, windowless monads and eternal recurrence, what could be more plain and obvious to the common fellow?"

One can always dream!

#

Some of you might disagree about the existence of the phenomenon I aim to explain. You'll think there is a thoroughly commonsensical metaphysics already on the market. Okay, so who might count as a thoroughly commonsensical metaphysician?

Aristotle, I've sometimes heard. Or Scottish "common sense" philosopher Thomas Reid. Or G.E. Moore, famous for his "Defence of Common Sense". Or "ordinary language" philosopher P.F. Strawson. Or the later Wittgenstein. But Aristotle didn't envision himself as developing a commonsensical view. In the introduction to the *Metaphysics,* Aristotle says that the conclusions of sophisticated inquiries such as his own will often seem "wonderful" to the untutored and contrary to their initial opinions; and Aristotle sees himself as aiming to distinguish the true from the false in common opinion.[177] Moore, though fierce in wielding common sense against his foes, seems unable to preserve all commonsense opinions when he develops his positive views in detail, for example in his waffling about "sense-data".[178] Strawson struggles similarly, especially in his 1985 book, where he can find no satisfactory account of mental causation. Wittgenstein does not commit to a detailed metaphysical system. Despite frequently championing the idea of "common sense" philosophy, Reid acknowledges that in some areas common sense goes badly wrong and his opinions conflict with those of "the vulgar". He argues, for example, that without the constant intervention of immaterial souls, causation is impossible and objects can't even cohere into stable shapes.[179]

Since I cannot be expert in the entire history of global philosophy, maybe there's someone I've overlooked – a thorough and broad-ranging metaphysician who nowhere violates the common sense at least of their own culture. I welcome the careful exploration of other putative counterexamples to my generalization.

My argument is an empirical explanatory or "abductive" one. The empirical fact to be explained is that, across the entire history of philosophy, all well-developed metaphysical systems defy common sense. Every one of them is in some respect jaw-droppingly bizarre. The best explanation of this striking empirical fact is that people's commonsensical metaphysical intuitions form an incoherent set.

#

What should we conclude from this? Not, I think, that common sense must be abandoned. It's too essential to philosophical projects – common sense, or at least something that resembles it: pre-theoretical intuition, a prior sense of plausibility, culturally shared common ground, a set of starting points that we reject only very reluctantly. One has to start somewhere. But we shouldn't be too surprised if despite starting commonsensically, we end up, after sufficient inquiry, stuck having to choose among competing bizarrenesses.

50. The Philosopher of Hair

When I was a graduate student at U.C. Berkeley in the 1990s, the philosophy lounge had a billboard on which the graduate students posted philosophical humor. Among the items that lived a span upon that board was a newspaper clipping in which a French coiffeur claims not to be a barber but rather a "philosopher of hair". The humor in this, presumably, derives from the seemingly pretentious strangeness of a barber describing himself as a philosopher of hair. Philosophers, one might think, normally work in philosophy departments, or sit on lonely hills, or stew profoundly in St. Petersburg basements; they don't stand behind people's chairs with scissors. Hair doesn't seem like the right kind of thing to get all philosophical about, much less to build an identity as a philosopher around.

But why not? First off, lots of people find hair very important. We certainly spend a lot of collective time and money on it! It's intellectual snobbery to think that hair styling is an art below philosophical notice. Besides, aren't philosophers also supposed to be interested in Beauty, right alongside Truth and Morality? If a good thinning shears helps with Beauty, let's celebrate it.

Moreover, the following are recognizably philosophical questions:

(1.) What exactly is a haircut? For example, what distinguishes a haircut from a trim or a styling?

(2.) Is a good haircut timelessly good, or does the quality of a haircut depend on the currents of fashion?

(3.) Must a haircut please its bearer to be good? Or could a haircut be objectively great although its bearer hates it?

(4.) Should the nature and quality of a haircut be judged in part by the intent of the hairdresser? Or is it more of a "strict liability" thing, such that the nature and quality of the

haircut supervenes on the state of the hair on the head (and maybe surrounding fashion and the bearer's desires) but not at all on the hairdresser's intentions.

(5.) How important is it to have a good haircut, compared to other things one might value in life? Assuming it is important to have a good haircut, *why* is it important?

These questions resemble questions one sees in other areas of philosophy, especially aesthetics, and yet they aren't merely derivative of those other questions. The answers will involve factors particular to hair, and won't follow straightaway from one's general stance about the role of authorial intent in literature, what grounds quality judgments about museum-style paintings, etc. And I suspect that our French philosopher of hair is just bristling with opinions about these types of issues! ("So, you dislike your haircut? You know nothing!")

I care that there's a philosophy of hair partly because I care that we not misconstrue philosophy as a subject area. Philosophy should not be conceptualized as reflection on a particular set of specific, "profound" topics. Rather, philosophy is a style of thinking, a willingness to plunge in to consider the more fundamental ontological, normative, conceptual, and theoretical questions about anything. Any topic – the mind, language, physics, ethics, hair, Barbie dolls, carpentry, auto racing – can be approached philosophically. For all X, there's a philosophy of X.

51. Obfuscatory Philosophy as Intellectual Authoritarianism and Cowardice

I've been told that Kant and Hegel were poor writers whose impenetrable prose style is incidental to their philosophy. I've also been told that their views are so profound as to defy expression in terms comprehensible even to smart, patient, well-educated people who are not specialists in the philosophy of the period. I've heard similar things about Laozi, Heidegger, Plotinus, and Derrida. (I won't name any living philosophers.) I don't buy it.

Philosophy is not wordless profound insight. Philosophy is prose. Philosophy happens not in numinous moments of personal genius but in the creation of mundane sentences. It happens on the page, in the pen, through the keyboard, in dialogue with students and peers, and to some extent but only secondarily in private inner speech. If what exists on the page is not clear, the philosophy is not clear. Philosophers, like all specialists, profit from a certain amount of jargon, but philosophy need not become a maze of jargon. If private jargon doesn't regularly touch down in comprehensible public meanings, one has produced not philosophy but merely hazy words of indeterminate content. There are always gaps, confusions, indeterminacies, hidden assumptions, failures of clarity, even in the plainest philosophical writers, like Mozi, Descartes, Hume, and David Lewis. Thus, these philosophers present ample interpretative challenges. But the gaps, confusions, indeterminacies, hidden assumptions, and even to some extent the failures of clarity, are right there on the page, available to anyone who looks conscientiously for them, not sliding away into a nebulous murk.

If a philosopher can convince the public to take them seriously, being obfuscatory yields three illegitimate benefits. First, they intimidate the reader and by intimidation don a mantle of undeserved intellectual authority. Second, they disempower potential critics by

having a view of such indeterminate form that any criticism can be written off as based on a misinterpretation. Third, they exert a fascination on the kind of reader who enjoys the puzzle-solving aspect of interpretation, thus drawing from that reader a level of attention that may not be merited by the quality of their ideas (though this third benefit may be offset by alienating readers with a low tolerance for unclarity). These philosophers exhibit a kind of intellectual authoritarianism, with themselves as the assumed authority whose words we must spend time puzzling out. And simultaneously they lack intellectual courage: the courage to make plain claims that could be proven wrong, supported by plain arguments that could be proven fallacious. These three features synergize: If a critic thinks she has finally located a sound criticism, she can be accused of failing to solve the fascinating interpretative puzzle of the philosopher's superior genius.

Few philosophers, I suspect, deliberately set out to be obfuscatory. But I am inclined to believe that some are attuned to its advantages as an effect of their prose style and for that reason make little effort to write comprehensibly. Maybe they find their prose style shaped by audience responses: When they write clearly, they are dismissed or refuted; when they produce a fog of words that hint of profound meaning underneath, they earn praise. Maybe they are themselves to some extent victims – victims of a subculture, or a circle of friends, or an intended audience, that regards incomprehensibility as a sign of brilliance and so demands it in their heroes.[180]

52. Kant on Killing Bastards, Masturbation, Organ Donation, Homosexuality, Tyrants, Wives, and Servants

Immanuel Kant is among the best respected, best loved philosophers who has ever lived. His contributions to ethics, metaphysics, epistemology, and aesthetics have helped structure all four of those fields in Europe and the Americas for over two centuries. In the *Philosopher's Index* abstracts, Kant* ("*" is a truncation symbol) appears more frequently than similar searches for Plato, Aristotle, Hume, Confucius, Aquinas, or any other philosopher whose name I've tried. Philosophy departments even advertise specifically for specialists in Kant's philosophy – a treatment that no other philosopher receives even a quarter as frequently.[181]

Kant's most famous writings are notoriously abstract and difficult to understand. That's part, I suspect, of what makes him appealing to some readers: Kant interpretation can be a fun puzzle, and his abstractions invite fleshing out with plausible, attractive details.

When Kant himself fleshes out the details, it's often not so pretty.

#

For example, in *The Metaphysics of Morals* (1797/1996; not to be confused with the more famous, more abstract, and somewhat earlier *Groundwork for the Metaphysics of Morals*), Kant expresses the following views:

(1.) Wives, servants, and children are possessed in a way akin to our possession of objects. If they flee, they must be returned to the owner if he demands them, without regard for the cause that led them to flee. (See esp. pages 278, 282-284 [original pagination].) Kant does acknowledge that the owner is not permitted to treat these people as mere objects to "use

up", but this appears to have no bearing on the owner's right to demand their return. Evidently, if such an owned person flees to us from an abusive master, we may admonish the master for behaving badly while we return what is rightly his.

(2.) Homosexuality is an "unmentionable vice" so wrong that "there are no limitations whatsoever that can save [it] from being repudiated completely" (p. 277).

(3.) Masturbation is in some ways a worse vice than the horror of murdering oneself, and "debases [the masturbator] below the beasts". Kant writes:

> But it is not so easy to produce a rational proof that unnatural, and even merely unpurposive, use of one's sexual attribute is inadmissible as being a violation of duty to oneself (and indeed, as far as its unnatural use is concerned, a violation in the highest degree). The ground of proof is, indeed, that by it a man surrenders his personality (throwing it away), since he uses himself as a means to satisfy an animal impulse. But this does not explain the high degree of violation of the humanity in one's own person by such a vice in its unnaturalness, which seems in terms of its form (the disposition it involves) to exceed even murdering oneself. It consists, then, in this: That a man who defiantly casts off life as a burden is at least not making a feeble surrender to animal impulse in throwing himself away (p. 425).

(4.) On killing bastards:

> A child that comes into the world apart from marriage is born outside the law (for the law is marriage) and therefore outside the protection of the law. It has, as it were, stolen into the commonwealth (like contraband merchandise), so that the commonwealth can ignore its existence (since it rightly should not have come to exist in this way), and can therefore also ignore its annihilation (p. 336).

On the face of it, similar reasoning might seem to apply to people who enter a country illegally. As far as I'm aware, though, Kant doesn't address that issue.

(5.) On organ donation:

> To deprive oneself of an integral part or organ (to maim oneself) – for example, to give away or sell a tooth to be transplanted into another's mouth... are ways of partially murdering oneself... cutting one's hair in order to sell it is not altogether free from blame. (p. 423)

(6.) Servants and women "lack civil personality and their existence is, as it were, only inherence" and thus should not be permitted to vote or take an active role in the affairs of state (p. 314-315).

(7.) Under no circumstances is it right to resist the legislative head of state or to rebel on the pretext that the ruler has abused his authority (p. 319-320). Of course, the ruler is *supposed* to treat people well – but (as with wives and servants under abusive masters) there appears to be no legitimate means of escape if he does not.

These views are all, I hope you will agree, odious.[182] So too, is Kant's racism, which doesn't show itself so clearly in the *Metaphysics of Morals* but is painfully evident in other of his writings, for example, "Of the Different Human Races", where he argues that the Negro, biologically, is "lazy, soft, and trifling" (1775/2007, p. 438) and "On the Feeling of the Beautiful and Sublime", where he asserts that the "Negroes of Africa have by nature no feeling that rises above the ridiculous" (1764/2007, p. 253).[183]

#

You might say, so what? Kant was a creature of his time, as are we all. No one is a perfect discoverer of moral truths. In two centuries, John Rawls, Peter Singer, Martha

Nussbaum, and Bernard Williams might look similarly foolish in some of their opinions (if they don't already).

Well, sure! Maybe that's comforting to know. And yet I don't think that all ethicists' worldviews age equally badly. There's a humaneness and anachronistic egalitarianism that I think I hear in the ancient Chinese philosopher Zhuangzi, and in the sixteenth-century French philosopher Montaigne, which has aged better, to my ear.[184]

Some interpreters of Kant ask the reader for a considerable patience and deference to his genius. Looking at the passages above, though, it should be clear that Kant's arguments do not always deserve patience and deference. When reading a passage of Kant's with which you are inclined to disagree, bear in mind that among the interpretive possibilities is this: He's just being a vile, boneheaded doofus.

From our cultural distance, it is evident that Kant's arguments against masturbation, for the return of wives to abusive husbands, etc., are shoddy stuff. This should make us suspicious that there might be other parts of Kant, too, where the arguments are shoddy, even if those arguments are too abstract to generate a vividly odious conclusion that alerts us to this fact. I'd suggest that among the candidates for being shoddy work, incoherent and bad, are the "transcendental deduction", which stands at the heart of Kant's *Critique of Pure Reason,* and which defies consensus interpretation; and Kant's claim, near the heart of his ethics, that his three seemingly obviously non-equivalent formulations of the "categorical imperative" are in fact equivalent.[185] For the reader unfamiliar with the transcendental deduction and the formulations of the categorical imperative, suffice to say that these are among the most famous and influential aspects of Kant's work, and consequently among the most famous and influential bits of philosophy done by anyone ever.

I might be entirely wrong in my suspicions about the transcendental deduction and the three formulations. It's perfectly reasonable to think that I am probably wrong! Interpreters

have put quite a bit of effort into trying to make sense of them, and the work has been influential and even inspirational to many subsequent philosophers. How could such influential, widely admired work be rotten? From a bird's eye view, so to speak, I think my suspicions must be wrong. But from a worm's eye view, looking directly at what's on the page, well….

Kant specialists will tend to disagree with my negative opinion of the transcendental deduction and of Kant's claim of the equivalence of the three formulations of the categorical imperative. Should I defer to them? You'd think they'd know, if anyone does! Yet I worry that a person does not embark on becoming a Kant specialist without already having a tendency to give Kant more trust and charity than he may deserve. Continuing the long journey far into Kant scholarship might then further aggravate the specialist's initial excessive charity and trust: After dedicating years of one's career to Kant, it might be difficult not to see his work as worth the immense effort one has poured into him. In disputes about the quality of Kant's work, Kant specialists are not neutral parties.

If you'll forgive me, here's a very uncharitable theory of Kant: He is a master at promising philosophers what they long for – such as a decisive refutation of radical skepticism or a proof that it's irrational to be immoral – and then effusing a haze of words, some confident statements, some obscure patches of seemingly relevant argumentation, some intriguing suggestions, with glimmers enough of hope that readers can convince themselves that a profound solution lies beneath, if only they understood it.[186]

53. Nazi Philosophers, World War I, and the Grand Wisdom Hypothesis

As described in Chapter 4, I've done a fair bit of empirical research on the moral behavior of ethics professors. My collaborators and I have consistently found that ethicists behave no better than socially comparable non-ethicists. However, the moral violations that we've examined have mostly been minor: stealing library books, neglecting student emails, littering, forgetting to call mom. Some behaviors are arguably much more significant – donating large amounts to charity, vegetarianism – but there's certainly no consensus about the moral importance of those things. Sometimes I hear the objection that the moral behavior I've studied is all trivial stuff – that even if ethicists behave no better in day-to-day ways, on issues of great moral importance – decisions that reflect on one's overarching worldview, one's broad concern for humanity, one's general moral vision – professional ethicists, and professional philosophers in general, might show greater wisdom. Call this the Grand Wisdom Hypothesis.

Now let's think about Nazis.

Nazism is an excellent test case of the Grand Wisdom Hypothesis since everyone now agrees that Nazism is extremely morally odious. Germany had a robust philosophical tradition in the 1930s, and excellent records are available on individual professors' participation in or resistance to the Nazi movement. So, we can ask: Did a background in philosophical ethics serve as any kind of protection against the moral delusions of Nazism? Or were ethicists just as likely to be swept up in noxious German nationalism as were others of their social class? Did reading Kant on the importance of treating all people as "ends in themselves" help philosophers better see the errors of Nazism, or instead did philosophers tend to appropriate Kant for anti-Semitic and expansionist purposes?

Heidegger's involvement with Nazism is famous and much discussed, but he's only one data point. There were also, of course, German philosophers who opposed Nazism, possibly partly – if the Grand Wisdom Hypothesis is correct – because of their familiarity with theoretical ethics. My question is quantitative: Were philosophers as a group *any more likely* than other academics to oppose Nazism, or any less likely to be enthusiastic supporters? I am not aware of any careful, quantitative attempts to address this question.

There's a terrific resource on ordinary German philosophers' engagement with Nazism: George Leaman's (1993) *Heidegger im Kontext,* which contains a complete list of all German philosophy professors from 1932 to 1945 and provides summary data on their involvement with or resistance to Nazism.

In Leaman's data set, I count 179 philosophers with "habilitation" in 1932 when the Nazis started to ascend to power, including "dozents" and "ausserordentlichers" but not assistants. ("Habilitation" is an academic achievement after the PhD, with no equivalent in the Anglophone world but roughly comparable in its requirements to gaining tenure in the U.S.) I haven't attempted to divide these philosophers into ethicists and non-ethicists, since the ethics/non-ethics division wasn't as sharp then as it is now in twenty-first century Anglophone philosophy. (Consider Heidegger again. In a sense he's an ethicist since he writes among other things on the question of how one should live, but his interests range broadly.) Of these 179 philosophers, 58 (32%) joined the Nazi party.[187] This compares with estimates of about 21%-25% Nazi party membership among German professors as a whole.[188] Philosophers were thus not underrepresented in the Nazi party.

To what extent did joining the Nazi party reflect enthusiasm for its goals vs. opportunism vs. a reluctant decision under pressure?

I think we can assume that membership in either of the two notorious Nazi paramilitary organizations, the SA or the SS, reflects either enthusiastic Nazism or an unusual

degree of self-serving opportunism: Membership in these organizations was by no means required for continuation in a university position. Among philosophers with habilitation in 1932, two (1%) joined the SS and another 20 (11%) joined (or were already in) the SA (one philosopher joined both), percentages approximately similar to the overall academic participation in these organizations. I suspect that this estimate substantially undercounts enthusiastic Nazis, since a number of philosophers (including briefly Heidegger) appear to have gone beyond mere membership to enthusiastic support through their writings and other academic activities, despite not joining the SA or SS. One further possible measure is involvement with Alfred Rosenberg, the notorious Nazi racial theorist. Combining the SA, SS, and Rosenberg associates yields a minimum of 30 philosophers (17%) on the far right side of Nazism – not even including those who received their posts or habilitation after the Nazis rose to power (and thus perhaps partly because of their Nazism). By this time, Hitler's book *Mein Kampf,* was widely known and widely circulated, proudly proclaiming Hitler's genocidal aims. Almost a fifth of professional philosophers thus embraced a political worldview that is now rightly regarded as a paradigm example of evil.

Among philosophers who were not party members, 22 (12%) were "Jewish" (by the broad Nazi definition). Excluding these from the total leaves 157 non-Jewish philosophers with habilitation before 1933. The 58 Nazis thus constituted 37% of established philosophers who had the opportunity to join the party. Of the remainder, 47 (30%) were deprived of the right to teach, imprisoned, or otherwise severely punished by the Nazis for Jewish family connections or political unreliability. (This second number excludes five philosophers who were Nazi party members but also later severely penalized.) It's difficult to know how many of this group took courageous stands vs. found themselves intolerable for reasons outside of their control. The remaining 33% we might think of as "coasters" – those who neither joined the party nor incurred severe penalty. Most of these coasters had at least token Nazi

affiliations, especially with the NSLB (the Nazi organization of teachers), but probably NSLB affiliation alone did not reflect much commitment to the Nazi cause.

If joining the Nazi party were necessary for simply getting along as a professor, membership in the Nazi party would not reflect much commitment to Nazism. The fact that about a third of professors could be coasters suggests that token gestures of Nazism, rather than actual party membership, were sufficient, as long as one did not actively protest or have Jewish affiliations. Nor were the coasters mostly old men on the verge of retirement (though there was a wave of retirements in 1933, the year the Nazis assumed power). If we include only the subset of 107 professors who were not Jewish, habilitated before 1933, and continuing to teach past 1940, we still find 30% coasters (or 28% excluding two emigrants).

The existence of unpublished coasters shows that philosophy professors were not forced to join the Nazi party. Nevertheless, a substantial proportion did so voluntarily, either out of enthusiasm or opportunistically for the sake of career advancement. A substantial minority, at least 19% of the non-Jews, occupied the far right of the Nazi party, as reflected by membership in the SS, SA, or association with Rosenberg. It is unclear whether pressures might have been greater on philosophers than on other disciplines, but there was substantial ideological pressure on many disciplines: There was also Nazi physics (no Jewish relativity theory, for example), Nazi biology, Nazi history, etc. Given the possible differences in pressure, and the lack of a dataset strictly comparable to Leaman's for the professoriate as a whole, I don't think we can conclude that philosophers were especially more likely to endorse Nazism than were other professors. However, I do think it is reasonable to conclude that they were not especially *less* likely.

Nonetheless, given that about a third of non-Jewish philosophers were severely penalized by the Nazis (including one executed for resistance and two who died in

concentration camps), it remains possible that philosophers are overrepresented among those who resisted or were ejected. I have not seen quantitative data that bear on this question.

#

In doing background reading for the analysis I've just described, I was struck by the following passage from Fritz Ringer's 1969 classic *Decline of the German Mandarins:*

> Early in August of 1914, the war finally came. One imagines that at least a few educated Germans had private moments of horror at the slaughter which was about to commence. In public, however, German academics of all political persuasions spoke almost exclusively of their optimism and enthusiasm. Indeed, they greeted the war with a sense of relief. Party differences and class antagonisms seemed to evaporate at the call of national duty.... intellectuals rejoiced at the apparent rebirth of "idealism" in Germany. They celebrated the death of politics, the triumph of ultimate, apolitical objectives over short-range interests, and the resurgence of those moral and irrational sources of social cohesion that had been threatened by the "materialistic" calculation of Wilhelmian modernity.
>
> On August 2, the day after the German mobilization order, the modernist [theologian] Ernst Troeltsch spoke at a public rally. Early in his address, he hinted that "criminal elements" might try to attack property and order, now that the army had been moved from the German cities to the front. This is the only overt reference to fear of social disturbance that I have been able to discover in the academic literature of the years 1914-1916.... the German university professors sang hymns of praise to the "voluntary

> submission of all individuals and social groups to this army." They were almost grateful that the outbreak of war had given them the chance to experience the national enthusiasm of those heady weeks in August (Ringer 1969, 180-181).

With the notable exception of Bertrand Russell (who lost his academic post and was imprisoned for his pacifism), philosophers in England appear to have been similarly enthusiastic. Ludwig Wittgenstein never did anything so cheerily, it seems, as head off to fight as an Austrian foot-soldier. Alfred North Whitehead rebuked his friend and co-author Russell for his opposition to the war and eagerly sent off his sons North and Eric. (Eric Whitehead died.) French philosophers appear to have been similarly enthusiastic. It's as though, in 1914, European philosophers rose as one to join the general chorus of people proudly declaring "Yay! World war is a great idea!"

If there is anything that seems, in retrospect, plainly, head-smackingly obviously *not* to have been a great idea, it was World War I, which destroyed millions of lives to no purpose. At best, it should have been viewed as a regrettable, painful necessity in the face of foreign aggression, which hopefully could soon be diplomatically resolved; yet that seems rarely to have been the mood of academic thought about war in 1914. Philosophers at the time were evidently no more capable of seeing the obvious (?) downsides of world war than was anyone else. Even if the downsides of war were, in the period, not entirely obvious upon careful reflection – the glory of Bismarck and all that? – with a few rare and ostracized exceptions, philosophers and other academics showed little of the special foresight and broad vision required by the Grand Wisdom Hypothesis.

Here's a model of philosophical reflection on which philosophers' enthusiasm for World War I is unsurprising: Philosophers – and everyone else – possess their views about the big questions of life for emotional and sociological reasons that have little to do with their

philosophical theories and academic research. They recruit Kant, Mill, Locke, Rousseau, Aristotle, etc., only after the fact to justify what they would have believed anyway. Moral and political philosophy is nothing but post-hoc rationalization.

Here's a model of philosophical reflection on which philosophers' enthusiasm for World War I is, in contrast, surprising: Reading Kant, Mill, Locke, Rousseau, Aristotle, etc., helps induce a broadly humanitarian view, helps you see that people everywhere deserve respect and self-determination, moves you toward a more cosmopolitan worldview that doesn't overvalue national borders, helps you gain critical perspective on the political currents of your own time and country, helps you better see through the rhetoric of demagogues and narrow-minded politicians.

Both models are of course too simple.

#

When I was in Berlin in 2010, I spent some time in the Humboldt University library, browsing philosophy journals from the Nazi era. The journals differed in their degree of alignment with the Nazi worldview. Perhaps the most Nazified was *Kant-Studien*, which at the time was one of the leading German-language journals of general philosophy (not just a journal for Kant scholarship). The old issues of *Kant-Studien* aren't widely available, but I took some photos. Below, Sascha Fink and I have translated the preface to *Kant-Studien* Volume 40 (1935):

> *Kant-Studien,* now under its new leadership that begins with this first issue of the 40th volume, sets itself a new task: to bring the new will, in which the deeper essence of the German life and the German mind is powerfully realized, to a

breakthrough in the fundamental questions as well as the individual questions of philosophy and science.

Guiding us is the conviction that the German Revolution is a unified metaphysical act of German life, which expresses itself in all areas of German existence, and which will therefore – with irresistible necessity – put philosophy and science under its spell.

But is this not – as is so often said – to snatch away the autonomy of philosophy and science and give it over to a law alien to them?

Against all such questions and concerns, we offer the insight that moves our innermost being: That the reality of our life, that shapes itself and will shape itself, is deeper, more fundamental, and more true than that of our modern era as a whole – that philosophy and science, which compete for it, will in a radical sense become liberated to their own essence, to their own truth. Precisely for the sake of truth, the struggle with modernity – maybe with the basic norms and basic forms of the time in which we live – is necessary. It is – in a sense that is alien and outrageous to modern thinking – to recapture the form in which the untrue and fundamentally destroyed life can win back its innermost truth – its rescue and salvation. This connection of the German life to fundamental forces and to the original truth of Being and its order – as has never been attempted in the same depth in our entire history – is what we think of when we hear that word of destiny: a new Reich.

If on the basis of German life German philosophy struggles for this truly Platonic unity of truth with historical-political life, then it takes up a European duty. Because it poses the problem that each European people must solve, as a necessity of life, from its own individual powers and freedoms.

Again, one must – and now in a new and unexpected sense, in the spirit of Kant's term, "bracket knowledge" [das Wissen aufzuheben]. Not for the sake of negation: but to gain space for a more fundamental form of philosophy and science, for the new form of spirit and life [für die neue Form ... des Lebens Raum zu gewinnen]. In this living and creative sense is *Kant-Studien* connected to the true spirit of Kantian philosophy.

So, we call on the productive forces of German philosophy and science to collaborate in these new tasks. We also turn especially to foreign friends, confident that in this joint struggle with the fundamental questions of philosophy and science, concerning the truth of Being and life, we will gain not only a deeper understanding of each other, but also develop an awareness of our joint responsibility for the cultural community of peoples.

– H. Heyse, Professor of Philosophy, University of Königsberg

#

Is it just good cultural luck – the luck of having been born into the right kind of society – that explains why 21st-century Anglophone philosophers reject such loathsome worldviews? Or is it more than luck? Have we somehow acquired better tools for rising above our cultural prejudices?

Or – as I'll suggest in Chapter 58 – ought we entirely refrain from self-congratulation, whether for our luck or our skill? Maybe we aren't so different, after all, from the early 20th-century Germans. Maybe we have our own suite of culturally-shared moral defects, invisible to us or obscured by a fog of bad philosophy.

54. Against Charity in the History of Philosophy

In 2016, Peter Adamson, host of the podcast *History of Philosophy Without Any Gaps*, posted twenty "Rules for the History of Philosophy". Mostly they are fine rules. I want to quibble with one.

Like almost every historian of philosophy I know, Adamson recommends that we be "charitable" to the text. Here's how he puts it in "Rule 2: Respect the text":

> This is my version of what is sometimes called the "principle of charity." A minimal version of this rule is that we should assume, in the absence of fairly strong reasons for doubt, that the philosophical texts we are reading make sense.... [It] seems obvious (to me at least) that useful history of philosophy doesn't involve looking for inconsistencies and mistakes, but rather trying one's best to get a coherent and interesting line of argument out of the text. This is, of course, not to say that historical figures never contradicted themselves, made errors, and the like, but our interpretations should seek to avoid imputing such slips to them unless we have tried hard and failed to find a way of resolving the apparent slip.

At a first pass, it seems like a good idea, if at all possible, to avoid imputing contradictions and errors, and to seek a coherent sensible interpretation of historical texts. This is how, it seems, to best "respect the text".

To see why I think charity isn't as good an idea as it seems, let me first mention my main reason for reading history of philosophy: It's to gain a perspective, through the lens of distance, on my own philosophical views and presuppositions, and on the philosophical attitudes and presuppositions of twenty-first century Anglophone philosophy generally. Twenty-first century Anglophone philosophers, for example, tend to assume that the world is wholly material – or if they reject that view, they tend to occupy one of a few well-known

alternative positions (e.g., Christian theism, naturalistic "property dualism"). I'm inclined to accept the majority's materialism. Reading the history of philosophy reminds me that a wide range of other views have been taken seriously over time. Similarly, twenty-first century Anglophone philosophers tend to favor a certain species of liberal ethics, with an emphasis on individual rights and comparatively little deference to traditional rules and social roles – and I tend to favor such an ethics too. But it's good to be vividly aware that historically important thinkers have often had very different moral opinions, which they felt they could adequately justify. Reading culturally distant texts reminds me that I am a creature of my era, with views that have been shaped by contingent social factors.

Others might read history of philosophy with very different aims, of course.

Question: If my main aim in reading history of philosophy is to appreciate the historical diversity of philosophical views, what is the most counterproductive thing I could do when confronting a historical text?

Answer: Interpret the author as endorsing a view that is familiar, "sensible", and similar to my own and my colleagues'.

Historical texts, like all philosophical texts – but more so, given our linguistic and cultural distance – tend to be difficult and ambiguous. Therefore, they will admit of multiple interpretations. Suppose, then, that a text admits of four possible interpretations, A, B, C, and D, where Interpretation A is the least challenging, least weird, and most sensible, and Interpretation D is the most challenging, weirdest, and least sensible. A simple application of the principle of charity seems to recommend that we favor the sensible, pedestrian Interpretation A, the interpretation that, in our view, makes the most sense and avoids the most errors. However, weird and wild Interpretation D might be the one we would learn most from taking seriously, challenging our presuppositions more deeply and giving us a

more helpfully divergent perspective. This is one reason to favor Interpretation D. Call this the *Principle of Anti-Charity*.

This way of defending Anti-Charity might seem bluntly instrumentalist. What about historical accuracy? Don't we want the interpretation that's most likely to be *correct* interpretation?

Bracketing post-modern views that reject truth in textual interpretation, I have four responses to that concern:

(1.) Being Anti-Charitable doesn't mean that anything goes. You still want to respect the surface of the text. If the author says "P", you don't want to attribute the view that not-P. In fact, it is the more "charitable" views that are likely to take an author's claims other than at face value: "Kant seems to say that it's permissible to kill children who are born out of wedlock, but really a charitable, sensible interpretation in light of X, and Y, and Z is that he really meant…." In one way, it is actually more respectful to texts *not* to be too charitable, and to interpret the text superficially at face value. After all, P is what the author literally said.

(2.) What seems "coherent" and "sensible" is culturally variable. You might reject excessive charitableness, while still wanting to limit allowable interpretations to one among several sensible and coherent ones. But this might already be too limiting. It might not seem "coherent" to us to embrace a contradiction, but some philosophers in some traditions seem happy to accept some bald contradictions.[189] It might not seem sensible to think that the world is nothing but a flux of ideas, the existence of rocks depending on the states of immaterial spirits; so, in the spirit of charity, if there's any ambiguity, you might prefer an interpretation that you find less metaphysically peculiar. But metaphysical idealism is now at a low ebb by world historical standards, so this strategy might lead you away from rather than toward interpretive accuracy.

(3.) Philosophy is hard and philosophers are stupid. This human mind is not well-designed for figuring out philosophical truths. Timeless philosophical puzzles tend to kick our collective butts. Sadly, this is going to be true of your favorite philosopher too. The odds are good that this philosopher, being a flawed human like you and me, made mistakes, fell into contradictions, changed opinions, and failed to see what seem in retrospect to be obvious consequences and counterexamples. Great philosophers can be great fools, and indeed the foolishness of rejecting assumptions that are widespread in your culture, without appreciating the alarming consequences for other views you hold, is sometimes exactly what propels philosophy forward. (An example might be Kant's egalitarian abstract ethics alongside his inegalitarian views of race, class, and gender; Chapter 52.) Respecting the text and respecting the person means, in part, not trying too hard to smooth this stuff away. The warts can even be part of the loveliness. Noticing them in your favorite philosopher can also be tonic against excessive hero worship and a reminder of your own likely warts and failings.

(4.) Some authors might not even *want* to be interpreted as having a coherent, stable view. Zhuangzi, Montaigne, Nietzsche, and the later Wittgenstein all might be interpreted as expressing philosophical opinions that they don't expect to form an entirely coherent set.[190] If so, attempting "charitably" to stitch together a coherent picture might be a failure to respect the philosopher's own aims and intentions as expressed in the text.

I prefer uncharitable interpretation and the naked wartiness of the text. Refuse to hide the weirdness and the plain old wrongness and badness. Refuse to dress the text in sensible twenty-first century garb.

55. Invisible Revisions

Imagine an essay manuscript: Version A. Monday morning, I read through Version A. I'm not satisfied. Monday, Tuesday, Wednesday, I revise and revise – cutting some ideas, adding others, tweaking the phrasing, trying to perfect the manuscript. Wednesday night I have the new version, Version B. My labor is complete. I set it aside.

Three weeks later, I re-read the manuscript – Version B, of course. It lacks something. The ideas I had made more complex seem now too complex. They lack vigor. Conversely, what I had simplified for Version B now seems flat and cartoonish. The new sentences are clumsy, the old ones better. My first instincts had been right, my second thoughts poor. I change everything back to the way it was, one piece at a time, thoughtfully. Now I have Version C – word-for-word identical with Version A.

To your eyes, Version A and Version C look the same, but I know them to be vastly different. What was simplistic in Version A is now, in Version C, elegantly simple. What I overlooked in Version A, Version C instead subtly finesses. What was rough prose in Version A is now artfully casual. Every sentence of Version C is deeper and more powerful than in Version A. A journal would rightly reject Version A but rightly accept Version C.[191]

56. On Being Good at Seeming Smart

Once upon a time, there was a graduate student at UC Riverside who I will call Student X. The general sense among the faculty that Student X was particularly promising. For example, after a colloquium at which the student had asked a question, one faculty member expressed to me how impressive the student was. I was struck by that remark because I had thought the student's question had actually been rather poor. But it occurred to me that the question had seemed, superficially, to be smart. That is, if you didn't think too much about the content but rather just about the tone and delivery, you probably would get a strong impression of smartness. In fact, my overall view of this student was that he was about average – neither particularly good nor particularly bad – but that he was a master of *seeming smart:* He had the confidence, the delivery, the style, all the paraphernalia of smartness, without an especially large dose of the actual thing.

Mostly, I've noticed, it's White men from upper- and upper-middle class backgrounds who are described in my presence as "seeming smart". It's really quite a striking pattern. (I've been taking a tally, since I first started becoming interested in the phenomenon.) This makes sense, in a way. When the topic of conversation is complex and outside of one's specific expertise – in other words, most of philosophy even for most professional philosophers – seeming smart is probably to a large extent about activating people's *associations* with intelligence as one discusses that topic. This can be done through poise, confidence (but not defensiveness), giving a moderate amount of detail but not too much, and providing some frame and jargon, and also, I suspect, unfortunately, in part by having the right kind of look and physical bearing, a dialect that is associated with high education levels, the right prosody (e.g., not the "Valley girl" habit of ending sentences with a rising intonation), the right body language. If you want to "seem smart", it helps immensely, I think, to just *sound* right, to have a "smart professor voice" in your toolkit, to be comfortable

in an academic setting, to just strike the listener at a gut level as someone who belongs. Who will tend to have those tools and habits and feelings of confidence, and who will feel like they naturally belong, and who will strike those in power as "one of us"? Unsurprisingly, it's typically the people who culturally resemble those who hold the majority of academic power. Me, for example – the White male professor's kid from an affluent suburb who went to Stanford. But philosophy as a discipline shouldn't be so dominated by my social group. The kid from the inner city whose parents never went to college will also have some interesting things to say, even if she's not so good at professor voice.

#

Student X actually ended up doing very well in the program and writing an excellent dissertation. He rose to his teachers' expectations – as students often do.[192] His terrific skill at seeming smart paid off handsomely in attracting positive attention and the support and confidence of his professors, which probably helped him flourish over the long haul of the PhD program. Conversely, the students not as good at the art of seeming smart often sink to their teachers' low expectations – frustrated, ignored, disvalued, criticized, made to feel not at home, as I have also seen.

I hereby renew my resolution to view skeptically all judgments of "seeming smart". Let's try to appreciate instead, the value and potential in the young scholar who seems superficially not to belong, who seems to be awkward and foolish and out of their depth, but who has somehow made it into the room anyway. They've probably fought a few more battles to get there, and they might have something different and interesting to say, if we're game to listen.

57. Blogging and Philosophical Cognition

Academic philosophers tend to have a narrow view of what constitutes valuable philosophical research. Hiring, tenure, promotion and prestige depend mainly on one's ability to produce journal articles in a particular theoretical, abstract style, mostly in reaction to a small group of canonical and recently influential thinkers, for a small readership of specialists. We should broaden our vision.

Consider the historical contingency of the journal article, a late-19th century invention. Even as recently as the middle of the 20th century, influential philosophers in Western Europe and North America did important work in a much broader range of genres: the fictions and difficult-to-classify reflections of Sartre, Camus, and Unamuno; Wittgenstein's cryptic fragments; the peace activism and popular writings of Bertrand Russell; John Dewey's work on educational reform. Popular essays, fictions, aphorisms, dialogues, autobiographical reflections and personal letters have historically played a central role in philosophy. So also have public acts of direct confrontation with the structures of one's society: Socrates' trial and acceptance of the lethal hemlock; Confucius' inspiring personal correctness. It was really only with the generation hired to teach the baby boomers in the 1960s and 1970s that academic philosophers' conception of philosophical work became narrowly focused on the technical journal article.

In the 21st century, we have an even wider selection of media. Is there reason to think that journal articles are uniformly better for philosophical reflection than videos, interactive demonstrations, blog posts, or multi-party conversations in social media? A conversation on Facebook, if good participants bring their best to the enterprise, has the potential to be a philosophical creation of the highest order, with a depth and breadth beyond the capacity of any individual philosopher to create. A video game could illuminate, critique, and advance a vision of worthwhile living, deploying sight, hearing, emotion and personal

narrative as well as (why not?) traditional verbal exposition – and it could potentially do so with all the freshness of thinking, all the transformative power and all the expository rigor of Aristotle, Xunzi, or Hume. Academic philosophers are paid to develop expertise in philosophy, to bring that expertise into the classroom and to contribute that expertise to society in part by advancing philosophical knowledge. A wide range of activities fit within that job description.

Every topic of human concern is open to philosophical inquiry. This includes not only subjects well represented in journals, such as the structure of propositional attitudes and the nature of moral facts, but also how one ought to raise children and what makes for a good haircut (Chapter 50). The method of writing and responding to journal-article-length expository arguments by fellow philosophers is only one possible method of inquiry. Engaging with the world, trying out one's ideas in action, seeing the reactions of non-academics, exploring ideas in fiction and meditation – in these activities we can not only deploy knowledge but cultivate, expand, and propagate that knowledge.

Philosophical expertise is not like scientific expertise. Although academic philosophers know certain literatures very well, on questions about the general human condition and what our fundamental values should be, knowledge of the canon gives academic philosophers no especially privileged wisdom. Non-academics can and should be respected partners in the philosophical dialogue. Too exclusive a focus on technical journal articles excludes non-academics from the dialogue – or maybe, better said, excludes us philosophers from non-academics' more important dialogue. The academic journal article as it exists today is too limited in format, topic, method, and audience to deserve so centrally privileged a place in philosophers' conception of the discipline.

If one approaches popular writing only as a means of "dumbing down" preexisting philosophical ideas for an audience of non-experts whose reactions one doesn't plan to take seriously, then one is squandering a great research opportunity. The popular essay can be a locus of philosophical creativity, in which ideas are explored in hope of discovering new possibilities, advancing (and not just marketing) one's own thinking in a way that might strike professionals too as interesting. Analogously for government consulting, Twitter feeds, TED videos, and poetry.

A *Philosophical Review* article can be an amazing thing. But we should see journal articles in that style, in that type of venue, as only one of many possible forms of important, field-shaping philosophical work.

#

The eight-hundred-word blog post deserves, I think, special praise, for three distinct reasons:

(1.) *Short, fat, tangled arguments.*

In her 2015 Dewey lecture at the Pacific Division meeting of the American Philosophical Association, philosopher of science Nancy Cartwright celebrated what she called "short, stocky, tangled" arguments over "tall, skinny" arguments.

Here's a tall, skinny, neat argument:

A

$A \rightarrow B$

$B \rightarrow C$

$C \rightarrow D$

$D \rightarrow E$

E → F

F → G

G → H

H → I

I → J

Therefore, J. (Whew!)

And here's a short, fat, tangled argument:

A1	A2	A3
A1 → B	A2 → B	A1 → A3
Therefore, B.	A2 → A1	A3 → A2
	Therefore, B.	A3 → B
		Therefore, B.

The tall, lean argument takes you straight like an arrowshot all the way from A to J. All the way from the fundamental nature of consciousness to the reforms and downfall of Napoleon. (Yes, I'm thinking of you, Georg Wilhelm Friedrich.[193]) All the way from seven abstract Axioms to Proposition V.42 "Blessedness is not the reward of virtue, but virtue itself; nor do we enjoy it because we restrain our lusts; on the contrary, because we enjoy it, we are able to restrain them". (Sorry, Baruch, I wish I were more convinced![194])

In contrast, the short, fat, tangled argument only takes you from versions of A to B. But it does so in three ways, so that if one argument fails, the others remain. It does so without needing a long string of possibly dubious intermediate claims. And the different premises lend tangly sideways support to each other. I think here of the ancient Chinese philosopher Mozi's dozen arguments for impartial concern or the ancient Greek philosopher Sextus's many modes of skepticism.[195]

In areas like mathematics, tall, skinny arguments can work. Maybe the proof of Fermat's last theorem is one – long and complicated, but apparently sound. (Not that I would be any authority.) When each step is secure, tall arguments succeed and take us to wonderful heights. But philosophy tends not to have such secure intermediate steps.

The human mind is great at determining an object's shape from its shading. The human mind is great at interpreting a stream of incoming sound as a sly dig on someone's character. The human mind is horrible at determining the soundness of philosophical arguments, and also at determining the soundness of most intermediate stages within philosophical arguments. (If these remarks sound familiar from my other chapters, that's part of the fat tangle I seek.) Tall, skinny philosophical arguments – this was Cartwright's point – will almost inevitably topple. Even the short arguments usually fail – as I believe they do in Mozi and Sextus – but at least they have a shot.

Individual blog posts are short. They are, I think, just about the right size for human philosophical cognition: 500-1000 words, long enough to put some flesh on an idea, making it vivid (pure philosophical abstractions being almost impossible to evaluate for multiple reasons), long enough to make one or maybe two novel turns or connections, but short enough that the reader can reach the end without having lost track of the path there, keeping more or less the whole argument simultaneously in view.

In the aggregate, blog posts are fat and tangled: Multiple posts can drive toward the same conclusion from diverse angles. Multiple posts can lend sideways support to each other. I've written many posts, for example, that are skeptical of expert philosophical cognition – several of which have been adapted for this book. I've written many posts, also adapted for this book, that are skeptical of moral self-knowledge, and many posts that aim to express my wonder at the possible weirdness of the world. They synergize, perhaps, in conveying my philosophical vision of the world – but they do not stack up into a tall, topply

tower. If some or even most of them fail individually, the general picture can still stand supported.

Of course, there's also something to be said for trying to build a ladder to the Moon. Maybe you'll be the one who finally gets there! Even if you don't, it might be quite a lovely ladder.

(2.) *The discipline of writing clearly for a broad audience.*

Specialists love jargon. And for good reason: When you work a lot with a concept or a tool, you want to have a specific name for it, and that specific name might not be part of the common language of non-specialists. So, stock analysts have their trailing P/E ratios and their EBITDAs, jazz musicians have their 9/8 time signatures and their A♭dim7s, and philosophers have their neutral monisms and their supererogation.

Specialists also love arguing about subsidiary issues four layers deep in a conversation no one else will understand. And for good reason: Those things matter, and when something works or doesn't work, especially if its success or failure is surprising, it's often because of some recondite detail: the previous quarter report reflects a huge legal liability that everyone knew was coming and makes the earnings look jumpy, the guitar sound will be fuller if the song is transposed into E which will open more strings. The last two journal articles I refereed concerned, in part, (a.) whether the "rubber hand" illusion is a good objection to Rory Madden's argument that there can be no scattered subjects of consciousness and (b.) whether the Tamar Gendler's concept of "alief" can successfully save, from a certain type of objection concerning absentmindedness, the view that the disposition to phenomenally experience the judgment that P is sufficient for believing that P. (Don't ask.[196]) These are important issues. Pulling on these threads threatens to unravel some big stuff, and I hope both articles are published.

And yet the specialist can spend too much time neck-deep in these issues, conversing with others who are also neck deep. Consequently, you can forget what is at stake – what makes the issues interesting to discuss in the first place. You can lose yourself in picayune details and forget the big picture that those detailed arguments are supposed to be illuminating. Furthermore, you can lose sight of the presuppositions, idealizations, or philosophical background beliefs that are implicit in your shared specialists' jargon.

I have found it to be excellent philosophical discipline to regularly articulate what I care about for educated but non-specialist readers. It forces me to convey what's interesting or important in what I'm doing, and it forces me to examine my jargon. If a thousand words isn't enough space to clearly say what I'm doing, why it might be interesting, and what my core argument is, then I'm probably un-compassed in some philosophical forest, turning circles in the underbrush.

(3.) *Feedback and revisability.*

If you're a big believer in philosophical expertise, then you might not put much stock in feedback from non-experts. No mathematician would value my feedback on a proof, and rightly so; I don't have the expertise to comment competently. You could think the same thing is true in philosophy. Philosophical combatants deep in the trenches of arguments about neutral monism and supererogation might reasonably believe that non-experts could have little of value to contribute to their exchange.

But I'm mistrustful of philosophical expertise, as you'll have gathered. Philosophers have *some* expertise, sure, and a lot more familiarity with some of the standard moves in their professional terrain – but the difference between Kant and our philosopher of hair (Chapter 50) is, I believe, not nearly as great as sometimes advertised. If Kant had approached the women and servants around him as philosophical equals, expressing his ideas in plain language, they might have given him some good advice (Chapter 52). In presenting ideas for

a broad audience, one receives feedback from a broad range of people. Those people can often see your specialist's dicey presuppositions, which really ought to be challenged, and they can see connections to issues beyond your usual purview.

And then – even better – in a blog, you can *continue* the discussion, gain a better understanding of where they're coming from and how they react to your responses. You can reply in the comments section, maybe even revise the post or write an addendum; or you can try again in a different way next year when your mind circles back around to writing another post on that topic.

The blog post is therefore the ideal medium for philosophy! Several hundred to a thousand words: just the right length for a human-sized philosophical thought with some detail, not built too tall. A public medium, encouraging clarity and a sense of what's important, and potentially drawing feedback from a wide range of thoughtful people, both expert and non-expert, in light of which you can revise and develop your view. It's how philosophy ought to be done.

Okay, I shouldn't be so imperialistic about it. Journal articles and books are good too, in their own way, for their own different purposes, since the detailed argumentative moves do often matter. And short stories, for their vividness and emotional complexity. And TV shows. And interviews, and dialogues, and….

58. Will Future Generations Find Us Morally Loathsome?

Ethical norms change. Although reading Confucius doesn't feel like encountering some bizarrely alien moral system, some ethical ideas do shift dramatically over time and between cultures. Genocide and civilian-slaughtering aggressive warfare are now considered among the evilest things people can do, yet they appear to be celebrated in the Bible and we still name children after Alexander "the Great".[197] Many seemingly careful thinkers, including notoriously Aristotle and Locke, wrote justifications of slavery.[198] Much of the world now recognizes equal rights for women, homosexuals, low-status workers, people with disabilities, and ethnic minorities.[199]

It's unlikely that we've reached the end of moral change. In a few centuries, people might view our current norms with the same mix of appreciation and condemnation that we now view norms common in ancient China and Early Modern Europe. Indeed, future generations might find our generation to be especially vividly loathsome, since we are the first generation creating an extensive video record of our day-to-day activities.

Let me highlight the point about vividness. I find it helpful, and intimidating, to imagine the microscope upon us. It's one thing to know, in the abstract, that Rousseau fathered five children with a lover he regarded as too dull-witted to be worth attempting to formally educate, and that he demanded against her protests that their children be sent to (probably high mortality) orphanages.[200] It would be quite another if we had baby pictures and video of Rousseau's interactions with Thérèse. It's one thing to know, in the abstract, that Aristotle had a wife and a life of privilege. It would be quite another to watch video of him proudly enacting elitist and patriarchal values we now find vile, while pontificating on the ethics of the man of wisdom. Future generations might detest our consumerism, or our casual destruction of the environment, or our neglect of the sick and elderly, and they might be horrified to view these practices in vivid detail – perhaps especially when enacted by

professional ethicists. By "our" practices and values, I mean the typical practices and values of readers living in early twenty-first century democracies – the notional readership of this book.

Maybe climate change proves to be catastrophic: Crops fail, low-lying cities are flooded, a billion desperate people are displaced or malnourished or tossed into war. Looking back on video of a philosopher of our era proudly stepping out of his shiny, privately-owned minivan, across his beautiful irrigated lawn in the summer heat, into his large chilly air-conditioned house, maybe wearing a leather hat, maybe sharing McDonald's ice-cream cones with his kids – looking back, that is, on what I (of course this is me) think of as a lovely family moment – might this seem to some future Bangladeshi philosopher as vividly disgusting as I suspect I would find, if there were video record of it, Heidegger's treatment of his Jewish colleagues?

#

If we are currently at the moral pinnacle, any change will be a change for the worse. Future generations might condemn our mixing of the races, for example. They might wince to see video of interracial couples walking together in public with their mixed-race children. Or they might loathe clothing customs they view as obscene. However, I feel comfortable saying that they'd be wrong to condemn us, if those were the reasons why. Only by an unusual exertion of imagination can I muster any real doubt about the moral permissibility of our culture's interracial marriage and tank tops.

But it seems unlikely that our culture is at the moral pinnacle; and thus it seems likely that future generations will have some excellent moral reasons to condemn us. More likely than our being at the moral pinnacle, it seems to me, is that either (a.) there has been a slow

trajectory toward better values over the centuries (as recently argued by Steven Pinker[201]) and that the trajectory will continue, or alternatively that (b.) shifts in value are more or less a random walk up, down, and sideways, in which case it would be unlikely chance if we happened to be at peak right now. I am assuming here the same kind of non-relativism that most people assume in condemning Nazism and in thinking that it constitutes genuine moral progress to recognize the equal moral status of women and men.

It doesn't *feel* like we're wrong, I assume, for those of us who share the values we currently find ordinary. It probably feels as though we are applying our excellent minds excellently to the matter, with wisdom and good sense. But might we sometimes be using philosophy to justify the twenty-first-century college-educated North American's moral equivalent of keeping slaves and oppressing women? Is there some way to gain some insight into this possibility – some way to get a temperature reading, so to speak, on our unrecognized evil?

Here's one thing I don't think will work: relying on the ethical reasoning of the highest status philosophers in our society. If you've read my chapters on Kant, Nazi philosophers, and the morality of ethics professors, you'll know why I say this.

#

I'd suggest, or at least I'd hope, that if future generations rightly condemn us, it won't be for something we'd find incomprehensible. It won't be because we sometimes chose blue shirts over red ones or because we like to smile at children. It will be for things that we already have an inkling might be wrong, and which some people do already condemn as wrong. As Michele Moody-Adams emphasizes in her discussion of slavery and cultural relativism, in every slave culture there were always *some* voices condemning the injustice of

slavery – among them, typically, the slaves themselves.[202] As a clue to our own evil, we might look to minority moral opinions in our own culture.

I tend to think that the behavior of my social group is more or less fine, or at worst forgivably mediocre (Chapters 4 and 8), and if someone advances a minority ethical view I disagree with, I'm philosopher enough to concoct some superficially plausible defenses. But I worry that a properly situated observer might recognize those defenses to be no better than Hans Heyse's defense of Nazism (Chapter 53) or Kant's justification for denying servants the vote (Chapter 52).

I find myself, as I write this final chapter, rereading the epilogue of Moody-Adams' 1997 book, *Fieldwork in Familiar Places*. Moody-Adams suggests that we can begin to rise beyond our cultural and historical moral boundaries through moral reflection of the right sort: moral reflection that involves… well, I'm going to bullet-point the list to slow down the presentation of it, since the list is so good:

- self-scrutiny,
- vivid imagination,
- a wide-ranging contact with other disciplines and traditions,
- a recognition of minority voices, and
- serious engagement with the concrete details of everyday moral inquiry.

This list, I should clarify, is what I extract from Moody-Adams' remarks, which are not presented in exactly these words or in a list format.

Instead of a narrow or papier-mâché seminar-room rationalism, we should treasure insight from the entire range of lived experience and from perspectives as different as possible from one's own, in a spirit of open-mindedness and self-doubt. Here lies our best chance of repairing our probable moral myopia.

If I have an agenda in this book, it's less to defend any specific philosophical thesis than to philosophize in a manner that manifests these virtues.

#

There's one thing missing from Moody-Adams' lovely list, though, or maybe it's a cluster of related things. It's wonder, fun, and a sense of the incomprehensible bizarreness of the world. We should have those in our vision of good philosophy too! Moral open-mindedness is not, I think, entirely distinct from epistemic and metaphysical open-mindedness. They mix (I hope) in this book. I think I see them mixing, too, in two of my favorite philosophers, the great humane skeptics Zhuangzi and Montaigne.

Uncomfortably self-critical reflections on jerkitude – they're apt to wear us down, and too much thinking of that sort might reinforce the exact type of moralizing jerkitude we hope to avoid. When we need a break from that, and some fun, and to cast ourselves into a very different sort of doubt, we could spend some time – you and me together if you like – dreaming of zombie robots.

Acknowledgements

The following chapters appeared first in venues other than my blog, and have been somewhat revised for inclusion in this book: “A Theory of Jerks”, “Cheeseburger Ethics”, and “We Have Greater Obligations to Robots” in *Aeon Magazine*; “Should Your Driverless Car Kill You”, “Dreidel: A Seemingly Foolish Game”, “Does It Matter if the Passover Story Is Literally True?”, “What Happens to Democracy”, and part of “Blogging and Philosophical Cognition” in the *Los Angeles Times*; “Cute AI and the ASIMO Problem” in Schwitzgebel and Garza 2015; and “Why Metaphysics Is Always Bizarre” in Schwitzgebel 2014b. “Is the United States Literally Conscious” is an abbreviated version of Schwitzgebel 2015a.

So many people have given me helpful comments and suggestions on the topics of these posts, in person and on social media and by email or instant messages and during presentations and interviews and telephone conversations and video calls (and even a couple of helpful old-fashioned pieces of mail about my op-ed on Passover traditions) that it would take a superhuman memory to properly thank everyone who deserves it. So I’ll just mention some names of people who saliently came to mind as I was revising and updating, for having said something memorably helpful on one or more of the chapters above. My embarrassed apologies – especially embarrassed in light of what I say about forgetting in Chapter 2! – to the many I unfairly omit. Thank you, Peter Adamson, Gene Anderson, Nick Alonso, Roman Altshuler, Nomy Arpaly, Yuval Avnur, Uziel Awret, John Baez, Nick Baiamonte, Dave Baker, Scott Bakker, Zach Barnett, Jon Baron, John Basl, Howard Berman, Daryl Bird, Izzy Black, Reid Blackman, Ned Block, Daniel Bonevac, Nick Bostrom, Kurt Boughan, “Brandon” (on Kant), Richard Brown, Tim Brown, Wesley Buckwalter, Nick Byrd, Joe Campbell, Sean Carroll, David Chalmers, Myisha Cherry, chinaphil, Michel Clasquin-Johnson, Brad Cokelet, Mich Curria, Helen De Cruz, Leif Czerny, Rolf Degen, Dan Dennett,

Keith DeRose, Cory Doctorow, Fred Dretske, David Duffy, Kenny Easwaran, Aryeh Englander, Daniel Estrada, Sascha Fink, John Fischer, Owen Flanagan, Julia Galef, Kirk Gable, Mara Garza, Dan George, David Glidden, Peter Godfrey-Smith, Sergio Graziosi, Jon Haidt, Rotem Hermann, Joshua Hollowell, Russ Hurlburt, Anne Jacobson, Aaron James, Eric Kaplan, Jonathan Kaplan, Kim Kempton, Jeanette McMullin King, Peter Kirwan, Roxana Kreimer, Ed Lake, Juliet Lapidos, Chris Laursen, Amod Lele, Neil Levy, P.D. Magnus, Angra Mainyu, Pete Mandik, Josh May, Randy Mayes, Joshua Miller, Ethan Mills, Kian Mintz-Woo, Alan Moore, Daniel Nagase, Eddy Nahmias, Adam Pautz, Bryony Pierce, Steve Petersen, Benjamin Philip, Peter Railton, Paul Raymont, Kris Rhodes, Sam Rickless, Regina Rini, Rebecca Roache, Josh Rust, Colleen Ryan, Sandy Ryan, Callan S., Susan Schneider, David Schwitzgebel, Kate Schwitzgebel, Lisa Shapiro, Henry Shevlin, Nichi Smith, Justin Smith, David Sobel, Dan Sperber, Eric Steinhart, Charlie Stross, Olufemi Taiwo, Yvonne Tam, June Tangney, Valerie Tiberius, Justin Tiwald, Clinton Tolley, Shelley Tremain, David Udell, Bryan Van Norden, Dan Weijers, Nathan Westbrook, Eric Winsberg, Charles Wolverton, Mark Wrathall, Aaron Zimmerman, and of course the cognitively diverse group mind Anonymous.

For comments on the whole draft, I am especially grateful to Linus Huang, P.J. Ivanhoe, Jeremy Pober, Cati Porter, and most of all my wife Pauline Price, ever patient, tolerant, critical, and forgiving. This book is the joint project of our co-authored lives.

Notes

[1] Yankovic 2006, "I'll Sue Ya"; Temple 1959 as cited in Schwitzgebel 2014a.

[2] *Etymology Online* https://www.etymonline.com/word/jerk; and *Oxford English Dictionary* online http://www.oed.com/ [both accessed Jun. 21, 2018]; Barnhart 1988.

[3] Nietzsche 1887/1998, I.4-5, p. 12-14.

[4] Frankfurt 1986/2005; James 2012.

[5] James 2012, p. 5.

[6] Lilienfeld and Andrews 1996; Paulhus and Williams 2002; Blair, Mitchell, and Blair 2005.

[7] On Machivavellianism and narcissism, see Christie and Geis 1970; Fehr, Samsom, and Paulhus 1992; Miller and Campbell 2008; Jones and Paulhus 2014.

[8] Vazire 2010; also John and Robins 1993; Gosling, John, Craik, and Robins 1998.

[9] Arendt 1963, p. 114. For a very different portrayal of Eichmann on which he might have only strategically feigned ignorance, see Stangneth 2014.

[10] I had heard that Michelangelo said this, but no such luck: O'Toole 2014.

[11] GiveWell, for example, estimates that effective malaria programs cost about $2000 per life saved. URL: https://www.givewell.org/giving101/Your-dollar-goes-further-overseas [accessed Jun. 21, 2018]. See also discussion in Singer 2009, ch. 6.

[12] Doctors report smoking at rates substantially lower than do members of other professions. However, the data on nurses are mixed and the self-reports of doctors are probably compromised to some extent by embarrassment (Squier et al. 2006; Jiang et al. 2007; Sezer, Guler, and Sezer 2007; Smith and Leggat 2007; Frank and Segura 2009; Sarna, Bialous, Sinha, Yang, and Wewers 2010; Saikh, Sikora, Siahpush, and Singh 2015; though see Abdullah et al. 2013). Studies of doctors' general health practices are mixed but

confounded by issues of convenience, embarrassment, high professional demands, and the temptation to self-diagnose and self-treat (Richards 1999; Kay, Mitchell, and Del Mar 2004; Frank and Segura 2009; George, Hanson, and Jackson 2014).

[13] Sluga 1993; Young 1997; Faye 2005/2009; Gordon 2014.

[14] With the exception of the Nazi study, which was never formally written except as a blog post (included here as Chapter 53), these studies are summarized Schwitzgebel and Rust 2016. Philipp Schoenegger and collaborators have recently replicated several of our findings among German-language philosophers (personal communication), although the results on vegetarianism appear to be rather different, perhaps due to US/German cultural differences and/or changes in attitude in the ten years between 2008 and 2018.

[15] Schwitzgebel and Rust 2014, section 7.

[16] Schwitzgebel and Rust 2014, section 10.

[17] Singer 1975/2009.

[18] Cialdini, Reno, and Kallgren 1990; Cialdini, Demaine, Sagarin, Barrett, Rhoads, and Winter 2006; Schultz, Nolan, Cialdini, and Griskevicius 2007; Goldstein, Cialdini, and Griskevicius 2008; Bicchieri and Xiao 2009; Bicchieri 2017. The most widely replicated of these findings is that home energy conservation practices shift depending on what people learn about their neighbors' conservation practices: Allcott 2011; Ayres, Raseman, and Shih 2013; Karlin, Zinger, and Ford 2015.

[19] Mazar and Zhong 2010; Brown et al. 2011; Jordan, Mullen, and Murninghand 2011; Conway and Peetz 2012; Clot, Grolleau, and Ibanez 2013; Cornelissen, Bashshur, Rode, and le Menestrel 2013; Susewind and Hoelzl 2014; Blanken, van de Ven, and Zeelenberg 2015; Mullen and Monin 2016.

[20] This is why I find it so interesting, for example, to examine Kant's *Metaphysics of Morals* alongside his more abstract ethical works. See Chapter 52.

[21] See Haidt 2012 for an interpretation of Josh's and my work along roughly these lines, and see Rust and Schwitzgebel 2015 for discussion of that and several other alternative interpretations (though we do not discuss the moral mediocrity interpretation).

[22] In many theoretical discussions (e.g., Schroeder 2004; Moore 2004/2013; Berridge and Kringelbach 2013), pleasure and displeasure are treated as having similar motivational weight, but I guess this doesn't seem phenomenologically correct to me. (This issue is distinct from the question of whether pleasure and pain are distinct dimensions of experience that are capable of being experienced simultaneously.)

[23] Nuanced readings of Stoicism and Buddhism normally acknowledge that the aim is not the removal of all hedonically-valenced states, but rather the cultivation of a certain type of calm positive state of joy or tranquility (e.g., Epictetus 1st c. CE/1944; Hanh 1998/2015). Smart 1958 critiques a version of "negative utilitarianism", drawn partly from Popper 1945/1994 (p. 548-549), on which relieving suffering has more moral value that increasing pleasure by a similar amount. However, negative utilitarianism is rarely endorsed. Griffin 1979 suggests similarly that a small amount of unhappiness may be much more undesirable than a fairly large amount of happiness is desirable.

[24] The classic treatments of "loss aversion" are Tversky and Kahneman 1991 and 1992. The huge subsequent literature in psychology and experimental economics is, I understand, broadly confirmatory, though I have not yet managed to read quite all of it.

[25] At least non-derivatively. Bentham 1781/1988. Such psychological or motivational hedonism is now rarely accepted. Influential critiques include Hume's discussion of "self-love" (1751/1975, Appendix II), Nozick's (1974) "experience machine" thought experiment, and Sober and Wilson's (1998) discussion of the evolutionary bases of altruism, and Batson's long series of psychological experiments on altruistic motivation

(summarized in Batson 2016). For a recent defense of motivational hedonism, see Garson 2016.

[26] For some (non-decisive) evidence of this, see Kahneman, Krueger, Schkade, Schwarz, and Stone 2004; Margolis, Rachel, and Mikko Myrskylä 2011; Hansen 2012; for a critical perspective see Herbst and Ifcher 2016.

[27] See Parfit's (1984) classic distinction between hedonic, desire-fulfillment, and objective list theories of well-being. For a review of the literature on well-being, see Crisp 2001/2017; for a recent defense of prudiential hedonism, see Feldman 2004. Three prominent objections against prudential hedonism are the "happy swine" objection (that life is not better for an enormously happy pig than for a person with a mix of ups and downs; see discussion in Bramble 2016); Nozick's 1974 "experience machine" objection (that it would be worse to be unwittingly trapped in an experience machine generating all sorts of fake experiences and consequent real pleasures, than to live a real life); and the idea that betrayal behind one's back harms you even if you never learn about it or have any bad experiences as a result (Nagel 1979; Fischer 1997).

[28] This is widely accepted in the dream literature, matches the personal experiences of the people I've discussed it with, and fits with a common model of dream recall on which shortly after waking the attempt to recall one's dreams consolidates them into long term memory. However, as noted in Aspy, Delfabbro, and Proeve 2015, direct and rigorous empirical evidence in favor of this conclusion is thin.

[29] The classic treatment of lucid dreaming is LaBerge and Rheingold 1990.

[30] I remember chatting with someone about these matters at a conference meeting a few weeks before the original post in 2012. In the fog of memory, I couldn't recall exactly who it was or to what extent these thoughts originated from me as opposed to my interlocutor. Apologies, then, if they're due.

[31] Mengzi 4th c. BCE/2008, 7A15.

[32] Mengzi 4th c. BCE/2008, 1A7.

[33] I owe this point to Yvonne Tam.

[34] I discuss this issue at length in Schwitzgebel 2007.

[35] Rousseau 1762/1979, p 235.

[36] Confucius 5th c. BCE/2003, 15.24. Due to its negative phasing "do not…", this is sometimes called the Silver Rule.

[37] Mozi 5th c. BCE/2013; Xunzi 3rd c. BCE/2014. However, Mozi does have an argument for impartial concern that starts by assuming that one is concerned for one's parents (Chapter 16)

[38] Schwitzgebel 2018a.

[39] Susan Wolf famously argues that sainthood "does not constitute a model of personal well-being toward which it would be particularly rational or good or desirable for a human being to strive" (1982, p. 419). What I mean to be discussing here is not sainthood in Wolf's strong sense but only moral excellence of the more pedestrian sort – the excellence, perhaps, of the few most overall morally excellent people you personally know.

[40] In the philosophical lingo, your actual and hypothetical choices reveal what you are aiming for *de re* (i.e., what state of affairs in the world you are really guiding your actions toward), which might be different from what you are aiming for *de dicto* (i.e., what sentence you would endorse to describe what you are aiming for). Or something like that. On the de re / de dicto distinction in general see McKay and Nelson 2010/2014. For a discussion of its application to moral cases see Arpaly 2003.

[41] See, for example, the essays in Bloomfield, ed., 2008.

[42] For a review, see Langlois, Kalakanis, Rubenstein, Larson, Hallam, and Smoot 2000.

[43] See references in [XXX licensing footnote in Cheeseburger Ethics chapter]

[44] See, for example, Erin Faith Wilson's list of seventeen anti-gay activists and preachers whose homosexual affairs were exposed from 2004-2017 (Wilson 2017).

[45] Mikkelson and Evon 2007/2017. Nota bene: If Gore were saying, "We need top down regulation but until that comes, individuals should feel free to consume as luxuriously as they wish" then his luxurious consumption would not be evidence of hypocrisy.

[46] Optimistic self-illusions: Taylor and Brown 1988; Shepperd, Klein, Waters, and Weinstein 2013. End-of-history thinking: Quoidbach, Gilbert, and Wilson 2013, though see Ellenberg 2013 for a critique.

[47] Batson 2016, p. 25-26.

[48] LeCun, Bengio, and Hinton 2015; Silver et al. 2016; for comparison and contrast with human cognitive architecture see Lake, Ullman, Tenenbaum, and Gershman 2017.

[49] Wheatley and Haidt 2005; Schnall, Haidt, Clore, and Jordan 2008; Eskine, Kacinik, and Prinz 2011; for a review and meta-analysis, see Landy and Goodwin 2015.

[50] Baron 1997.

[51] Haidt 2012.

[52] Emailed McMeel re permission on July 13, 2018, but no reply.

[53] In ethics, see McDowell 1985; Railton 1986; Brink 1989; Casebeer 2003; and Flanagan, Sarkissian, and Wong 2008.

[54] Kruger and Dunning 1999. (Figure adapted from page 1129.)

[55][55] I haven't noticed a systematic discussion of cases where Dunning-Kruger doesn't apply, though Kahneman and Klein 2009 is related.

[56] Taylor and Brown 1988; Sedikes and Gregg 2008; Shepperd, Klein, Waters, and Weinstein 2013.

[57] Regenerate figures and make the curve in the middle more visible than the dots.

[58] On self-enhancement, see references in note XXX. On low correlation, see references in note XXX.

[59] In Schwitzgebel 2007, I argue that Xunzi's and Hobbes's models of moral education fit this pattern.

[60] In Schwitzgebel 2007, I argue that Mengzi's and Rousseau's models of moral education fit this pattern. See also Ivanhoe 1990/2002 on the metaphor of cultivation in the Confucian tradition. I also read most of the foundational figures in 20th century moral psychology as endorsing this type of approach, including Piaget 1932/1975, Kohlberg 1981, and Damon 1988. See also Baumrind 1971 on the "authoritative" parenting style.

[61] Here, and in most of my examples, names are chosen randomly from names of former students in my lower division classes, excluding Jesus, Mohammed, and very unusual names. For unnamed characters, the gender is the opposite of the named characters, to improve pronoun clarity.

[62] For example, Damon 1988; de Waal 1996; Haidt 2012; Bloom 2013.

[63] Mengzi (4th c. BCE/2008) is my favorite advocate of this type of view and emphasizes the seeds and cultivation metaphor, as emphasized and clarified in Ivanhoe 1990/2002 and Schwitzgebel 2007.

[64] Respondents in one survey (Bonnefon, Shariff, and Rahwon 2016) reported a median 50% likelihood of purchasing an autonomous vehicle programmed to save its passengers even at the cost of killing ten times as many pedestrians, compared to a median 19% likelihood of purchasing one programmed to minimize the total number of deaths even if it meant killing you and a family member.

[65] I originally published the reflections above in the *Los Angeles Times* in 2015, while Google was actively making the case for "self-certification" in California. Google and Tesla were threatening to move testing out of California to more lenient states if the Department of

Motor Vehicles didn't relax its attitude. It appears that they subsequently convinced the California DMV to permit self-certification rather than make the algorithms and standards public or have them at least evaluated by an independent regulator (McFarland 2017). However, it's not yet too late for nation-level regulators to step in, to reduce the race-to-laxity competition among the states.

[66] Clark 2008.

[67] Weizenbaum 1976.

[68] Darling 2017; Vedantam 2017.

[69] Johnson 2003; Meltzoff, Brooks, Shon, and Rao 2010; Fiala, Arico, and Nichols 2012.

[70] But see references in note XXX for an alternative view.

[71] Snodgrass and Scheerer 1989.

[72] Bryson 2010, 2013.

[73] For further discussion, see Schwitzgebel and Garza 2015. Jeremy Pober has argued in personal communication that artificial selection in dog breeding might partly violate the Emotional Alignment Design Policy by creating dogs that solicit emotional reactions disproportionate to their real moral status – in contrast to, say, coyotes. On the other hand, according to the reasoning of Chapter 19 (as also observed by Jeremy), we might owe dogs special moral consideration, more than we owe to coyotes, due to the role we have played in shaping them.

[74] This chapter was inspired by a conversation with Cory Doctorow about how a kid's high-tech rented eyes might be turned toward favored products in the cereal aisle.

[75] Especially Asimov 1950, 1976. See Petersen 2012 for a philosophical essay defending this solution.

[76] Adams 1980/2002, p. 224.

[77] Adams 1980/2002, p. 225.

[78] As usual, the *Stanford Encyclopedia of Philosophy* is an excellent starting place for general reviews of these topics: Sinnott-Armstrong 2003/2015; Hursthouse and Pettigrove 2003/2017; Johnson and Cureton 2004/2016.

[79] However, in Schwitzgebel and Garza (forthcoming), we argue that the deontological concern here is closest to the root issue.

[80] Asimov 1976; Snodgress and Scheerer 1989; Sparrow 2004; Basl 2013; Bostrom and Yudkowsky 2014.

[81] Shelley 1818/1965, p. 95.

[82] Searle 1980, 1992.

[83] See Schwitzgebel 2014b for my defense of medium-term dubiety of any general theory of consciousness that applies broadly across possible types of natural and artificial beings. I assume that consciousness is required for human-level moral considerability. Kate Darling (2016), Daniel Estrada (2017), and Greg Antill (in a non-circulating draft article) have argued (each on different grounds) that AI need not even be potentially conscious to deserve at least some moral consideration.

[84] For more extended discussion of these issues see Schwitzgebel and Garza 2015, forthcoming.

[85] Nozick 1974.

[86] Sparrow 2004; in short story format: Schwitzgebel and Bakker 2013.

[87] By random chance, it is a heterosexual marriage. See note XXX on my policy for choosing names in philosophical examples.

[88] For more detailed philosophical treatments of simulated and virtual worlds see Bostrom 2003; Chalmers 2003/2010, forthcoming; Steinhart 2014; Schwitzgebel

forthcoming. For a science-fictional portrayal of such worlds, with philosophical consequences in plain view, see Egan 1994, 1997.

[89] Bostrom 2003.

[90] Chalmers 2003/2010. See also Steinhart 2014. The famous entrepreneur Elon Musk endorses the possibility in a 2016 interview here: https://www.youtube.com/watch?v=2KK_kzrJPS8 [accessed Jul. 3, 2018].

[91] Chalmers does not specifically discuss dream scenarios, but I see no reason to think he would treat it differently as long as it met fairly stringent conditions of stability and shared collective experience.

[92] Bostrom 2011.

[93] For a more detailed defense of sim-based skepticism and cosmological skepticism, see Schwitzgebel 2017.

[94] See Schneider forthcoming. On philosophical "zombies" see Kirk 2003/2015. I use "zombie" here in a somewhat looser sense than is orthodox, to refer to an entity who is functionally and behaviorally similar to an ordinary person (bracketing the issue of biological vs. robotic or virtual embodiment) but who entirely lacks a stream of conscious experience or "phenomenology".

[95] Block 1978/2007, 2002/2007; Searle 1980, 1984.

[96] See Schneider forthcoming, ch. 4.

[97] Descartes 1641/1984. For an extended argument that introspection even of current conscious experience is highly untrustworthy, see Schwitzgebel 2011, esp. ch. 7.

[98] This would follow from the well-known Integrated Information Theory of Consciousness (Oizumi, Albantakis, and Tononi 2014)..

[99] For related arguments, but with the opposite conclusion, see Cuda 1985 and Chalmers 1996, ch. 7. For helpful discussion of Schneider's chip test, thanks to David Udell.

[100] See Heschel 2003 for details of the story.

[101] I transcribed this quote from an old CLU newspaper clipping, but I can no longer find the original source.

[102] For example Skinner 1948/1976; Leary 1988/2003.

[103] As famously portrayed in *One Flew Over the Cuckoo's Nest* (Kesey 1962).

[104] In the psychology literature, this is called "conversational shadowing". In his career as a psychologist, my father was involved in some early shadowing studies: Schwitzgebel and Taylor 1980.

[105] Here was I was influenced by Harry Frankfurt's lectures and discussions at UC Riverside, some of which became Frankfurt 2004.

[106] McGeer 1996. See also Zawidzki 2013.

[107] On cognitive dissonance: Festinger 1957; Cooper 2007. Relatedly, in "choice blindness" studies, people choose one response and through sleight of hand, the experimenter makes it appear as they had made the opposite response. Participants often don't notice the swap and even justify their "choice" with reasons indistinguishable from the types of reasons that are given for ordinary choices: Johansson, Hall, Sikström, Tärning, and Lind 2006; Hall, Johansson, and Strandberg 2012.

[108] Timmer, Westerhof, Dittmann-Kohli 2005. Ware 2011 is a popular discussion based the author's experiences in elder care, which better fits the standard picture.

[109] Adams 1980/2002, p. 163.

[110] [Need permission.]

[111] Roese and Vohs 2012.

[112] The final words of a death row inmate who was not given the dignity of a careful fulfillment of his requested last meal (Melton 2001-2009, p. 90).

[113] Sacks 1985; Bruner 1987; Dennett 1992; Fischer 2005; Velleman 2005.

[114] Herodotus 5th c. BCE/2017, book I, esp. I.30-32, p. 15-17 and I.86, p. 38-40.

[115] Pun intended. See Chapter 27.

[116] Roache 2015, 2016.

[117] Roache 2015, 31:20.

[118] Image from: https://www.someecards.com/users/profile/Serena2015282 [need permission].

[119] Carlin 1972. Evidently, there was no official list, but I can attest that *I* never heard these words on U.S. TV in the 1970s, and broadcasters used to be fined for violations. Carlin was arrested for "disturbing the peace" when he performed the act in the Milwaukee in 1972 (Dimeo 2008). To be clear, despite my preference for reducing the usage of "fuck" I would not support arresting anyone for using it.

[120] Google NGram, smoothing of 3, downloaded Jul. 8, 2015. URL: https://books.google.com/ngrams. [NOTE to publisher: This needn't be color, but I can't figure out how to change it to grayscale without making the lines indistinguishable.]

[121] https://trends.google.com/trends/?geo=US [accessed Jul. 4, 2018]. Usage has declined somewhat since peaking in 2015, but it's unclear whether that marks the beginning of a long-term trend. Similar increases, and 2015 peaks, show in data from the U.K., Canada, and Australia. Worldwide usage has almost doubled over the period.

[122] NOTE TO EDITOR: This isn't really a note to the publisher, and should be kept in the body of the text.

[123] NOTE TO EDITOR: Image can be converted to grayscale if necessary.

[124] Photos courtesy of California Lutheran University.

[125] Bradbury 1997-2000.

[126] See also Bostrom 2014; Steinhart 2014; Schneider forthcoming.

[127] This is the orthodox "functionalist" view of consciousness: Levin 2004/2018.

[128] This is the "simulation" scenario, which I explore in more detail in Chapters 21 and 22.

[129] Gazzaniga 2005.

[130] Ever since Chalmers 1996.

[131] Since the whole affair is secret, there's no point in searching Koch's website for information on it. Sorry! In fact, I am only revealing this episode now because I trust that you will misinterpret it as fiction.

[132] Example adapted from Block 1978/2007.

[133] For example Putnam 1975; Burge 1979; Millikan 1984; Drestske 1988, 1995.

[134] See Moravec 1999; Kurzweil 2005; Hilbert and López 2011. It is probably too simplistic to conceptualize the connectivity of the brain as though all that mattered were neuron-to-neuron connections; but those who favor complex models of the internal interactivity of the brain should, I think, for similar reasons, be drawn to appreciate complex models of the interactivity of citizens and residents of the United States.

[135] As in Dennett 1991; Metzinger 2003.

[136] Few theorists have attempted to explain why they think that the United States isn't literally conscious. Tononi is one; see my responses in Schwitzgebel 2014d and Schwitzgebel 2015a (from which the present chapter is adapted). François Kammerer is another; see our exchange in Kammerer 2015; Schwitzgebel 2016. In Schwitzgebel 2015a, I also reply to objections via email from David Chalmers, Daniel Dennett, and Fred Dretske.

[137] Boltzmann 1897; Gott 2008; Carroll 2010, 2017; De Simone, Guth, Linde, Noorbala, Salem, and Vilenkin 2010; Crawford 2013; Boddy, Carroll, and Pollack 2014/2016.

[138] On "externalist" views of the mind, specific thought contents or even the presence of consciousness itself can require certain things to be factually true about one's history or

environment, independently of any difference in one's current locally-described brain structure. For this chapter, I am assuming the falsity of externalism. Classic externalist arguments include Putnam 1975; Burge 1979; Millikan 1984; Dretske 1995. For a review, see Lau and Deutsch 2002/2016. For modest externalisms that require only a moderately large chunk of environment, we can modify the case to include random fluctuations of at least that size. Externalisms that require a substantial evolutionary history, and thus stability over millions of years, might be so much more improbable, by chance, as to require a fresh assessment of the relative probabilities, for example, if the long-estabilished stability makes it likely that, by chance, a stable system was spawned. Even if we grant a version of externalism that disallows genuine thoughts about cosmology in freak systems, we might only have kicked the epistemic problem up a level, turning it into a problem about how we know that we are actually having thoughts about cosmology (possible only with the right history) instead of sham-thoughts that only seem to have that content (as a freak might, on an externalist view of mental content; adapting McKinsey 1991's well-known challenge to externalism).

[139] Boddy, Carroll, and Pollack 2014/2016.

[140] De Simone, Guth, Linde, Noorbala, Salem, and Vilenkin 2010.

[141] The classic presentation is Putnam 1981.

[142] Descartes 1641/1984.

[143] See Schwitzgebel 2017 for a fuller version of this argument, as well as a parallel argument concerning the skeptical worry that you might currently be dreaming.

[144] Gott 2008; Tegmark 2014.

[145] Carroll 2017; also Davenport and Olum 2010, Crawford 2013.

[146] The example is originally from Davidson 1987. See also Dretske 1995; Neander 1996; Millikan 2010.

[147] The classic source of transporter puzzles about personal identity is Parfit 1984. For a recent discussion see Langford and Ramachandran 2013.

[148] I further develop this thought experiment in my short story "The Dauphin's Metaphysics" (Schwitzgebel 2015b).

[149] Greene 2011; Tegmark 2014.

[150] The classic statement of this principle is in Bondi 1952/1960.

[151] The name "butterfly effect" traces to Lorenz 1972 and was popularized in Gleick 1987. Wolfram 2002 traces mathematical proofs of the amplification of small effects back to James Clerk Maxwell in 1860 and Henri Poincaré in 1890.

[152] Please don't think about Evil Emily.

[153] Williams 1973. For the contrary case, see especially Fischer 1994; Fischer and Mitchell-Yellin 2014. The thought behind this post is that Fischer's case in terms of "repeatable" pleasures is even easier to make than he allows, since forgetting is inevitable and every pleasure is repeatable upon forgetting.

[154] Gaiman 1996/1998, p. 81.

[155] If you haven't yet read Borges's *Labyrinths* – a translation of some of his most philosophical mid-career stories – I urge you to put this book down and pick up that one instead.

[156] For a formal discussion of the physics and logic of recurrence, see Wallace 2015.

[157] For a fictional exploration of some related themes, see Schwitzgebel 2017b.

[158] Most of the material in this section is drawn from Chase 2002. For more details and fuller references see Schwitzgebel 2018c.

[159] Carruthers 1996, 2018; Dennett 1996; Strawson 2006; Oizumi, Albantakis, and Tononi 2014.

[160] Arpaly and Barnett 2017.

[161] This is an empirical claim about what is usually the case. Sometimes a Truth philosopher will strike upon a truth so weird and rarely recognized that others think that they can't really sincerely believe the position they advance. Such discoveries should, by their nature, be rare among philosophers with realistic self-assessments. Of course, one such truth, once found, can become the centerpiece of a whole career.

[162] Possible worlds really exist: Lewis 1986. All matter is conscious: Goff 2017. We're morally obliged to let humanity go extinct: Benatar 2006.

[163] USA consciousness: Chapter 39 and Schwitzgebel 2015a. Bad introspectors: Schwitzgebel 2011. Short-lived AIs: Chapter 22 and Schwitzgebel 2017. Freaks: Chapter 40 and Schwitzgebel 2013.

[164] In a comment on the blog post this chapter was drawn from, Randy Mayes suggests that the Hair Splitter may be an even more common type, and Nichi Smith suggests the Hole Poker. Clearly there remains much important taxonomic work to be done.

[165] I'm assuming you didn't just jump to Chapter 51 after reading the preface. If you did jump here immediately after reading the preface, you probably already trust your sense of fun, and so you don't need to read this chapter. But then, if you're like that, you probably aren't reading this endnote – unless you find endnotes fun, in which case you're some sort of *double* dork (just like me for writing this dorky footnote).

[166] External world: Schwitzgebel and Moore 2015. Ethics books: Schwitzgebel 2009. Dreaming in black and white: Schwitzgebel 2011. Jerkitude: Chapter 1. Self-ignorance: Schwitzgebel 2012. Bizarre aliens: Schwitzgebel 2015a. Babies: Schwitzgebel 1999.

[167] If you mention the code phrase "philosophy dork", I will befriend you on my own social media. Hi!

[168] Feynman 1985, p. 173ff.

[169] Karpel 2014. The rest of the Whedon quote also applies quite well to philosophy, I think: "Some people will disagree, but for me if I've written a meaty, delightful, wonderful bunch of scenes and now I have to do the hard, connective, dog's body work of writing, when I finish the dog's body work, I'll have a screenplay that I already love. I used to write chronologically when I started, from beginning to end. Eventually I went, That's absurd; my heart is in this one scene, therefore I must follow it. Obviously, if you know you have a bunch of stuff to do, I have to lay out this, all this dull stuff, and I feel very uncreative but the clock is ticking. Then you do that and you choose to do that. But I always believe in just have as much fun as you can so that when you're in the part that you hate there's a light at the end of the tunnel, that you're close to finished."

[170] For methodological details see Hurlburt 2011.

[171] At our most recent presentation of this material in Tucson in 2018, Russ and I noticed that, quite unusually, that in all three samples the audience member reported content having to do with the lecture. I'm unsure whether this is a fluke, or some difference in the audience (though we didn't notice this in our previous presentation in Tucson), or due to the fact that early in the presentation I had mentioned that audience members rarely reported attending to the content of the lectures (and I thereby may have tainted the procedure).

[172] Hurlburt 1979 and personal communication July 18, 2018, as well as my own experience in beeping studies in both the participant and the researcher roles.

[173] One time I was sitting on a raised platform behind a table with three other speakers. I was nervously wiggling my knee against the edge of the table, not noticing that the table was edging closer to the lip of the stage. Five minutes before the session's scheduled end, the table suddenly flipped off the stage, tossing notes and water pitcher and water glasses into the air, then landing face down at the feet of the people in the first row,

smashing the glasses. Though it certainly caught the audience's attention, on balance I don't recommend this approach as a regular practice.

[174] For example, Frege 1884/1953, 1918/1956 (though see Reck 2005); Lewis 1986; Yablo 1987.

[175] Besides the strangeness of realism about possible worlds, which I've already mentioned, Lewis's view of consciousness is bizarre in ways that I don't think have been fully appreciated: See Schwitzgebel 2015c and 2014, sec. 3.

[176] Critiques of the role of common sense or philosophical intuition as a guide to metaphysics and philosophy of mind can be found in, for example, Churchland 1981; Stich 1983; Kornblith 1998; Dennett 2005; Ladyman and Ross 2007; and Weinberg, Gonnerman, Buckner, and Alexander 2010. Hume 1740/1978 and Kant 1781/1787/1998 are also interesting on this issue. Even metaphilosophical views that treat metaphysics largely as a matter of building a rigorous structure out of our commonsense judgments often envision conflicts within common sense so that the entirety of common sense cannot be preserved: e.g., Ayer 1967; Kriegel 2011.

[177] Aristotle 4th c. BCE/1928, 983a; θαυμαστόν: wonderful in the sense of tending to cause wonder, or amazing.

[178] See Moore 1925 on common sense and Moore 1922, 1953, 1957 on sense data.

[179] Reid 1774-1778/1995; 1788/2010; though he says this mistake of the "vulgar" does them no harm: 1788/2010, IV.3.

[180] See also Sperber 2010 on gurus. A fuller treatment of the topic would also mention the positive reasons for obscurity (some of which are nicely listed by "boomer" in a comment on the original post from which this chapter was adapted. URL: https://schwitzsplinters.blogspot.com/2011/10/obfuscatory-philosophy-as-intellectual.html.

[181] For example, in a search of the PhilJobs database from June 1, 2015, to June 18, 2018, I count 30 advertisements mentioning "Kant*", three each for Plato* and Aquin*, two for Aristot*, and one for Confuc* [accessed June 19, 2018].

[182] For more charitable interpretations of these passages, a starting place might be the dozens of annoyed comments that my original post on this topic received on my blog (some with helpful references). I recommend checking it out for the other side of the story! URL https://schwitzsplinters.blogspot.com/2010/03/kant-on-killing-bastards-on.html.

[183] For discussion of Kant's racism see Mills 2005; Bernasconi 2011; Allais 2016. Allais argues that on the issue of race Kant is simply "not noticing obvious contradictions" in his thinking (p. 20), illustrating that even great philosophers are liable to ordinary human self-deception. For more of my reflections on philosophers' capacity for rationalization, see Schwitzgebel and Ellis 2017.

[184] Zhuangzi 4th c. BCE/2009; Montaigne 1580/1595/1957. Admittedly, Montaigne is a bit disappointing on gender.

[185] Kant 1785/1996.

[186] For a much more charitable reading of one aspect of Kant, see Schwitzgebel forthcoming.

[187] A few joined the SA or SS but not the Nazi party, but since involvement in one of these dedicated Nazi organizations reflects at least as much involvement in Nazism as does Nazi party membership alone, I have included them in the total.

[188] Jarausch and Arminger 1989.

[189] See Priest, Berto, and Weber 2008/2018.

[190] For example, on contradictoriness in Zhuangzi: Schwitzgebel 2018b; Montaigne: Miernowski 2016; Nietzsche: Müller-Lauter 1971/1999; Wittgenstein: Pichler 2007.

[191] I have done several invisible revisions of this chapter to make it more consistent with Chapters 51 and 54. I think you'll find it much better now.

[192] This is called the "Pygmalion effect" in educational psychology. The classic study is Rosenthal and Jacobson 1968/1992. For a recent review see Murdock-Perriera and Sedlacek 2018.

[193] Hegel 1807/1977; on Hegel's reaction to Napoleon: Pinkard 2000, p. 246, 311. One Hegel expert has commented that this might be the first time he had heard Hegel's arguments associated with "tidy". However, this expert did not dispute that it is a long, tenuous path from Hegel's A to his J.

[194] Spinoza 1677/1994, p. 180.

[195] Mozi 5th c. BCE.2013, ch. 16; Sextus circa 200 CE/1994, Book I.

[196] Or ask, I suppose, if you want: eschwitz@ucr.edu. Update: Both articles have been accepted: Chomanski forthcoming, Schiller forthcoming.

[197] One instance of brutal slaughter by Alexander and his forces was after the famous siege of Tyre (Green 1972/2013, ch. 7).

[198] Aristotle is clearest about this in Politics I: 4th c BCE/1995, 1254a-b, p. 6-7. Locke has been charitably interpreted as opposed to slavery by some scholars, and he has been interpreted as defending slavery by others (Glausser 1990; Armitage 2004). I believe it is clear that in his *Second Treatise of Government* (1689/2016) Locke defends the capture and holding of slaves as long as master and slave are "at war" with each other, and justly so if the slavery results from a just war, while acknowledging that slavery should end as soon as master and slave enter into a "pact" (which might possibly not occur during the slave's lifetime). Locke writes, for example, "*Slaves*, who being Captives taken in a just War, are by the Right of Nature subjected to the absolute Dominion and arbitrary Power of their Masters. These Men having, as I say, forfeited their Lives, and with it their Liberties, and lost their

Estates; and being in the *State of Slavery*, not capable of any Property, cannot in that State be considered as any part of *Civil Society;* the chief end whereof is the Preservation of Property" (1689/2016, VII.86, p. 43; though in Chapter 16 Locke confusingly seems to take back the part about forfeiting goods: XVI.182, p. 92). Locke also had part ownership of slave-trading enterprises and probably helped compose the 1669 Fundamental Constitutions of Carolina which stated that "Every freeman of Carolina shall have absolute power and authority over his negro slaves".

[199] Steven Pinker (2011) makes this point vividly in terms of what he calls the "Rights Revolutions". However, his depiction of traditional, "nonstate" societies might be inaccurate, excessively emphasizing their violence: Ferguson 2013a&b; Gómez, Verdú, González-Megías, and Méndez 2016.

[200] See Books VII and IX of Rousseau's *Confessions* (1769/1995), esp. VII, 331-333, p. 278-279 and VII, 343-345, p. 289. On high mortality rates in French orphanages or "foundling homes" during the period, see Colón and Colón 2001, p. 323-324, 503-504. Colón and Colón remark that "for overtly abandoned infants being nursed in foundling homes, death was a predictable outcome". Under the influence of new friends over a decade later, Rousseau made inquiries about the fate of his eldest child, seeking reunion. However, no record could be found.

[201] Pinker 2011.

[202] Moody-Adams 1997.

References

Aaronson, Scott (2014a). Why I am not an Integrated Information Theorist (or, the unconscious expander). Blog post at *Shtetl Optimized* (May 21). URL: https://www.scottaaronson.com/blog/?p=1799 [accessed Jul. 5, 2018].

Aaronson, Scott (2014b). Giulio Tononi and me: A Phi-nal exchange. Blog post at *Shtetl Optimized* (May 30). URL: https://www.scottaaronson.com/blog/?p=1823 [accessed Jul. 5, 2018].

Abdullah, Abu S., Feng Qiming, Vivian Pun, Frances A Stillman, and Jonathan M Samet (2013). A review of tobacco smoking and smoking cessation practices among physicians in China. *Tobacco Control, 22,* 9-14.

Adams, Douglas (1980/2002). *The Restaurant at the End of the Universe.* In *The ultimate Hitchhiker's Guide to the Galaxy.* New York: Random House.

Adamson, Peter (2016). All 20 "rules for the history of philosophy". At the *History of Philosophy Without Any Gaps* website (Dec. 31). URL: https://historyofphilosophy.net/rules-history-philosophy [accessed Jul. 5, 2018].

Allais, Lucy (2016). Kant's racism. *Philosophical papers, 45,* 1-36.

Allcott, Hunt (2011). Social norms and energy conservation. *Journal of Public Economics, 95,* 1082-1095.

Arendt, Hannah (1963). *Eichmann in Jerusalem.* New York: Penguin.

Aristotle (4th c. BCE/1928). *The works of Aristotle, vol VII: Metaphysica,* trans. W.D. Ross. Oxford: Oxford.

Aristotle (4th c. BCE/1995). *Politics, books I and II,* trans. T.J. Saunders. Oxford: Oxford.

Armitage, David (2004). John Locke, Carolina, and the *Two Treatises of Government*. *Political Theory, 32,* 602-627.

Arpaly, Nomy (2003). *Unprincipled virtue.* Oxford: Oxford.

Arpaly, Nomy, and Zach Barnett (2017). Philosophy: Truth or dare? Blog post at *The View from the Owl's Roost* (Sep. 28). URL: https://theviewfromtheowlsroost.com/2017/09/28/philosophy-truth-or-dare [accessed Jul. 5, 2018].

Asimov, Isaac (1950). *I, robot.* New York: Random House.

Asimov, Isaac (1976). *Bicentennial man.* Herts, UK: Victor Gollancz.

Aspy, Denholm J., Paul Delfabbro, and Michael Proeve (2015). Is dream recall underestimated by retrospective measures and enhanced by keeping a logbook? A review. *Consciousness & Cognition, 33,* 364-374.

Augustine (4th c. CE/1993). *On free choice of the will,* trans. T. Williams. Indianapolis: Hackett.

Ayer, A.J. (1967). *Metaphysics and common sense.* San Francisco: Freeman Cooper.

Ayres, Ian, Sophie Raseman, and Alice Shih (2013). Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage. *The Journal of Law, Economics, and Organization, 29,* 992-1022.

Barnhart, Robert K. (1988). *The Barnhart dictionary of etymology.* Bronx, NY: H.W. Wilson.

Baron, Robert A. (1997). The sweet smell of... helping: Effects of pleasant ambient fragrance on prosocial behavior in shopping malls. *Personality and Social Psychology Bulletin, 23,* 498-503.

Basl, John (2013). The ethics of creating artificial consciousness. *APA Newsletter on Philosophy and Computers, 13* (1), 23-29.

Batson, C. Daniel (2016). *What's wrong with morality?* Oxford: Oxford.

Baumrind, Diana (1971). Current patterns of parental authority. *Developmental Psychology, 4*, 1-103.

Bayle, Pierre (1697/1965). Paulicians. In Historical and Critical Dictionary: Selections, trans. R.H. Popkin. Indianapolis: Bobbs-Merrill.

Bazargan, Mohsen, Marian Makar, Shahrzad Bazargan-Hejazi, Chizobam Ani, and Kenneth E. Wolf (2009). Preventive, lifestyle, and personal health behaviors among physicians. *Academic Psychiatry, 33,* 289-295.

Benatar, David (2006). *Better never to have been.* Oxford: Oxford.

Bentham, Jeremy (1781/1988). *The principles of morals and legislation.* New York: Prometheus.

Bernasconi, Robert (2011). Kant's third thoughts on race. In S. Elden and E. Mendieta, *Reading Kant's Geography.* SUNY Press.

Berridge, Kent C., and Morten L. Kringelbach (2013). Neuroscience of affect: brain mechanisms of pleasure and displeasure. *Current Opinion in Neurobiology, 23,* 294-303.

Bicchieri, Cristina, and Erte Xiao (2009). Do the right thing: But only if others do so. *Journal of Behavioral Decision Making, 22,* 191-208.

Bicchieri, Cristina (2017). *Norms in the wild.* Oxford: Oxford.

Blair, James, Derek Robert Mitchell, and Karina Blair (2005). *The psychopath.* Malden, MA: Blackwell.

Blanken, Irene, Niels van de Ven, and Marcel Zeelenberg (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin, 41,* 540-558.

Block, Ned (1978/2007). Troubles with functionalism. In N. Block, *Consciousness, function, and representation*. Cambridge, MA: MIT.

Block, Ned (2002/2007). The harder problem of consciousness. In N. Block, *Consciousness, function, and representation*. Cambridge, MA: MIT.

Bloom, Paul (2013). *Just babies.* New York: Penguin.

Bloomfield, Paul, ed., (2008). *Morality and self-interest.* Oxford: Oxford.

Boddy, Kimberly K., Sean Carroll, and Jason Pollack (2014/2016). *De Sitter space without quantum fluctuations.* Manuscript at arXiv:1405.0298 [accessed Jul. 4, 2018].

Boltzmann, Ludwig (1887). Zu Hrn. Zermelo's Abhandlung "Ueber die mechanische Erklärung irreversibler Vorgänge". *Annalen der Physik, 296* (2), 392-398.

Bondi, Hermann (1952/1960). *Cosmology, 2nd ed.* Cambridge: Cambridge.

Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan (2016). The social dilemma of autonomous vehicles. *Science, 352,* 1573-1576.

Borges, Jorge Luis (1962/2007). *Labyrinths,* trans. D.A. Yates and J.E. Irby. New York: New Directions.

Bostrom, Nick (2003). Are we living in a computer simulation? *Philosophical Quarterly, 53,* 243-255.

Bostrom, Nick (2011). Personal communication publicly shared with permission as "Bostrom's response to my discussion of the simulation argument" at *The Splintered Mind* blog, Sep. 2, 2011. http://schwitzsplinters.blogspot.com/2011/09/bostroms-response-to-my-discussion-of.html.

Bostrom, Nick (2014). *Superintelligence.* Oxford: Oxford.

Bostrom, Nick, and Eliezer Yudkowsky (2014). The ethics of artificial intelligence. In K. Frankish and W.M. Ramsey, eds., *Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge.

Bradbury, Robert J. (1997-2000). *Matriosha Brains.* Unpublished manuscript. URL: https://www.gwern.net/docs/ai/1999-bradbury-matrioshkabrains.pdf [accessed Jul. 4, 2018].

Bramble, Ben (2016). A new defense of hedonism about well-being. *Ergo, 3* (4). DOI: http://dx.doi.org/10.3998/ergo.12405314.0003.004

Brink, David O. (1989). *Moral realism and the foundations of ethics.* Cambridge: Cambridge.

Brown, Ryan P, Michael Tamborski, Xiaoqian Wang, Collin D. Barnes, Michael D. Mumford, Shane Connelly, and Lynn D. Devenport (2011). Moral credentialing and the rationalization of misconduct. *Ethics & Behavior, 21,* 1-12.

Bruner, Jerome (1987). Life as narrative. *Social Research, 54,* 11–32.

Bryson, Joanna J. (2010). Robots should be slaves. In Y. Wilks, *Close engagements with artificial companions.* Amsterdam: John Benjamins.

Bryson, Joanna J. (2013). Patiency is not a virtue: Intelligent artifacts and the design of ethical systems. Online MS: https://www.cs.bath.ac.uk/~jjb/ftp/Bryson-MQ-J.pdf.

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

Calvin, John (1559/1960). *Institutes of the Christian religion, vol. 1,* ed. J.T. McNeill, trans. F.L. Battles. Louisville: Westminster John Knox.

Carlin, George (1972). *The seven words you can never say on television.* Comedy routine. Transcript available at https://genius.com/George-carlin-the-seven-words-you-can-never-say-on-television-annotated [accessed Jul. 4, 2018].

Carroll, Sean (2010). *From eternity to here.* New York: Penguin Random House.

Carroll, Sean (2017). *Why Boltzmann brains are bad.* Manuscript at arXiv:1702.00850 [accessed Jul. 4, 2018].

Carruthers, Peter (1996). *Language, thought and consciousness.* Cambridge: Cambridge.

Carruthers, Peter (2000). *Phenomenal consciousness.* Cambridge: Cambridge.

Carruthers, Peter (2018). *The problem of animal consciousness.* Online MS. URL: http://faculty.philosophy.umd.edu/pcarruthers/

Cartwright, Nancy (2015). Philosophy of social technology: Get on board. *Proceedings and addresses of the American Philosophical Association., 89*, 98-116.

Casebeer, William D. (2003). *Natural ethical facts.* Cambridge, MA: MIT.

Chalmers, David J. (1996). *The conscious mind.* Oxford: Oxford.

Chalmers, David J. (2003/2010). *The Matrix* as metaphysics. In *The character of consciousness.* Oxford: Oxford.

Chalmers, David J. (forthcoming). The virtual and the real. *Disputatio.*

Chase, Ronald (2002). *Behavior and its neural control in gastropod molluscs.* Oxford: Oxford.

Chignell, Andrew (2010/2018). The ethics of belief. *Stanford Encyclopedia of Philosophy* (Spring 2018 Edition). URL: https://plato.stanford.edu/archives/spr2018/entries/ethics-belief.

Chomanski, Bartek (forthcoming). Could there be scattered subjects of consciousness? *Phenomenology and the Cognitive Sciences.*

Christie, Richard, and Florence L. Geis (1970). Studies in Machiavellianism. New York: Academic.

Churchland, Paul M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy, 78,* 67-90.

Cialdini, Robert B., Raymond R. Reno, and Carl A. Kallgren (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58,* 1015-1026.

Cialdini, Robert B., Linda J. Demaine, Brad J. Sagarin, Daniel W. Barrett , Kelton Rhoads, and Patricia L. Winter (2006). Managing social norms for persuasive impact. *Social Influence, 1*, 3-15.

Clark, Andy (2008). *Supersizing the mind.* Oxford: Oxford.

Clot, Sophie, Gilles Grolleau, and Lisette Ibanez (2013). Self-licensing and financial rewards: Is morality for sale? *Economics Bulletin, 33,* 2298-2306.

Colón, A. R. and P. A. Colón (2001). *A history of children.* Westport, CT: Greenwood.

Confucius (5th c. BCE/2003). *The analects,* trans. E. Slingerland. Indianapolis: Hackett.

Conway, Paul and Johanna Peetz (2012). When does feeling moral actually make you a better person? Conceptual abstraction moderates whether past moral deeds motivate consistency or compensatory behavior. *Personality and Social Psychology Bulletin, 38,* 907-919.

Cooper, Joel (2007). *Cognitive dissonance.* Los Angeles: Sage.

Cornelissen, Gert, Michael R. Bashshur, Julian Rode, and Marc Le Menestrel1 (2013). Rules or consequences? The role of ethical mind-sets in moral dynamics. *Psychological Science, 24,* 482-488.

Crawford, Lyle (2013). Freak observers and the simulation argument. *Ratio, 26,* 250-264.

Crisp, Roger (2001/2017). Well-being. *Stanford Encyclopedia of Philosophy* (Fall 2017 Edition). URL: https://plato.stanford.edu/archives/fall2017/entries/well-being.

Cudd, Ann, and Seena Eftekhari, Seena (2000/2018). Contractarianism. *Stanford Encyclopedia of Philosophy* (Summer 2018 Edition). URL: https://plato.stanford.edu/archives/sum2018/entries/contractarianism.

Damon, William (1988). *The moral child.* New York: Simon & Schuster.

Darling, Kate (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior toward robotic objects. In R. Calo, A.M. Froomkin, and I. Kerr, eds., *Robot Law.* Glos, UK: Edward Elgar.

Darling, Kate (2017). "Who's Johnny?" Anthropomorphic framing in human-robot interaction, integration, and policy. In P. Lin, G. Bekey, K. Abney, R. Jenkins, eds., *Robot Ethics 2.0*. Oxford: Oxford.

Davenport, Matthew, and Ken D. Olum (2010). *Are there Boltzmann brains in the vacuum?* Manuscript at arXiv:1008.0808 [accessed Jul. 4, 2018].

Davidson, Donald (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association, 61,* 441–58.

De Simone, Andrea, Alan H. Guth, Andrei Linde, Mahdiyar Noorbala, Michael P. Salem, and Alexander Vilenkin (2010). Boltzmann brains and the scale-factor cutoff measure of the multiverse. *Physical Review D, 82,* 063520.

de Waal, Frans (1996). *Good natured.* Cambridge, MA: Harvard.

Dennett, Daniel C. (1991). *Consciousness explained.* Boston: Little, Brown.

Dennett, Daniel C. (1992). The self as a center of narrative gravity. In F. Kessel, P. Cole and D. Johnson, eds, *Self and consciousness.* Hillsdale, NJ: Erlbaum.

Dennett, Daniel C. (1996). *Kinds of minds.* New York: Basic Books.

Dennett, Daniel C. (2005). *Sweet dreams.* Cambridge, MA: MIT.

Descartes, René (1641/1984). *Meditations on first philosophy.* In *The Philosophical Writings of Descartes, vol. II,* trans. J. Cottingham, R. Stoothoff, and D. Murdoch. Cambridge: Cambridge.

Dimeo, Nate (2008). Iconoclastic comic George Carlin dies at 71. *NPR: Morning Edition*, radio transcript (Jun. 23). URL: https://genius.com/George-carlin-the-seven-words-you-can-never-say-on-television-annotated [accessed Jul. 4, 2018].

Dretske, Fred (1988). *Explaining behavior.* Cambridge, MA: MIT.

Dretske, Fred (1995). *Naturalizing the mind.* Cambridge, MA: MIT.

Dzokoto, Vivian Afi ,and Sumie Okazaki (2006). Happiness in the eye and the heart: Somatic referencing in West African emotion lexica. *Journal of Black Psychology, 32,* 17-140.

Egan, Greg (1994). *Permutation City.* London: Millennium.

Egan, Greg (1997). *Diaspora*. London: Millennium.

Ellenberg, Jordan S. (2013). Do we really underestimate how much we'll change? (Or: Absolute value is not linear!). *Quomodocumque* blog (Jan. 5). URL: https://quomodocumque.wordpress.com/2013/01/05/do-we-really-underestimate-how-much-well-change-or-absolute-value-is-not-linear/ [accessed Jul. 3, 2018].

Eskine, Kendall J., Natalie A. Kacinik, and Jesse J. Prinz (2011). A bad taste in the mouth: Gustatory disgust influences moral judgment. *Psychological Science, 22, 2*95-299.

Estrada, Daniel (2017). Robot rights. cheap, yo! *Made of Robots,* episode 1. (May 24) URL: https://www.madeofrobots.com/2017/05/24/episode-1-robot-rights-cheap-yo/

Epictetus (1st c. CE/1944). *Discourses and enchiridion,* trans. T.W. Higginson. Roslyn, NY: Walter J. Black.

Evans, C. Stephen (2013). *God and moral obligation.* Oxford: Oxford.

Faye, Emmanuel (2005/2009). *Heidegger,* trans. M.B. Smith. New Haven: Yale.

Fehr, Beverley, Deborah Samsom, and Delroy L. Paulhus (2013). The construct of Machiavellianism: Twenty years later. *Advances in Personality Assessement, vol. 9,* 77-116.

Feldman, Fred (2004). *Pleasure and the good life.* Oxford: Oxford.

Ferguson, R. Brian (2013a). Pinker's list: Exaggerating prehistoric war mortality. In D.P. Fry, ed., *War, peace, and human nature.* Oxford: Oxford.

Ferguson, R. Brian (2013b). The prehistory of war and peace in Europe and the Near East. In D.P. Fry, ed., *War, peace, and human nature.* Oxford: Oxford.

Festinger, Leon (1957). *A theory of cognitive dissonance.* Stanford, CA: Stanford.

Fetterman, Adam K., Tianwei Liu, and Michael D. Robinson (2015). Extending color psychology to the personality realm: Interpersonal hostility varies by red preferences and perceptual biases. *Journal of Personality, 83*, 106–116.

Feynman, Richard P. (1985). "Surely you're joking, Mr. Feynman", ed. E. Hutchings. New York: W.W. Norton.

Fiala, Brian, Adam Arico, and Shaun Nichols (2012). The psychological origins of dualism. In E. Slingerland and M. Collard, eds., *Creating consilience.* Oxford: Oxford.

Fischer, John Martin (1994). Why immortality is not so bad. *International Journal of Philosophical Studies, 2,* 257 – 270

Fischer, John Martin (1997). Death, badness, and the impossibility of experience. *Journal of Ethics, 1,* 341-353.

Fischer, John Martin (2005). Free will, death, and immortality: The role of narrative. *Philosophical Papers, 34,* 379-403.

Fischer, John Martin and Benjamin Mitchell-Yellin (2014). Immortality and boredom. *Journal of Ethics, 18,* 353-372.

Flanagan, Owen, Hagop Sarkissian, and David Wong (2007). Naturalizing ethics. In W. Sinnott-Armstrong, ed., *Moral psychology, vol. 1*. Cambridge, MA: MIT.

Foulkes, David (1999). *Children's dreaming and the development of consciousness.*

Frank, Erica, and Carolina Segura (2009). Health practices of Canadian physicians. *Canadian Family Physician, 55,* 810-1.e1-7.

Frankfurt, Harry G. (1986/2005). *On bullshit.* Princeton, NJ: Princeton.

Frankfurt, Harry G. (2004). *The reasons of love.* Princeton, NJ: Princeton.

Frege, Gottlob (1884/1953). *The foundations of arithmetic,* trans. J.L. Austin. New York: Philosophical Library.

Frege, Gottlob (1918/1956). The thought: A logical inquiry, trans. P.T. Geach. *Mind, 65,* 289-311.

Gaiman, Neil (1996/1998). The goldfish pool and other stories. In N. Gaiman, Smoke and Mirrors. New York: HarperCollins.

Gazzaniga, Michael S. (2005). Forty-five years of split-brain research and still going strong. *Nature Reviews Neuroscience, 6,* 653-659.

Garson, Justin (2016). Two types of psychological hedonism. *Studies in History and Philosophy of Biology and Biomedical Sciences, 56,* 7-14.

George, Susan, Janice Hanson, and Jeffrey L. Jackson (2014). Physician, heal thyself: A qualitative study of physician health behaviors. *Academic Psychiatry, 38,* 19-25.

Glausser, Wayne (1990). Three approaches to Locke and the slave trade. *Journal of the History of Ideas, 51,* 199-216.

Gleick, James (1987). *Chaos: Making a new science.* New York: Penguin.

Goff, Philip (2017). *Consciousness and fundamental reality.* Oxford: Oxford.

Goldstein, Noah J., Robert B. Cialdini, and Vladas Griskevicius (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research, 35*, 472-482.

Gómez, José María, Miguel Verdú, Adela González-Megías, and Marcos Méndez (2016). The phylogenetic roots of human lethal violence. *Nature 538,* 233-237.

Gordon, Peter E. (2014). Heidegger in Black. *New York Review of Books* (Oct. 9). URL: http://www.nybooks.com/articles/2014/10/09/heidegger-in-black/ [accessed Jun. 22, 2018].

Gosling, Samuel D., Oliver P. John, Kenneth H. Craik, and Richard W. Robins (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. *Journal of Personality and Social Psychology, 74*, 1337-1349.

Gott, J. Richard (2008). *Boltmann brains – I'd rather see than be one.* Manuscript at arXiv:0802.0233 [accessed Jul. 4, 2018].

Green, Peter (1972/2013). *Alexander of Macedon.* Berkeley: University of California Press.

Greene, Brian (2011). *The hidden reality.* New York: Vintage.

Griffin, James (1979). Is unhappiness morally more important than happiness? *Philosophical Quarterly, 29,* 47-55.

Haidt, Jonathan (2012). *The righteous mind.* New York: Random House.

Hall, Lars, Petter Johansson, and Thomas Strandberg (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PLOS ONE,* 7(9), e45457.

Hanh, Thich Nhat (1998/2015). *The heart of Buddha's teaching.* New York: Harmony.

Hansen, Thomas (2011). Parenthood and happiness: A review of folk theories versus empirical evidence. *Social Indicators Research, 108,* 29-64.

Hanson, Robin (2016). *The age of em.* Oxford: Oxford.

Hegel, G.F.W. (1807/1977). *Phenomenology of spirit,* trans. A.V. Miller. Oxford: Oxford.

Herbst, Chris M., and John Ifcher (2016). The increasing happiness of U.S. parents. *Review of the Economics of the Household, 14,* 529-551.

Herodotus (5th c. BCE/2017). *The essential Herodotus,* trans. W.A. Johnson. Oxford: Oxford.

Heschel, Susannah (2003). Orange on the Seder Plate. In S.C. Anifeld, T. Mohr, and C. Spector, eds., *The women's Passover companion.* Woodstock, VT: Jewish Lights.

Hilbert, Martin, and Priscila López (2011). The world's technological capacity to store, communicate, and compute information. *Science, 332,* 60-65.

Hill, Christopher S. (2009). *Consciousness*. Cambridge: Cambridge.

Hindriks, Frank, and Igor Douven (2018). Nozick's experience machine: An empirical study. *Philosophical Psychology, 31,* 278-298.

Hume, David (1740/1978). *A treatise of human nature,* ed. L.A. Selby-Bigge and P.H. Nidditch. Oxford: Oxford.

Hume, David (1751/1975). *Enquiry concerning the principles of morals.* In *Enquiries concerning human understanding and concerning the principles of morals, 3rd ed.,* ed. L.A. Selby-Bigge and P.H. Nidditch. Oxford: Oxford.

Hurlburt, Russell T. (1979). Random sampling of cognitions and behavior. *Journal of Research in Personality, 13,* 103-111.

Hurlburt, Russell T., and Eric Schwitzgebel (2007). *Describing inner experience? Proponent meets skeptic.* Cambridge, MA: MIT.

Hurlburt, Russell T. (2011). *Investigating pristine inner experience.* Cambridge: Cambridge.

Hursthouse, Rosalind, and Glen Pettigrove (2003/2017). Virtue ethics. *Stanford Encyclopedia of Philosophy* (Winter 2016 Edition). URL: https://plato.stanford.edu/archives/win2016/entries/ethics-virtue.

Ivanhoe, Philip J. (1990/2002). *Ethics in the Confucian tradition.* Indianapolis: Hackett.

James, Aaron (2012). *Assholes: A theory.* New York: Penguin.

Jarausch, Konrad H., and Gerhard Arminger (1989). The German teaching profession and Nazi party membership: A demographic logit model. *Journal of Interdisciplinary History, 20,* 197-225.

Jiang, Yuan, Michael K. Ong, Elisa K. Tong, Yan Yang, Yi Nan, Quan Gan, and Teh-wei Hu (2007). Chinese physicians and their smoking knowledge, attitudes, and practices. *American Journal of Preventive Medicine, 33,* 15-22.

Johansson, Petter, Lars Hall, Sverker Sikström, Betty Tärning, and Andreas Lind (2006). How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition, 15,* 673–692.

John, Oliver P., and Richard W. Robins (1993). Determinants of inter-judge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61*, 521-551.

Johnson, Robert, and Adam Cureton (2004/2016). Kant's moral philosophy. *Stanford Encyclopedia of Philosophy* (Spring 2018 Edition). URL https://plato.stanford.edu/archives/spr2018/entries/kant-moral.

Johnson, Susan C. (2003). Detecting agents. *Philosophical Transactions of the Royal Society B, 358,* 549-559.

Jones, Daniel N., and Delroy L. Paulhus (2014). Introducing the Short Dark Triad (SD3): A brief measure of dark personality traits. *Assessment, 21,* 28-41.

Jordan, Jennifer, Elizabeth Mullen, and J. Keith Murnighand (2011). Striving for the moral self: the effects of recalling past moral actions on future moral behavior. *Personality and Social Psychology Bulletin, 37,* 701-713.

Kahneman, Daniel, Alan B. Krueger, David A. Schkade, Norbert Schwarz, and Arthur A. Stone (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science, 306,* 1776-1780.

Kahneman, Daniel, and Gary Klein (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64,* 515-526.

Kammerer, François (2015). How a materialist can deny that the United States is probably conscious – response to Schwitzgebel. *Philosophia. 43,* 1047-1057.

Kant, Immanuel (1764/2007). Observations on the feeling of the beautiful and sublime. In H. Wilson and G. Zöller, eds., *Anthropology, history, and education.* Cambridge: Cambridge.

Kant, Immanuel (1775/2007). Of the different races of human beings. In H. Wilson and G. Zöller, eds., *Anthropology, history, and education.* Cambridge: Cambridge.

Kant, Immanuel (1781/1787/1998). *Critique of pure reason,* ed. and trans. P. Guyer and A.W. Wood. Cambridge: Cambridge.

Kant, Immanuel (1785/1996). *Groundwork of the metaphysics of morals,* trans. M.J. Gregor. In I. Kant, *Practical philosophy,* ed. M.J. Gregor. Cambridge: Cambridge.

Kant, Immanuel (1797/1991). *The metaphysics of morals,* trans. M.J. Gregor. In I. Kant, *Practical philosophy,* ed. M.J. Gregor. Cambridge: Cambridge.

Karlin, Beth, Joanne F. Zinger and Rebecca Ford. (2015). The effects of feedback on energy conservation: A meta-analysis. *Psychological Bulletin, 141,* 1205-1227.

Karpel, Ari (2014). How to be prolific: Guidelines for getting it done from Joss Whedon. *Fast Company* website (Jan. 9). URL: https://www.fastcompany.com/1683167/how-to-be-prolific-guidelines-for-getting-it-done-from-joss-whedon [accessed Jul. 5, 2018].

Kesey, Ken (1962). *One flew over the cuckoo's nes*t. New York: Penguin.

Kirk, Robert (2003/2015). Zombies. *Stanford Encyclopedia of Philosophy* (Summer 2015 Edition). URL: https://plato.stanford.edu/archives/sum2015/entries/zombies.

Kohlberg, Lawrence (1981). *The philosophy of moral development.* San Francisco: Harper & Row.

Kornblith, Hilary (1993). Epistemic normativity. *Synthese, 94,* 357-376.

Kornblith, Hilary (1998). The role of intuition in philosophical inquiry: An account with no unnatural ingredients. In *Rethinking intuition,* ed. M.R. DePaul and W. Ramsey. Lanham: Rowman and Littlefield.

Kriegel, Uriah (2011). Two defenses of common-sense ontology. *Dialectica, 65,* 177-204.

Kruger, Justin, and David Dunning (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121-34.

Kurzweil, Ray (2005). *The singularity is near.* New York: Penguin.

LaBerge, Stephen, and Howard Rheingold (1990). *Exploring the world of lucid dreaming.* New York: Ballatine.

Ladyman, James, and Don Ross (2007). *Every thing must go.* Oxford: Oxford.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenebaum, and Samuel J. Gershman (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences,* e253. DOI: 10.1017/S0140525X16001837.

Landy, Justin F., and Geoffrey P. Goodwin (2015). Does ncidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science, 10,* 518-536.

Langford, Simon, and Murali Ramachandran (2013). The products of fission, fusion, and teletransportation: An occasional identity theorist's perspective. *Australasian Journal of Philosophy, 91,* 105-117.

Langlois, Judith H., Lisa Kalakanis, Adam J. Rubenstein, Andrea Larson, Monica Hallam, and Monica Smoot (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin, 126,* 390-423.

Lau, Joe and Max Deutsch (2002/2016). Externalism about mental content. *Stanford Encyclopedia of Philosophy* (Fall 2018 Edition). URL: https://plato.stanford.edu/archives/fall2018/entries/functionalism.

Leaman, George (1993). *Heidegger im Kontext.* Hamburg: Argument-Verlag.

Leary, Timothy (1988/2003). *Musings on human metamorphoses*. Berkeley, CA: Ronin.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). Deep learning. *Nature, 521,* 436-444.

Levin, Janet (2004/2018). Functionalism. *Stanford Encyclopedia of Philosophy* (Winter 2016 Edition). URL: https://plato.stanford.edu/archives/win2016/entries/content-externalism.

Lewis, David K. (1980). Mad pain and Martian pain. In N. Block (ed.), *Readings in philosophy of psychology.* Cambridge, MA: Harvard.

Lewis, David K. (1986). *On the plurality of worlds.* Malden, MA: Blackwell.

Lilienfeld, Scott O., and Brian P. Andrews (1996). Development and preliminary validation of a self-report measure of psychopathic personality traits in noncriminal populations. *Journal of Personality Assessment, 66,* 488-524.

Locke, John (1689/2016). *Second treatise of government and A letter concerning toleration.* Ed. M. Goldie. Oxford: Oxford.

Lorenz, Edward U. (1972). Predictability: Does the flap of a butterfly's wings in Brazil set off a tornado in Texas? Address to the American Association for the Advancement of Science (Dec. 29). Text available at https://www.ias.ac.in/article/fulltext/reso/020/03/0260-0263 [accessed Jul. 4, 2018].

Margolis, Rachel, and Mikko Myrskylä (2011). A global perspective on happiness and fertility. *Population and Development Review, 37,* 29-56.

Mazar, Nina, and Chen-Bo Zhong (2010). Do green products make us better people? *Psychological Science, 21,* 494-498.

McDowell, John (1985). Values and secondary qualities. In *Morality and Objectivity*, ed. Ted Honderich. Routledge & Kegan Paul.

McFarland, Matt (2017). California is officially embracing the self-driving car. CNN Tech (Mar. 10). URL: https://money.cnn.com/2017/03/10/technology/california-dmv-self-driving-car/index.html [accessed Jul. 13, 2018].

McGeer, Victoria (1996). Is "self-knowledge" an empirical problem? Renegotiating the space of philosophical explanation. *Journal of Philosophy, 93,* 483-515.

McKay, Thomas and Michael Nelson, Michael (2010/2014). Propositional attitude reports: The de re / de dicto distinction. *Stanford Encyclopedia of Philosophy* (Spring 2014 Edition) URL https://plato.stanford.edu/entries/prop-attitude-reports/dere.html.

McKinsey, Michael (1991). Anti-individualism and privileged access. *Analysis, 51,* 9-16.

Melton, Wayne Rollan (2001-2009). *How to eliminate fear of global economic recession and terrorism.* Reno, NV: Fix Bay.

Meltzoff, Andrew N., Rechele Brooks, Aaron P. Shon, and Rajesh P.N. Rao (2010). "Social" robots are psychological agents for infants: A test of gaze following. *Neural Networks, 23,* 966-972.

Mengzi (4th c. BCE/2008). *Mengzi,* trans. B.W. Van Norden. Indianapolis: Hackett.

Metzinger, Thomas (2003). *Being no one.* Cambridge, MA: MIT.

Miernowski, Jan (2016). Montaigne on truth and skepticism. In P. Desan, ed., *The Oxford handbook of Montaigne.* New York: Oxford.

Mikkelson, David, and Dan Evon (2007/2017). Al Gore's home energy use. Snopes.com. URL: https://www.snopes.com/fact-check/al-gores-energy-use/ [accessed Jul. 3, 2018].

Miller, Joshua D., and W. Keith Campbell (2008). Comparing clinical and social-personality conceptualizations of narcissism. *Journal of Personality, 76,* 449-476.

Millikan, Ruth (1984). *Language, thought, and other biological categories.* Cambridge, MA: MIT.

Millikan, Ruth Garrett (2010). On knowing the meaning: With a coda on Swampman. *Mind, 119,* 43-81.

Mills, Charles W. (2005). Kant's *Untermenschen.* In A. Valls, ed., *Race and racism in modern philosophy.* Cornell.

Montaigne, Michel de (1580/1595/1957). *The complete essays,* trans. D.M. Frame. Stanford: Stanford.

Moody-Adams, Michele M. (1997). *Fieldwork in familiar places.* Cambridge, MA: Harvard.

Moore, Andrew (2004/2013). Hedonism. *Stanford Encyclopedia of Philosophy* (Winter 2013 Edition). URL: https://plato.stanford.edu/archives/win2013/entries/hedonism.

Moore, G.E. (1922). *Philosophical studies.* London: Kegan, Paul, Trench, Trubner.

Moore, G.E. (1925). A defence of common sense. In *Contemporary British philosophy,* ed. J.H Muirhead. London: George Allen & Unwin. Pp. 191-223.

Moore, G.E. (1953). *Some main problems of philosophy.* London: George Allen & Unwin.

Moore, G.E. (1957). Visual sense data. In *British philosophy in mid-century,* ed. C.A. Mace. London: George Allen & Unwin.

Moravec, Hans (1999). Rise of the robots. *Scientific American, 781* (6), 124-135.

Mozi (5th c. BCE/2013). *Mozi: A study and translation of the ethical and political writings,* trans. J. Knoblock and J. Riegel. Institute of East Asian Studies.

Murdock-Perriera, Lisel Alice, and Quentin Charles Sedlacek (2018). Questioning Pygmalion in the twenty-first century: The formation, transmission, and attributional influence of teacher expectancies. *Social Psychology of Education, 21,* 691-707.

Mullen, Elizabeth, and Benoît Monin (2016). Consistency versus licensing effects of past moral behavior. *Annual Review of Psychology, 67,* 363-385.

Müller-Lauter, Wolfgang (1971/1999). Nietzsche: His philosophy of contradictions and the contradictions of his philosophy, trans. D.J. Parent. New York: University of Illinois.

Nagel, Thomas (1979). *Mortal questions.* Cambridge: Cambridge.

Neander, Karen (1996). Swampman meets swampcow. *Mind & Language, 11,* 118-129.

Nietzsche, Friedrich (1887/1998). *On the genealogy of morality,* trans. M. Clark and A.J. Swensen. Indianapolis: Hackett.

Nozick, Robert (1974). *Anarchy, state, and utopia.* New York: Basic Books.

O'Neill, Elizabeth (2017). Kinds of norms. Philosophy Compass, 12, e12416. DOI: 10.1111/phc3.12416.

O'Toole, Garson (2014). You just chip away everything that doesn't look like David. *Quote Investigator* (June 22). URL: https://quoteinvestigator.com/2014/06/22/chip-away/ [accessed Jun. 21, 2018].

Oizumi, Masafumi, Larissa Albantakis, and Giulio Tononi (2014). Consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology.* https://doi.org/10.1371/journal.pcbi.1003588

Parfit, Derek (1984). *Reasons and persons.* Oxford: Oxford.

Paulhus, Delroy L., and Kevin M. Williams (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality, 36,* 556-563.

Petersen, Steve (2012). Designing people to serve. In P. Lin, K. Abney, and G.A. Bekey, eds., Robot ethics. Cambridge, MA: MIT.

Piaget, Jean (1932/1975). *The moral judgment of the child.* Trans. M. Gabain. New York: Free Press.

Pichler, Alois (2007). The interpretation of the *Philosophical Investigations:* Style, therapy, *Nachlass*. In G. Kahane, E. Kanterian, and O. Kuusela, eds., *Wittgenstein and his interpreters.* Malden, MA: Blackwell.

Pinker, Steven (2011). *The better angels of our nature.* New York: Penguin.

Popper, Karl (1945/1994). *The open society and its enemies.* Princeton, NJ: Princeton.

Priest, Graham, Francesco Berto, and Zach Weber (1998/2018). Dialetheism. *Stanford Encyclopedia of Philosophy* (Fall 2018 Edition). URL: https://plato.stanford.edu/archives/fall2018/entries/dialetheism.

Putnam, Hilary (1965). Psychological predicates. In *Art, mind, and religion,* ed. W.H. Capitan & D.D. Merrill. Liverpool: University of Pittsburgh Press / C. Tinling.

Putnam, Hilary (1975). The meaning of 'Meaning'. *Philosophical Papers, vol. 2*. Cambridge: Cambridge.

Putnam, Hilary (1981). *Reason, truth, and history.* Cambridge: Cambridge UP.

Quoidbach, Jordi, Daniel T. Gilbert, and Timothy D. Wilson (2013). The end of history illusion. *Science, 339,* 96-98.

Railton, Peter (1986). Moral realism. *Philosophical Review, 95,* 163-207.

Reck, Erich (2005). Frege on numbers: Beyond the Platonist picture. *Harvard Review of Philosophy, 13* (2), 25-40.

Reid, Thomas (1774-1778, 1995). Materialism. In *Thomas Reid on the animate creation,* ed. P. Wood. University Park, PA: Pennsylvania State University.

Reid, Thomas (1788/2010). *Essays on the active powers of man,* ed. K Haakonssen and J.A. Harris. University Park, PA: Pennsylvania State University.

Richards, J. G. (1999). The health and health practices of doctors and their families. *New Zealand Medical Journal, 112,* 96-99.

Ringer, Fritz K. (1969). *The decline of the German mandarins.* Hanover, NH: University Press of New England.

Roache, Rebecca (2015). Rebecca Roache on swearing. *Philosophy Bites* podcast (Mar. 29). URL: http://philosophybites.com/2015/03/rebecca-roache-on-swearing.html [accessed Jul. 4, 2018].

Roache, Rebecca (2016). Naughty words. Aeon Magazine (Feb. 22). URL: http://philosophybites.com/2015/03/rebecca-roache-on-swearing.html [accessed Jul. 4, 2018].

Roese, Neal J. and Kathleen D. Vohs (2012). Hindsight bias. *Perspectives on Psychological Science, 7,* 411-426.

Rosenthal, Robert, and Lenore Jacobson (1968/1992). *Pygmalion in the classroom.* Norwalk, CT: Crown House.

Rousseau, Jean-Jacques (1755/1997). *Discourse on the origins of inequality*. In. J.-J. Rousseau, *The discourses and other early political writings,* trans. V. Gourevitch. Cambridge: Cambridge.

Rousseau, Jean-Jacques (1762/1979). *Emile,* trans. A. Bloom. Basic Books.

Rousseau, Jean-Jacques (1769/1995). *The confessions.* In *The collected writings of Rousseau, vol. 5,* trans. C. Kelly, ed. C. Kelly, R.D. Masters, and P.D. Stillman. Hanover, NH: University Press of New England.

Rust, Joshua, and Eric Schwitzgebel (2015). The moral behavior of ethicists and the power of reason. In *Advances in experimental moral psychology,* ed. H. Sarkissian and J.C. Wright. London: Bloomsbury.

Sacks, Oliver (1985). *The man who mistook his wife for a hat.* London: Duckworth.

Sarna, Linda, Stella Aguinaga Bialous, Karabi Sinha, Qing Yang, and Mary Ellen Wewers (2010). Are health care providers still smoking? Data From the 2003 and 2006/2007 Tobacco Use Supplement-Current Population Surveys. *Nicotine and Tobacco Research, 12,* 1167-1171.

Schiller, Henry Ian (forthcoming). Phenomenal dispositions. *Synthese.*

Schnall, Simone, Jonathan Haidt, Gerald L. Clore, and Alexander H. Jordan (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin, 34,* 1096-1109.

Schneider, Susan (forthcoming). *The future of the mind.* Manuscript.

Schroeder, Timothy (2004). *Three faces of desire.* Oxford: Oxford.

Schultz, P. Wesley, Jessica M. Nolan, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science, 18,* 429-434.

Schwitzgebel, Eric (1999). Gradual belief change in children. *Human Development, 42,* 283-296.

Schwitzgebel, Eric (2007). Human nature and moral education in Mencius, Xunzi, Hobbes, and Rousseau. *History of Philosophy Quarterly, 24,* 147-168.

Schwitzgebel, Eric (2009). Do ethicists steal more books? *Philosophical Psychology, 22,* 711-725.

Schwitzgebel, Eric (2011). *Perplexities of consciousness.* Cambridge, MA: MIT.

Schwitzgebel, Eric (2012). Self-Ignorance. In J. Liu and J. Perry, eds., *Consciousness and the self.* Cambridge: Cambridge.

Schwitzgebel, Eric (2013). My Boltzmann continuants. Blog post at *The Splintered Mind* (Jun. 6). URL: http://schwitzsplinters.blogspot.com/2013/06/my-boltzmann-continuants.html [accessed Jul. 5, 2018].

Schwitzgebel, Eric, and R. Scott Bakker (2013). Reinstalling Eden. *Nature, 503,* 562.

Schwitzgebel, Eric (2014a). A theory of jerks. *Aeon Magazine* (June 4). URL: https://aeon.co/essays/so-you-re-surrounded-by-idiots-guess-who-the-real-jerk-is [accessed Jun. 22, 2018].

Schwitzgebel, Eric (2014b). The crazyist metaphysics of mind. *Australasian Journal of Philosophy, 92,* 665-682.

Schwitzgebel, Eric (2014d). Tononi's Exclusion Postulate would make consciousness (nearly) irrelevant. Blog post at *The Splintered Mind* (Jul. 16). URL: https://schwitzsplinters.blogspot.com/2014/07/tononis-exclusion-postulate-would-make.html [accessed Nov. 1, 2018].

Schwitzgebel, Eric, and Joshua Rust (2014). The moral behavior of ethics professors: Relationships among self-reported behavior, expressed normative attitude, and directly observed behavior. *Philosophical Psychology, 27,* 293-327.

Schwitzgebel, Eric (2015a). If materialism is true, the United States is probably conscious. *Philosophical Studies, 172*, 1697-1721.

Schwitzgebel, Eric (2015b). The Dauphin's metaphysics. *Unlikely Story,* issue 12. URL: http://www.unlikely-story.com/stories/the-dauphins-metaphysics-by-eric-schwitzgebel [accessed Jul. 4, 2018].

Schwitzgebel, Eric (2015c). The tyrant's headache. *Sci Phi Journal,* issue 3, 78-83.

Schwitzgebel, Eric, and Mara Garza (2015). A defense of the rights of Artificial Intelligences. *Midwest Studies in Philosophy, 39,* 98-119.

Schwitzgebel, Eric, and Alan T. Moore (2015). Experimental evidence for the existence of an external world. *Journal of the American Philosophical Association, 1,* 564-582.

Schwitzgebel, Eric (2016). Is the United States Phenomenally Conscious? Reply to Kammerer. *Philosophia, 44,* 877-883.

Schwitzgebel, Eric, and Joshua Rust (2016). The behavior of ethicists. In *A Companion to Experimental Philosophy,* ed. J. Sytsma and W. Buckwalter. Malden, MA: Wiley Blackwell.

Schwitzgebel, Eric (2017a). 1% skepticism. *Noûs, 51,* 271-290.

Schwitzgebel, Eric (2017b). Fish dance. *Clarkesworld, 118.* URL: http://clarkesworldmagazine.com/schwitzgebel_07_16

Schwitzgebel, Eric, and Jonathan E. Ellis (2017). Rationalization in moral and philosophical thought. In J.-F. Bonnefon and B. Tremoliere, eds., *Moral inferences.* Psychology Press.

Schwitzgebel, Eric (2018a). *Aiming for moral mediocrity.* Unpublished draft manuscript. URL: http://www.faculty.ucr.edu/~eschwitz/SchwitzAbs/MoralMediocrity.htm [accessed Jul. 2, 2018].

Schwitzgebel, Eric (2018b). Death, self, and oneness in the incomprehensible Zhuangzi. In P.J. Ivanhoe, O.J. Flanagan, V.S Harrison, S. Sarkissian, and E. Schwitzgebel, eds., *The oneness hypothesis.* New York: Columbia University.

Schwitzgebel, Eric (2018c). Is there something it's like to be a garden snail? Unpublished draft manuscript. URL: http://www.faculty.ucr.edu/~eschwitz/SchwitzAbs/Snails.htm [accessed Nov. 2, 2018]

Schwitzgebel, Eric (forthcoming). Kant meets cyberpunk. *Disputatio.*

Schwitzgebel, Eric, and Mara Garza (forthcoming). "Designing AI with rights, consciousness, self-respect, and freedom". In S.M. Liao, ed., *The Ethics of Artificial Intelligence.* New York: Oxford.

Schwitzgebel, Ralph K. (1965). *Streetcorner research.* Cambridge, MA: Harvard.

Schwitzgebel, Ralph K., and Robert W. Taylor (1980). Impression formation under conditions of spontaneous and shadowed speech. *Journal of Social Psychology, 110,* 253-263.

Searle, John R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3,* 417-457.

Searle, John R. (1984). *Minds, brains, and science.* Cambridge, MA: Harvard.

Searle, John R. (1992). *The rediscovery of the mind.* Cambridge, MA: MIT.

Sedikides, Constantine, and Aiden P. Gregg (2008). Self-enhancement: Food for thought. *Perspectives on Psychological Science, 13,* 102-116.

Sextus Empiricus (circa 200 CE/1994). *Outlines of skepticism,* trans. J. Annas and J. Barnes. Cambridge: Cambridge.

Sezer, Hafize, Nuran Guler, and R. Erol Sezer (2007). Smoking among nurses in Turkey: Comparison with other countries. *Journal of Health, Population, and Nutrition, 25,* 107-111.

Shaikh, Raees A., Asia Sikora, Mohammad Siahpush, and Gopal K. Singh (2015). Occupational variations in obesity, smoking, heavy drinking, and non-adherence to physical activity recommendations: Findings from the 2010 National Health Interview Survey. *American Journal of Industrial Medicine, 58,* 77-87.

Shelley, Mary (1818/1965). *Frankenstein.* New York: Signet.

Shepperd, Klein, Waters, and Weinstein (2013). Taking stock of unrealistic optimism. *Perspectives on Psychological Science, 8,* 395-411.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529,* 484-489.

Singer, Peter (1975/2009). *Animal liberation.* New York: HarperCollins.

Singer, Peter (2009). *The life you can save.* New York: Random House.

Sinnott-Armstrong, Walter (2003/2015). Consequentialism, *Stanford Encyclopedia of Philosophy* (Winter 2015 Edition). URL: https://plato.stanford.edu/archives/win2015/entries/consequentialism.

Skinner, B. F. (1948/1972). *Walden two.* New York: Macmillan.

Sluga, Hans (1993). *Heidegger's crisis.* Cambridge, MA: Harvard.

Smart, R.N. (1958). Negative utilitarianism.

Smith, Derek R., and Peter A. Leggat (2007). Tobacco smoking by occupation in Australia: Results from the 2004 to 2005 National Health Survey. *Journal of Occupational and Environmental Medicine, 49,* 437-445.

Snodgrass, Melinda M., and Robert Scheerer (1989). The measure of a man. *Star Trek: The*

Sober, Elliott, and David Sloan Wilson (1998). *Unto others.* Cambridge, MA: Harvard.

Sparrow, Robert (2004). The Turing triage test. *Ethics and information technology, 6,* 203-213.

Sperber, Dan (2010). The guru effect. *Review of Philosophy and Psychology, 1*, 583-592.

Spinoza, Benedict de (1677/1994). Ethics, trans. E. Curley. New York: Penguin.

Sprigge, T.L.S. (1999). The relationship between Jeremy Bentham's psychological, and his ethical, hedonism. *Utilitas, 11,* 296-319.

Squier, Christopher, Vicki Hesli, John Lowe, Victor Ponamorenko, and Natalia Medvedovskaya (2006). Tobacco use, cessation advice to patients and attitudes to tobacco control among physicians in Ukraine. *European Journal of Cancer Prevention, 15,* 548-563.

Steinhart, Eric (2014). *Your digital afterlives.* New York: Palgrave.

Stich, Stephen (1983). *From folk psychology to cognitive science.* Cambridge, MA: MIT.

Strawson, Galen (2006). *Consciousness and its place in nature,* ed. A. Freeman. Exeter: Imprint Academic.

Strawson, P.F. (1985). *Skepticism and naturalism.* New York: Columbia.

Stross, Charles (2005). *Accelerando.* New York: Ace.

Susewind, Moritz, and Erik Hoelzl (2014). A matter of perspective: why past moral behavior can sometimes encourage and other times discourage future moral striving. *Journal of Applied Social Psychology, 44,* 1722-1731.

Stangneth, Bettina (2014). *Eichmann before Jerusalem.* New York: Penguin.

Swinburne, Richard (1977/2016). *The coherence of theism, 2nd ed.* Oxford: Oxford.

Taylor, Shelley E., and Jonathon D. Brown (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103,* 193-210.

Tegmark, Max (2014). *Our mathematical universe.* New York: Random House.

Thompson, Evan (2015). *Dreaming, waking, being.* New York: Columbia.

Timmer, Erika, Gerben J. Westerhof, and Freya Dittmann-Kohli (2005). "When Looking Back on My Past Life I Regret…": Retrospective Regret in the Second Half of Life. *Death Studies, 29*, 625-644.

Tononi, Giulio (2012). The integrated information theory of consciousness: An updated account. *Archives Italiennes de Biologie, 150,* 290-326.

Tononi, Giulio (2014). *Why Scott should stare at a blank wall and reconsider (or, the conscious grid).* Unpublished manuscript. URL: http://www.scottaaronson.com/tononi.docx [accessed Jul. 5, 2018].

Tversky, Amos, and Daniel Kahneman (1991). Loss aversion in riskless choice: A reference dependent model. *Quarterly Journal of Economics, 107*, 1039-1061.

Tversky, Amos, and Daniel Kahneman (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5,* 297-323.

Tye, Michael (1997). A representational theory of pains and their phenomenal character. In N. Block, O. Flanagan, and G. Güzeldere (eds.), *The nature of consciousness.* Cambridge, MA: MIT.

Tye, Michael (2000). *Consciousness, color, and content.* Cambridge, MA: MIT.

Tye, Michael (2005a). Another look at representationalism about pain. In M. Aydede (ed.), *Pain*. Cambridge, MA: MIT.

Tye, Michael (2005b). In defense of representationalism: A reply to commentaries. In M. Aydede (ed.), *Pain*. Cambridge, MA: MIT.

Van Patten, Tim, and Charlie Brooker (2017). *Black mirror,* Hang the DJ (S4:E4). House of Tomorrow.

Vedantam, Shankar (2017). Can robots teach us what it means to be human? *Hidden Brain* podcast (Jul. 10). Transcript at URL: https://www.npr.org/templates/transcript/transcript.php?storyId=536424647 [accessed Jul. 2, 2018].

Velleman, J. David (2005). The self as narrator. In J. Christman and J. Anderson, eds., *Autonomy and the challenges to liberalism.* Cambridge: Cambridge.

Wallace, David (2015). Recurrent theorems: A unified account. *Journal of Mathematical Physics 56,* 022105. doi: 10.1063/1.4907384.

Ware, Bronnie (2011). *Top five regrets of the dying.* Hay House.

Weijers, Dan (2014). Nozick's experience machine is dead, long live the experience machine! *Philosophical Psychology, 27,* 513-535.

Weinberg, Jonathan M., Chad Gonnerman, Cameron Buckner, and Joshua Alexander (2010). Are philosophers expert intuiters? *Philosophical Psychology, 23,* 331-355.

Weizenbaum, Joseph (1976). *Computer power and human reason.* San Francisco: W.H. Freeman.

Wheatley, Thalia, and Jonathan Haidt (2005). Hypnotically induced disgust makes moral judgments more severe. *Psychological Science, 16,* 780-784.

Williams, Bernard (1973). *Problems of the self.* Cambridge: Cambridge.

Wilson, Erin Faith (2017). 17 antigay leaders exposed as gay or bi. *The Advocate* (Nov. 21). URL: https://www.advocate.com/politics/2017/11/21/17-antigay-leaders-exposed-gay-or-bi [accessed Jul. 3, 2018].

Wolf, Susan (1982). Moral saints. *Journal of Philosophy, 79,* 419-439.

Wolfram, Stephen (2002). *A new kind of science.* Champaign, IL: Wolfram media.

Xunzi (3rd c. BCE/2014). *Xunzi,* trans. E. Hutton. Princeton, NJ: Princeton.

Yablo, Stephen (1987). Identity, essence, and indiscernibility. Journal of Philosophy, 84, 293-314.

Yankovic, Weird Al (2006). *Straight outta Lynwood.* Volcano Entertainment.

Young, Julian (1997). *Heidegger, philosophy, Nazism.* Cambridge: Cambridge.

Zawidzki, Tadeusz (2013). *Mindshaping.* Cambridge, MA: MIT.

Zhuangzi (4th c. BCE/2009). *The essential writings,* trans. B. Ziporyn. Indianapolis: Hackett.