A Life Hack with a Future

Elijah Millgram
Department of Philosophy
University of Utah
215 Central Campus Drive
Salt Lake City UT 84112
elijah.millgram@gmail.com

July 24, 2026

This is a DRAFT. Do not cite, circulate or copy without permission. Comments, criticism and suggestions welcome.

Contents

Acknowledgements

Chrisoula Andreou, Sarah Buss and C. Thi Nguyen worked through first versions of the great majority of this material; I'm very appreciative of the combination of skeptically raised eyebrows and the willingness to wrap one's mind around whatever was coming down the pike.

For comments on drafts of chapters, I'm grateful to Carla Bagnoli, Erin Beeghly, Jon Bendor, Szymon Bogacz, Margaret Bowman, Phoebe Chan, Sarah Childs, Steve Downes, Christoph Fehige, Leslie Francis, Svantje Guinebert, Matt Haber, Alan Hajek, Eric Hutton, Kimberly Johnston, Linh Mac, Jeff McMahan, Adam Morton, Madeleine Parkinson, Amy Schmitter, Maggie Shea, Aubrey Spivey, Michael Wilhelm, and several anonymous reviewers.

In addition, I benefited from helpful conversation with Eli Dresner, Michael Gill, David Humphries, Jenann Ismael, Buket Korkut-Raptis, Alasdair MacIntyre, Hillel Millgram, Shaun Nichols, Dan Russell, Jan Schiller, Sydney Shoemaker, Christian Skirke, Karola Stotz, Ulla Wessels, and Laura Wharton.

I was able to try out rough cuts of this material in front of a good many thoughtful and critical audiences: at ISHPSSB 2011, the University of Bremen, the University of Wisconsin/Milwaukee, the UC/Berkeley EMST Reasoning Group, the University of Canterbury, Tel-Hai College, the University of Amsterdam, Stanford University, Universität Konstanz, the University of the Saarland, Ben Gurion University, Tel Aviv University, the Australian National University, the University of Alberta, the Epistemology of Expert Judgement lecture series, the Università degli Studi di Modena e Reggio Emilia, Weber State University, Monash University, the University of Leipzig, Mannheim University, the University of Luxembourg, the Chattanooga Workshop on Aggregation, Rational Choice and Agency over the Lifecycle, the Joint Colloquium in Theoretical Philosophy, and the University of Utah's Practical Reasoning Kaffeeklatsch, as well as its Philosophy Club. I benefitted as well from feedback at the Liberty Fund symposium on corruption. I'm grateful to the Fund, as well to the participants for the lively conversation.

Work on sundry stages of the project was supported by the hospitality of a number of institutions: the Center for the Philosophy of Freedom at the University of Arizona, the University of the Saarland's Chair of Practical Philosophy, the Australian National University's Research School of Social Sciences, the Hebrew University, which provided a Lady Davis Fellowship, the Sackler Institute of Advanced Studies at Tel Aviv University, the Univer-

sity of Canterbury, through a Visiting Erskine Fellowship, and the Bogliasco Foundation. At my home institution, the University of Utah's Philosophy Department contributed through Sterling M. McMurrin Esteemed Faculty Awards, and its College of Humanities provided travel funding. For all of these, I'm enormously thankful.

Earlier versions of some of this material appeared as

Private Persons and Minimal Persons, *Journal of Social Philosophy* 45(3), Fall 2014: 323–347.

Hypophilosophy, *Social Philosophy and Policy* 35(2), Winter 2018: 138–157.

'. . . '. In P. Brösell, A.-M. Eder and T. Grundmann, *The Epistemology of Experts* (New York: Routledge, 2026): 203–222.

Cass Sunstein, *Decisions About Decisions: Practical Reason in Ordinary Life. Notre Dame Philosophical Reviews.* February 2025.

Chapter 1

Introduction

Full disclosure at the very outset: we have a complicated and ambitious agenda, and it comes in layers. There's an overarching methodological layer, advancing an approach on which metaphysics is understood as a design and engineering science. There's an extended investigation of one of the central topics of metaphysics, the *self*. There is the appropriation and adaptation of a mode of argumentation—to anticipate, creature construction arguments—that I'm going to position as especially suitable to the approach and the topic. There is a map to be drawn, one that will locate selves in the intersection of a handful of philosophical debates: just flagging these with keywords for the moment (and I'll introduce all of this in due course, so don't worry if they're not yet familiar), personal identity, first-person authority, and extended mind, but also, in philosophy of logic, defeasible inference, as well as what has recently come to be called the adoption problem. And there's a followon agenda under development: the progressively more complex society we live in today, with its deeper and more demanding forms of division of labor, requires us to rethink how selves are implemented and managed.

I'm going to be making a case that all of that is *one* topic, and has to be handled all at once, but it's also a good many moving parts to keep track of, and I'll need to do more than usual to orient my readers. Right now, I'm first going to provide a high-level sketch of the self and what, on the account we'll be developing, it's there to do. Then I'll preview and motivate the organization of the coming argument.

1.1

Take two contradictory propositions, for instance, that on the one hand a locavore restaurant, which sources its ingredients from and prepares dishes specific to the surrounding countryside, can't be a chain; that on the other, any successful restaurant can of course be scaled up into a franchise. (But let's be schematic, and go ahead with ps and qs to represent propositions like these.) If I believe the former—p—and I also believe that $\sim p$ (with the '$\sim$' for "not," that's shorthand for the latter), I'm contradicting myself; I'm inconsistent; if I don't change my mind when it's pointed out to me, I'm irrational.

But now, if *I* believe that p, and *you* believe that $\sim p$, we disagree, but no one is, so far, inconsistent; no one is contradicting himself; no one is—so far—required to change his mind. That's to say that our selves are serving as the scopes for anyway some (in fact, *many*) of the consistency requirements we impose and administer. That's one of the primary functions of a self, and it justifies overlaying this class of virtual object on the human animals that we are. (And does it also justify analogous overlays on the AI agents that seem to be in the process of joining us? I'll consider that question briefly as we finish up the discussion.)

What's the payoff of doing that? In the first place, the demand that inconsistencies be resolved drives inference. If you believe that p, that if p, then q, but also that $\sim q$, you're inconsistent. And if you resolve the inconsistency by deleting $\sim q$ from your field of representations, and replacing it with q, you're inferring q from your other beliefs.

But more generally and more importantly, the pressure to resolve inconsistencies motivates people to *figure things out*. Consistency regimes don't *just* force inferences in the way at which I just gestured. If I believe that p, and I believe that $\sim p$, I'm required to give one of them up, but obviously I shouldn't decide which by simply flipping a coin. Rather, identifying the inconsistency launches an investigation: maybe reviewing my sources of information or my reasoning, to see where I might have made a mistake; but maybe empirical fieldwork, or experimentation, to find out which one is actually correct. Sometimes, in cases that are among the deepest and most interesting, it becomes necessary to revise the concepts in play, and even the intellectual tools you operate to work with your concepts. Going back to that illustration, you might have been inclined simply to sacrifice the suggestion that, as a metaphysician might say it, these sorts of particulars (restaurants) can be converted into universals (franchises) as broadly as that. But Eataly—a corporation, that is, a legal person, subject to con-

sistency constraints modeled on those of selves—found a way to split the difference: to be a chain, but also a locavore eating establishment and grocery business. Inconsistencies are not merely mistakes to be expunged, but rather function as cognitive fuel.

Pairing the demand for consistency with selves as its scope has reshaped human life as a whole, because it is such an effective and fertile expedient for driving innovation and advances in all walks of life, both theoretical and practical. Across the ways that humans do things, the tools they use, and the theories they adopt, the explanation in the background is, frequently enough, that someone figured it out.[1] Animals eat; we eat in restaurants, and the procedures involved in running a restaurant arose—almost all of them—out of problems that someone had to solve. A very important way to flag problems as *problems* is to identify a violation of a consistency requirement. So the cumulative effects of our practices of consistency management have reshaped and inflected our repertoire of technique pretty much across the board.

Consistency management consequently proves central to an understanding of just about everything we do. Selves are an indispensible support for consistency management, because we can't require consistency without specifying *what* has to be kept consistent. Still, why *selves*? Why not simply require that all the information that *anyone* has, taken altogether, be adjusted into a state of global consistency? As we now know, from half a century of computational complexity theory, that's not possible. Humans are boundedly rational, and that fact is a constraint on the procedures we implement, and so on the virtual objects that figure into them.

Metaphysics has traditionally been conceived as a theory of exotic invisible objects that are somehow the basic or primary components of the world. However, the sense of what it is to be a bottom-level component in this way is very much at cross-purposes to that of physicists.[2] I'll need a shorthand for this way of understanding what metaphysics is after, and so, throughout, when I mention *old school* metaphysics, this is what I have in mind. Now, the framing assumption of the upcoming discussion is that this is a misconception. In fact metaphysicians have been, unawares, tracing the outlines of a cluster of interconnected virtual objects that play central roles in our reasoning and decisionmaking; that is, one of the larger ambitions of the discussion will be to show metaphysics to all along have been, not an impractical business of building castles in the sky, but rather, as I'm going

[1] Not always: a point that I'll take up in our second Interlude.

[2] That contrast is spelled out further, with an illustration, in Millgram, 2013, p. 400.

to put it, intellectual ergonomics.

It's worth keeping in mind that inference—and figuring things out more generally—is in good part about what to do, and about the values agents use to orient their choices. (I.e., it's not confined to theory.) The consistency regimes that drive inference include practical constraints. It's most usual to invoke the requirement of means-end consistency, and economists rely on consistency of preferences, but we also sometimes require compatibility of values; moreover, once selves are in play, the question of how to manage, as someone might imagine it, fleets of selves becomes pressing. So there are questions about how to coordinate commitments and choices across individuals. However, I'm going to for the most part defer the treatments of the problems of specifically practical rationality and with them, issues in moral and political philosophy, to a sequel.

1.2

The investigation we're embarking on has both forward- and backward-facing components. This book will be devoted to the latter, backward-facing task, but that can only be executed successfully if we keep the forward-facing agenda clearly in view throughout. And so I will also be providing a preview of the forward-facing complement.

The backward-facing enterprise is to provide engineering analyses of the virtual constructions we currently deploy in inference management. Here we'll aim at what used to be called rational reconstructions, with the difference that we are not chasing old-school definitions, that is, packages of necessary and sufficient conditions, but are in the quite different business of design analysis. For this purpose, creature construction arguments are an appropriate tool, and will serve as our primary resource.

Quick background: Creature construction arguments were introduced into the philosophical conversation with little fanfare or uptake at the time, by Paul Grice in 1975; they have in the meantime been further developed and deployed by philosophers such as Michael Bratman. In such an argument, a simplified model of an agent, one from which a feature or capability that is being investigated has been stripped out, is inspected for performance shortfalls. When those shortfalls are shown to be remediated by reinserting the feature or capability, a functional account of the target feature is on hand—and typically, a legitimization of it. You'll likely recognize creature construction arguments as close kin to state of nature arguments, but they bypass the latter's requirement that the agents being modeled appreciate

and consent to the proposed solution. The creature construction arguments to be worked up here will differ somewhat from those currently in the field, in that they will pay especially close attention to bounded rationality constraints, and in particular to computational feasibility.

We're thinking of a self as a virtual object, overlaid on and augmenting cognitive engines (that is, brains) that the now-dominant tradition in AI modeled as neural networks; we're trying to make design sense of the device. Now design problems have as ingredients not simply objectives but constraints, and these constraints are broadly of two kinds. First there is of course the issue of what's possible or feasible. I hope you'll excuse a trivial for-instance: we use rockets to put satellites into orbit, rather than a space elevator, because we don't have, anyway yet, materials with the tensile strength that would let us build skyhooks. Here, as already intimated, I'll mostly be attending to the limitations that get grouped under the heading of bounded rationality: that we humans can only think so quickly, can only remember so much, can only connect the dots so thoroughly, and so on.

But second, there are constraints generated by further problems—problems that need to be solved jointly with the problem you're engaging. Another trivial example: nowadays, gasoline-powered vehicles take unleaded gas, because the problem of cutting back on pollution in urban environments was added into the mix of the automotive design problem. You might think that adding additional problems makes coming up with a solution to your first problem all that much harder, and to be sure, sometimes it does. But sometimes, and this may come as a surprise, connecting your problem to other problem spaces gives it more structure and so turns out to make it more tractable. (That is, it makes a well-defined problem out of a *mess*.) I'll suggest that that's the way it's going to play out, when we're making sense of the design of the self.

To spell this out, we'll be introducing a handful of questions that are already on the philosophical agenda, and which will turn out to serve as ingredients of the problem of planning the self. As each lap is completed, I'll pause to take stock, to remind the reader of where we are in the development of the larger argument, and also to flag larger concerns to which we will need to face up, once we've given them a chance to emerge. (What sort of concerns? For instance, and to anticipate, just in case you're already getting nervous: *A self is a virtual overlay? Are you telling me that I'm a data structure? an* **app***?*) In order to facilitate engagement by readers with specialized interests, the chapters taking up the more familiar questions, though not the stocktaking interludes, are meant to be readable as standalones. But I'll have to ask for extra patience from those specialized

readers, as I fill in background they won't need, but which will be there to keep other readers on board.

What's more, by way of allowing the big-picture problem to fall into place piece by piece, I'll do two passes over the problem of first-person authority that I'm about to use to launch the discussion: this is the puzzle of how it is that you know (you *just* know) what you think, what you're in the middle of doing, and so on. The first pass will map out the way this problem is tied to the design of the self, to problems in philosophy of logic, and to issues in recent philosophy of mind; the second will introduce the creature construction methodology I mentioned a couple of steps back, and show how it can make headway on the class of exercise we're attempting to complete.

1.3

At this point, we can see how there is a forward-facing side to the project. Suppose that selves are, as I've started to suggest, virtual overlays on us human beings. As I suggested a moment back, that is to say that they are devices (which is not to say that *you* are a device—as we'll in due course see, there's more to you than your self). But no device is suitable for all tasks, or for any given task in all circumstances. Since I've mentioned cars that run on unleaded gas, that example will illustrate this point, too: they require pumps that deliver unleaded, which were not provided in the 1960s; the battery powered cars that are going to replace vehicles with internal combustion engines similarly require networks of charging stations; cars of both kinds do their job only once there is a road network. All of that means that a current-model automobile would be useless without those gas stations or charging stations, roads and so on. And while we're at it, remember Tom Waits, in a track titled "Step Right Up," amusingly lauding a single consumer product that, impossibly, does *everything*: it chops, dices, slices, mows your lawn, picks up the kids from school, entertains visiting relatives, turns a sandwich into a banquet, walks your dog, gets rid of your gambling debts, and much, much more. But, and to drive home the point, this is why the song is funny: nothing *can* do everything, and if nothing, not selves, either.

That is to say that if the demands our environment makes on us have changed, the self may need to be rethought, and perhaps reimplemented. When you look at our intellectual ergonomics, you see a long and impressive track record. However, the uses we make of our inferential accessories are not

well understood. The particular task I'll be making the focus of the second stretch of our argument is an aspect of the functionality of selves that has always been something of a mystery—despite which, we were hitherto able to, as the idiom goes, wing it. I'm going to argue that we can no longer ignore what we don't understand, because there is reason to think that our capabilities are in the course of being undercut by an emerging threat, namely, the division of intellectual and evaluative labor.

Let's consider these closely connected issues in turn. First I have to introduce that older task, and in doing so, complicate somewhat the simplified picture of selves that I've just given you.

1.4

The schematic illustrations of consistency requirements we've given so far have had the flavor of deductive logic. But both in day-to-day life and in the sciences, by far most of the consistency regimes we impose are *defeasible*: they apply only *ceteris paribus* ('other things equal,' or pending a 'defeater'). For instance, if someone has a ticket for the plane to Bologna, and then for the train to Reggio Emilia, they'll be in Reggio Emilia by evening; and you do have both tickets. Keeping yourself consistent, you're perfectly in order when you conclude you'll be there by evening. But then there are all sorts of glitches you could encounter: turning up at the airport having forgotten your identity documents, or a conductors' strike, or a mechanical problem with the aircraft, or construction on the tracks, or an airspace closure, or a shuttle bus breaking down on the highway, or a medical emergency, or taking the airport connector to the wrong terminal and missing your flight. Any one of these might make you back off that conclusion (and also a pending additional conclusion, that you can plan to meet someone for dinner); if one of these defeaters is live, not drawing the conclusion does not make you inconsistent, after all.

So the largest part of consistency management involves responsiveness to *soft* constraints, and in an especially problematic way: if you try listing the things that might go wrong in such an inference, you will find that no matter how many of them you have thought of, you can always come up with one more. It is a recurring complaint in the several literatures that engage this problem that this sort of soft constraint is unenforceable: how can you insist that someone is inconsistent, and has to revise his position or plans, when he has an open-ended escape clause? If selves and other devices in this metaphysical family are scopes for consistency regimes, how have we

been able to operate them effectively?

We can't retreat to deductive or even probabilistic constraints; treating consistency requirements and the inferences they underwrite as defeasible has a very practical point. In large part because human beings are boundedly rational, they overlook things. (Of course, not only for that reason: besides what we know about but don't manage to keep track of, at any time, there is also a great deal we simply haven't run into or observed at all.) Defeasibility is making room in advance to accommodate that fact, and because we can't generally anticipate what we overlook (not even enough, realistically, to estimate how probable it is that we've overlooked this or that type of issue), we allow for flexible and improvisational adjustments, which humans are assumed to come already equipped to produce *somehow*. However, these consistency-management techniques, together with the virtual objects they employ, date to an older time in human history, when the world that was visible to people was smaller, less complex, and conceptually less demanding. In that apparently simpler world, it was reasonable to rely on unexamined native capacities for flexible improvisation.

In order to correctly characterize the administration of the virtual objects we are considering, we have to explain the management of defeasible consistency requirements. Previous work in this area has for the most part attempted to extend the familiar model-theoretic approach in formal logic into nonmonotonic logics; here, however, we are taking a design approach, and so the orienting question is going to be what defeasibility is *for*, and what methods will get that job done. Because the ignorance we are managing is also metaignorance—you not only overlook things, but also overlook the sorts of things you might be overlooking—the general problem is hard, and one way to make progress is to zoom in on subdomains in which there is additional structure to exploit.

1.5

The world we live in makes up a vastly complicated environment, and one of the techniques we have collectively adopted to cope with it has had as a side effect making it still more complicated. Division of labor was no doubt originally driven by the sort of manufacturing efficiencies that Adam Smith was so enthusiastic about, but soon enough it also became a way of dealing with our own bounded rationality, by parcelling up the impossible task of knowing everything into manageably sized packages; division of labor has turned into division of intellectual and evaluative labor. The upshot is that

we now live in an even more complicated and *much* more cluttered world than that already too-complicated past. Specialization has brought it about that there is much more to be known: both 'knowledge-that,' i.e., of facts, and 'knowledge-how,' i.e., skills. There are times when the techniques you employ to address a problem somehow make the problem much harder than it was; this is one of those times.

Part of mastering an area of expertise is coming to control a vocabulary that those outside the field do not understand, a range of facts that you keep at your fingertips, and techniques for bringing those facts to bear; part is learning to use discipline-specific assessment tools that outsiders don't know how to employ. Thus the flip side of division of intellectual labor is massive ignorance, of most of what it's not your job to know; the flip side of division of evaluative labor is massive evaluative incompetence, with respect to almost all those things it's not your job to assess. That means that, in the world we have remade, when someone is trying to figure something out, there is *much* more for him to overlook, and as specialization progresses, we can only expect the problem to become more pressing.

If in the past it was reasonable to rely on those unexamined native capacities as support for the inferential techniques underwritten by our metaphysics, it is reasonable no longer. And now we can look ahead to the forward-facing side of the investigation on which we are embarking; the first and most straightforward aspect of it will evidently be scaffolding consistency management and the reasoning it drives, by flagging and vetting out-of-field defeaters for inferences-in-progress. Reassuringly, this is a more restricted task than the general problem of defeasibility management, in which a consistency requirement can be upended by a fact or a priority that no one knows about, or has ever known; in the special sort of ignorance generated by the division of intellectual and evaluative labor, what you don't know about, because it's not your area of expertise, someone else, in a differently demarcated field, does—the information is available, but elsewhere. Thus the form that specialization-driven defeasibility takes should be more tractable than defeasibility in the wild: *someone* out there has the relevant information, even if it's not you.

If selves are virtual objects, introduced to facilitate inference, they are not—as old-school metaphysics has implicitly supposed—peculiar objects whose features are what they are, prior to and regardless of human intervention, and that somehow come attached from the get-go to each and every human being. (More generally, the objects of metaphysics are not permanent, unchangeable fixtures of the universe.) Intellectual ergonomics can be improved, and intellectual accessories can be rethought. To be sure,

doing so successfully has as a precondition the characterization of extant virtual objects and the procedures that embed them; it's a familiar observation that our conceptual scheme or our 'caveman metaphysics'—as it formerly was phrased—is the product of long use and implicit debugging. How we do things now is a valuable starting point, not to be casually discarded. But once those rational reconstructions are in place, we can consider what sorts of modification are in order. For instance, and we'll focus on this issue in chs. 2 and 3, if the self is a virtual construct, one consisting of a scope specification paired with a repertoire of inferential consistency regimes, we can ask whether broadening or narrowing those scopes would improve intellectual performance. Just to give you a sense of the sort of thing I have in mind, if those consistency requirements had better be ones that humans are able to comply with, and if they're imposed only on what is, putting it colloquially, in someone's head, living up to them had also better be, still colloquially, the sort of thing you can do in your head. Suppose we broaden the scope of those consistency demands to cover persons taken together with their personal electronic devices. Could data in the digital penumbra be algorithmically vetted for compliance with more demanding consistency requirements?

Here our task is preparatory. This book is primarily devoted to what I framed as the backward-facing component of a larger enterprise, where I mean to be offering the rational reconstructions that I've been insisting we need. But just as with prototyping other varieties of software, or for that matter engineering of almost any kind, redesigning the virtual accessories we deploy ought to interleave experimentation at almost every stage, in order to see what really works and what doesn't. Moreover, that redesign ought to be conducted in view of the way that modern AI relaxes the bounded rationality constraints that I am suggesting shaped the design we have. Attention, and the ability to notice what needs to be noticed, has always been a scarce resource and limiting factor, but machine learning and generative AI make attention—or, if you insist, a mechanical surrogate for it—more plentiful and less expensive. Moreover, large language models are knowledgeable in ways no human could be. To be clear, however, even with advances in big-data connectionist computation, rationality remains bounded—just differently. However, as I write, the technology is moving rapidly, and it's too early to tell what sorts of cognitive and inferential scaffolding it will enable.

1.6

Let's look ahead at the order of proceedings. First off, you can't be called out on your inconsistencies—not usefully, not so that you can rectify them—if you don't know what you think. And so it turns out that knowing what you think gives us the leverage we want, in trying to account for the reach or extent of the self.

Not just any version of that will do. There's knowing what you think, in complicated and painfully protracted ways; one of Woody Allen's characters, complaining about his decade-long psychoanalysis, announces that he'll give the analyst just one more year. (If there's been no progress, he announces... he'll find another analyst.) And then there's knowing what you think, just like that. The instant, just-like-that version has in recent decades become its own subfield within philosophy, with an ongoing professional literature, variously labeled "self-knowledge," "first-person authority," or "privileged access". A growing number of philosophers seem to have decided that the topic is the key to many mysteries, and for once, they're right—but not, I'll argue, in quite the way they imagine. We'll start off, in ch. 2, by asking what to make of first-person authority.

That discussion will carry us over into the following chapter as well, because it will quickly take us to questions about methodology and logic. If you are going to rely on your self-knowledge to extirpate inconsistencies, *which* inconsistencies should you be prioritizing? (Or rather, which inconsistencies should we be insisting that you prioritize?) Since, as we've already noticed, inferences are driven by consistency requirements, that is more or less the same as to ask: what inferences are we going to demand that people make? And so ch. 3 will introduce the methodological frame that I'll be using to tackle this sort of question: here, that will be to ask, from the point of view of someone designing a reasoner, what patterns of inference it would make sense to ensure end up in the repertoire of his creation. That won't just give us a wish list. Designers have to take into account the limitations of the materials they have to work with, along with the environment in which they intend whatever it is they're making to operate. And they have to prioritize, because there will be tradeoffs. What demands have to be met? Which are most urgent? Most important? And where can they allow themselves to cut corners?

At this point, I'll pause for our first interlude. We'll do some stock-taking, but also step back to register and address apprehensions I expect to have been building up in the back of a reader's mind. What is the status of the argument I'll have been building? Can we really be on our way to an

understanding of *logic*? Of the *self*? At this stage in developing the view, I'll be asking my readers to suspend disbelief in large quantities, and by acknowledging the size of the ask, while also providing further motivation for it, I hope to go some way toward keeping you on board.

Once we have that picture of the self in place, together with the accompanying framing of questions of logic and methodology, we will need to return to those soft consistency conditions: that is, to the problems they pose for administering inferential requirements, and for functioning as an effective self. As you by this point have no doubt realized, I think of those last two problems as, really, just one problem; managing an effective self, in the vocabulary I'm adopting, is a matter of successfully administering inferential requirements.

The predicament, of realizing you might suddenly notice something, and have to take back an otherwise perfectly legitimate conclusion, is real and pervasive in human life. And accordingly it crops up under different names, not only in one after another philosophical discussion, but elsewhere as well. Philosophers of science argue amongst themselves about how to treat ceteris paribus laws, that is, laws that tell you what happens "other things equal". Legal theorists worry about equity: how do we make space for deciding that a law, as written or hitherto interpreted, doesn't make sense in *this* case? (And how do we do that compatibly with the very basic need, in legal contexts, to arrive at decisions that settle a dispute and allow those involved to move on?) Computer scientists and logicians try to model what they call nonmonotonic reasoning; some philosophers reporting on their efforts call it default reasoning. All of these are slightly differently accented ways of speaking about one and the same phenomenon, and moral philosophers, especially those influenced directly or indirectly by Aristotle, tend to talk (as I've started to) about *defeasibility*.

As announced, the main chapters of this book are meant to be read as standalones, and so I'll reintroduce that core concept each time through. Please don't be impatient; doing that will be my excuse for putting a longish series of illustrations on your plate. Those examples serve a function in the overall argument, that of providing evidence—anecdotal, but I mean it to be convincing nonetheless—that defeasibility is a feature of human life and of our reasoning that you can't sidestep, and so one that merits the attention we're going to give it here.

Now, in real life, as opposed to logic classes, the consistency demands that selves are asked to live up to are, almost all of them, defeasible (that is, other-things-equal) requirements, and so we need to understand what they're doing there. Accordingly, the next two chapters are devoted to making sense

of those demands as a feature of our design, and as part of that, making sense of that deepest and most puzzling logical feature of defeasibility.

If you are considering one of those soft-consistency trains of reasoning, and you start making a list of issues you might plausibly need to doublecheck before you go ahead, you'll soon enough realize that your list doesn't come to an end on its own. No matter how many potential problems you've already jotted down, you can always—anyway, if you feel like it and are even moderately inventive—come up with one more. On the one hand, we're told, avoid paralysis by analysis; on the other, look before you leap. How do we decide when enough is enough? As we will see, we have native but visibly makeshift strategies for handling these situations. One of them is vagueness, a topic much discussed by metaphysicians and philosophers of language in recent decades. Another is to appeal to likelihoods—and here it will be useful and important to see how it is that probability is a poor conceptual fit for the phenomenon we're trying to understand. So we'll work through both of these topics in the course of those two chapters.

At that stage we'll pause once again, this time to distinguish the methods we'll have seen in action from neighboring approaches. The second Interlude will juxtapose the questions we're asking with those pressed by other research programs: evolutionary psychology, meme theory, what is currently the standard way of framing logic, and conceptual engineering. To make the contrasts tangible, I'll take some time to step through a couple of substantial illustrations. So we'll see what Newcomb's Problem, peer disagreement, and one of Sydney Shoemaker's explanations of first-person authority look like to a bounded-rationality theorist.

Once the full difficulty of explaining how we handle soft consistency is appreciated, even moderate success at it can start to seem like magic. But nothing is magic. The following chapters take up two very different ways in which our competence with defeasibility has been underwritten. Part of the story is selection effects in the social world shaping the institutions, practices and technologies that we interact with, so as to make them in this particular respect more user friendly. And part of it is the way we have taken the selves that, for all I have said so far, might as well be momentary, or anyway, quite short lived, and built them out into persons, who can be around for, these days, as long as a century. Thus a nonstandard slant on the traditional philosophical problem of personal identity—of what makes a person now the same person he will be later—drops out of the account we're embarking on; temporally extended persons have been an improvisation that was needed to make defeasible inferences a workable technique for facing an often obscured world.

And at that point, we'll be positioned to set that forward-facing agenda. The stakes will be clear, and they are high. Our selfhood, our personhood, and our ability to draw conclusions reliably, in the course of figuring things out: all of this is threatened by, ironically, our success in turning ourselves into the highly specialized experts that together can run a technologically and economically advanced society. I'll make some suggestions as to the direction that research ought to take, if we want to look out for our selves. We'll also be positioned to look back at the philosophical register in which the argument has been conducted, so as to consider the questions it will have raised: Can we say more about the appropriateness of creature construction arguments for problems of this kind? Does our account of selfhood allow us to adopt a new perspective on what it is to philosophize?

Now, over the course of this preview, the reader may have had a growing sense that I'm trying to do much too much for a single book. After all, if a philosopher has an account (and especially, a revisionist account) of personal identity, isn't that a book on its own? If he has an account of vagueness (and especially, one that takes issue with the recent literature on the subject), surely *that's* a book. And likewise, for first-person authority, for how we should argue about logic, and for a view about what the objects metaphysicians investigate really are—again, especially since it is one that a majority of metaphysicians will resist. How can we expect all this to fit into what you're about to read right here?

In a short volume that is gradually coming to be recognized as a philosophical classic, Elizabeth Anscombe drew a picture of actions as nested or embedded one inside the other, such that a seemingly narrower action *is*, once we attend to a bit more of the surrounding context, another and broader action, which is, in turn, once still further context is pulled into view, a yet broader action... To give you a for-instance, I'm pointing at the contents of a glass-covered, climate-controlled display cabinet. Pulling back one step, that just is my selecting the contents of a box of chocolates. Pulling back one step more, that is picking up a house gift. One more step back, and it's taking care of a social obligation, that of bringing something to a dinner to which one is invited. So far, we have four distinctly enumerated actions, but right now, what you see is just me pointing out this and that praline to the counter clerk.

The agenda I've been previewing is tractable because it is structured like *that*. The objectives are nested: a higher-level objective is executed in that an objective one level down is executed. *By* characterizing defeasibility management from a design stance, we will be characterizing consistency

regimes that drive inference and intellectual innovation. *By* characterizing these consistency regimes, we will be providing design-stance analyses of the self and the person—that is, of some of the central objects of traditional metaphysics. *By* providing these design-stance analyses, we will be making concrete and validating the intellectual ergonomics construal of and approach to metaphysics.

And looking ahead, to that forward-facing agenda: *by* validating that approach, we will be underwriting experimentation with readjusting and scaffolding the objects of traditional metaphysics, to take advantage of the novel possibilities offered by modern AI. And *by* performing that experimentation, we will be putting people who have to live in our highly specialized world—all of us!—in a better position to figure things out.

1.7 A note on notes

Footnotes are where I'll be putting material that lies off the main line of argument. Some of that will just be pointers to the literature: citations and the like. Occasionally it will be necessary to engage an author, where inserting the back-and-forth into the text would be distracting. Some notes are for handing objections that might or might not occur to a reader, where addressing them in the text would make it harder to follow the flow of the argument. It's our stylistic practice, in the world of English-speaking philosophy, to front-load qualifications to claims, by way of anticipating objections, and where that would interupt the flow of the argument—or just turn out to be tedious—sometimes I've pushed them down into the notes. And finally, some of them will be asides that I think are interesting enough to make optional digressions. The footnotes are there, and I hope you enjoy them, but don't worry that they're where the argument is happening.

Chapter 2

Outsourcing the Mind

Over the years, a good deal of ink was spilled over what looks like an odd-ball puzzle in epistemology, the problem of first-person authority or self-knowledge. Now that ink is out of fashion, toner cartridges are getting emptied over so-called extended or wide cognition. I want to bring these two topics to bear on each other, so as to explain why first-person authority is not *just* a puzzle, and why the problem posed by extended cognition is not merely a pseudo-problem. First-person authority, I am going to claim, amounts to a constraint on how one addresses the reach or extent of the mind or self, and the problem of setting the limits of the self turns out to be an engineering problem—and the flip side of some of the deepest questions in logic.

It's often helpful for philosophers to have a new proposal anchored to familiar topics or problems, and that's my reason for starting off with what may seem to outsiders to be exotic questions. If you're not already immersed in the back and forth, however, I hope I can make good on the promise that the questions will prove interesting, and I'll ask you to indulge me while I mark connections to other discussions-in-progress. If you *are* already a player, let me ask for forebearance as I get everyone up to speed, and forgive me if I don't tax other readers with the details of professional infighting.

I'll begin by rehearsing the puzzles, and why they can so easily prompt the dismissive response that they are no more than puzzles.

2.1

The problem of first-person authority is that of explaining how it is that, normally, at any rate, you know what you believe, what you feel, what you're in the middle of doing, what you want, and so on; and further, how it is that you seem to know what you think, want and so on in a way that's different than other people do. If you want to know what Roger Wade is up to, you have to hire Philip Marlowe to follow him around; but if you want to know what *you're* up to, you don't have to hire anyone—you *just know*. Philosophers have differed among themselves about whether you can be mistaken about such matters; back before 'Cartesianism' became a dirty word, your opinions about your own beliefs, sensations, and so on were held to be incorrigible; these days, while you can still find philosophers insisting that you simply couldn't be wrong about whether you have an itch, most of them allow that people can be mistaken about most of their other attitudes. The vain, and squeamish bigots, and other self-deceivers don't think they have the beliefs and other attitudes that they do.

Incorrigible or not, there does seem to be a substantial body of self-knowledge, knowledge that we come by apparently without engaging in observation, and it's easy to see how accounting for it could make up an obligation that it would be incumbent on an epistemological theory to discharge. (Given that knowledge is so-and-so, and has to be acquired in such and such a way, what about your views about your own mental life? How could they amount to knowledge?) Or again, there are theories of the mind on which it is hard to see how anyone could know, without observation, what is in his: old-fashioned logical behaviorism, on which a mental state just is a disposition to certain sorts of outward behavior, or the view Wittgenstein seems to have been exploring towards the end of his life, on which having a mind (or being a person) is conceptually on a par with being a smiley or a Necker cube (others *see you as* a person, minded, and so on). Or again, externalism in philosophy of language—on which the contents of your thoughts depend on facts outside the mind, possibly distant facts you may not be in a position to observe—raises the question of how you could know what your thoughts are about. And there are still further research programs that similarly need something to say about self-knowledge.[1]

[1]For an introductory survey of the field, see Gertler, 2003; Cassam, 1994, is a somewhat dated overview focusing on the problems posed by externalism. For a sophisticated behaviorist worrying about the problem, see Ryle, 1967, ch. 6.

The claim that we *do* have the sort of self-knowledge that we're interested in has to be qualified, and has been in various ways denied, by, for instance, Lear, 1990, and as Robert

When it's looked at in one of these ways, however, first-person authority is only of interest to someone in the middle of constructing a theory of knowledge, or of the mind, or a theory of reference, or what have you, and the primary responsibility of a solution is to the problem first-person authority poses for that background theory. If you don't happen to be committed to the background theory, you don't need to be interested in the account of self-knowledge that it motivates. First-person authority appears to be a puzzle *for* one local philosophical theory or another, and thus a puzzle that is optional to the extent that that theory is.

Let's now turn to wide cognition, the idea that if a mind is a computational system, then there's no principled reason for the borders of your mind to follow the contours of your skull: after all, if your datebook and calculator figure into the computations that produce your beliefs and your actions, why aren't they parts of your mind as well... in as literal a sense as you like?[2] Philosophers who discuss extended cognition worry about what they call the problem of cognitive ooze: after all, if your datebook can be part of your mind, why not all of the internet? So, at first glance, the question that extended cognition poses for us is: what counts as part of a mind? Where are its edges?

At second glance, however, that question is unmotivated and not well defined—so far, not even a puzzle. Consider an analogous set of questions, about the borders of the US economy: Are the Chinese subsidiaries of American corporations part of the US economy? How about their Chinese suppliers? And isn't there are problem of economic ooze? Why isn't the entire world part of the US economy? Questions like these can sound urgent, but the urgency is phony; to do a sensible job of asking what the boundaries of the US economy are, we first need to know what the answer is *for*. (Perhaps a better and prior question would be, What is wrong with hollow economies? Thinking about *that* question might tell us why, exactly, we want answers to the questions we first introduced.[3]) No one has told us what we are asking about wide cognition for, and until we know, the urgent sound that questions about extended cognition can take on is just as phony. Really, why

Schwartz reminded me, it's part of the stereotype that Jewish mothers don't take one to have first-person authority with respect to one's own sensations. ("Mom, I'm not cold; I don't need a jacket!" "Don't *you* tell *me* you're not cold!")

[2]Clark and Chalmers, 1998; Clark, 2003. Computational (or 'functionalist') theories of mind were introduced by Putnam, 1975b, and Putnam, 1975c. For back-and-forth about extended cognition, see Rupert, 2009, and Adams and Aizawa, 2008.

[3]A very suggestive and historically informed answer to this question is given by Braudel, 1992.

worry about extended cognition? Like first-person authority, it looks to be a philosophically optional concern—or rather, what Carnap long ago called a pseudo-problem.

When you outsource bits and pieces of your mind, do you normally have first-person authority with respect to the bits and pieces? Perhaps not: I know what is in my datebook by looking, and the sort of self-knowledge philosophers have been trying to explain is characteristically introduced by stipulating that you know *without* looking. But, I am about to suggest, offshoring your cognition requires a surrogate for first-person authority (in something like the way that outsourcing your production facilities to Indonesia requires a surrogate for the self-monitoring your corporation conducts at home). And thinking about the role the surrogate plays, and how it constrains what can be outsourced, will allow to us say what philosophical issues lie at the bottom of what turn out not to be merely optional philosophical puzzles.

2.2

Extended cognition is mental offshoring, so let's consider what it takes to make real offshoring work. When a corporation outsources, it has a choice. It can simply buy the product, without intervening in the manufacturing process, in which case it needs only to monitor the quality and cost of the deliverables. Or it can opt to regulate the production processes itself. It is natural to think of the first type of case as purchasing from a supplier, and of the second, as the corporation manufacturing the product itself. If it is regulating the production processes, it has to be able to intervene in them, and in order to intervene, it has to know what is going on.[4] Knowing what is going on means being able to monitor the production process.

Three points are important here. First of all, defaultly, the monitoring must produce outputs that track what is happening on the ground: that is, when the report says that p, then if p were true, that's what it *would* say, and if p *weren't* true, it wouldn't; in other words, the outputs counterfactually covary with the state of affairs on the ground.[5] But there is a complication: notice that the direction of influence need not be from the events being reported to the report. Suppose that an overseas manager tells the home

[4]To be sure, often the point of outsourcing is *not* knowing what is going on: that's what gives you plausible deniability about the absurdly low wages and miserable working conditions of the employees, the environmental impacts, and so on.

[5]This is Robert Nozick's sense of the term (1981, ch. 3).

office that a new policy has been put in place, even though it hasn't yet been adopted. If that commits his branch to following through, and if the policy is then hurriedly implemented, from the point of view of the home office, that report is just as good as a report that is causally posterior to the implementation of the policy. What matters is that the information is usable, not how it gets that way.

Second, if regulation and control are to be feasible, the overhead must be low, which means that information must normally be usable *as is*, without further examination: when a manager back home gets a report, he must normally be able to presume that it *does* tell him what the situation on the ground is. Organizations whose internal reporting systems are so unreliable that reports from the field are only the starting point for an investigation into whether what they convey is true can't effectively manage remote operations.

Third, the corporation's monitoring can take many different forms. Managers may conduct site visits and inspect the operations. They may upload data from terminals via a satellite link. They may receive lengthy, handwritten letters from overseas, delivered by slow boat. (Long ago, the British East India Company relied on this expedient.) And there are indefinitely many other ways that the required monitoring could be implemented.

What these many methods will share is that they are low overhead, in terms of the further attention that must normally be paid to the reliability of the outputs. (Which is not to say that they are always correct.) The monitoring machinery takes care of itself, recedes into the background, and, most of the time, can be ignored. Remote operations—as opposed to purchasing goods and services from a subcontractor—require a surrogate for or analog of first-person authority. The analogy, I am about to suggest, is a guide to what the important features of first-person authority really are.

2.3

Let's now turn to a simple example of extended cognition. Suppose that, instead of keeping track of your appointments by remembering them, you use a datebook. On the wide-cognition way of seeing things, when your datebook is properly integrated into your decision making, it is offshore memory, literally a small piece of your mind that resides outside your skull.

Consider for a moment what proper integration requires, helping ourselves to the corporate analogy for guidance. You have to be able to tell what your datebook says, pretty much just by looking, or it's not going to work for you: using your datebook has to be low-overhead, and in partic-

ular, questions of the reliability of its contents must—normally—not come up. For the datebook to function as prosthetic memory, you have to be able to control what goes into it: if *you're* not the one entering and deleting the appointments, you have, not a datebook, but an engagement calendar used to give you your marching orders. So you have to be able to take it pretty much for granted that you're responsible for the entries.[6] Briefly, the contents of the datebook have to track your commitments, and the overhead involved in accessing the information has to be low. You don't have to puzzle over whether the entry ("4:30: meet Jack") commits you to meeting Jack at 4:30; you look at what's written, and you *just know* that you're planning to meet Jack.[7]

We shouldn't expect there to be only one explanation for how these demands are met. Your entries may track your commitments because, after you have formed the commitments, you systematically enter them in the book; or they may do so because, as you write an entry into your datebook, it becomes your commitment. You may be able to take it for granted that only you have modified your datebook because it's obvious to you when an entry is in your own handwriting, or, alternatively, because you have ensured that only you have access to your datebook, or, again alternatively, because you can implicitly rely on others who do have access (these days, system administrators) not to tamper with it. And we should not expect there to be any metaphysically mysterious explanation of your 'datebook authority' in the offing. The reason is that your low-overhead, reliable-enough access to the contents of your datebook is a *requirement* that must be met, if your datebook is to be considered an external bit of your mind, and there are indefinitely many ways to meet such a requirement. As so often in philosophy, it is confusing a requirement with the fact that it is typically met that makes an old-school metaphysical explanation seem appropriate.

This point is important enough to deserve a second illustration. Drivers are required to stay under the speed limit, and so they're required to know how fast they're driving. (Ignorance is no excuse.) Because driving involves the real-time control of dangerous machinery, the requirement has

[6]That matters mostly for whether you regard the datebook as *memory*; with appropriate social practices in place, the online engagement calendar may still count, for other purposes, as part of your mind.

[7]For a discussion of the claim that you are entitled to accept some claims without observation or inference, see Burge, 1993; it should be clear that while I share Burge's view that there are such claims, and that the fact is philosophically important, I do not accept his substantive claims as to what underwrites the contrast, and what we are defaultly entitled to accept.

to be satisfied without diverting cognitive resources from other important tasks; it should demand scarcely any attention at all. Partly it's managed by having working speedometers (checked during the annual safety inspection). Partly it's managed by subliminal awareness of the flow of traffic, and cues such as textural gradients, splay, and optical flow.[8] There are the radar signs placed by traffic enforcement. A driver can know that he was under the limit because he was making a point of driving slowly; here, the default relation between the knowledge and its object is being reversed. All in all, while you may insist to the policeman that you *just know* that you were definitely under the speed limit, there is no metaphysically mysterious fact of 'driving-speed authority' that requires the sorts of explanations we see bruited about for first-person authority. And the reason is that what we have is, in the first place, not a fact but a demand, and a demand that can be met in any one of indefinitely many ways.

It's time to return to first-person authority, and the hypothesis we're now placed to entertain is that it, too, is in the first place not a fact but a demand. If that is correct, then there is no reason that demand cannot be met in any one of indefinitely many ways. Perhaps it can be met, some of the time, by the low-cost expedient of learning that, when you are disposed to assert p, you may prefix "I believe that" to your utterance.[9] Perhaps some of it is handled by cognitive machinery that might as well be a mental camera pointed at the inner workings of your brain—in the jargon, 'detectivism' (more or less, that you literally introspect your thoughts)—might be the correct account of *some* first-person authority.[10] Some of it might be met by committing yourself to followup; once you say that you believe such and such, you act, until further notice, in ways that constitute believing such and such.[11] Some of it might be met by reassessing the object of the attitude about which you're being queried; asked if you believe it is about to rain, you do a quick assay of the evidence for and against its being about to rain.[12] And some of it might be a matter of blurting out self-descriptions that are in the first place expressive behavior, and continuous with behavior such as laughing and crying.[13]

The point is that if first-person authority *is* after all a fact, it is a fact because a demand is being met (or met often enough), and the explanation

[8] Wickens and Hollands, 2000, pp. 160f.
[9] Shoemaker, 1996, p. 36.
[10] For pointers to instances of detectivism, see ch. 3, n. 19.
[11] McGeer, 1996.
[12] Moran, 2001, following Evans, 1982, pp. 225f.
[13] Finkelstein, 2003, Bar-On, 2005.

worth going after is of the *demand*.[14] I'll take up the question of just what the demand is and why it is being made in a moment, but before I do, notice that if I am right, the philosophical project of constructing a unified, uniform, old-school metaphysical theory of the fact (or, the 'fact') that we have first-person authority with respect to our beliefs, desires, sensations, and so on is misguided. Most work in the field tries to explain *the* way that first-person authority works, but we should expect implementation strategies to vary. The terms of the problem that the standard treatments have taken as their frame are mistaken.[15]

Now, you might also be inclined to draw the further conclusion that the implementation does not matter at all for philosophical purposes; as we will see, that would be to overread the point I am making. *How* first-person authority is managed does not really matter, but engineering limits to how *much* of it is managed will turn out to matter a great deal.

2.4

Inference is a prescriptive notion; it's about what you *are to do* when you're reasoning your way through a question; it is something that you can do right or wrong, and if you do it too badly, we stop calling it inference at all. Now, there is normally not a lot of point in imposing demands on people who cannot comply with them. (Of course there are exceptions, as when the point of the ideal is not to be achieved, but to elicit that extra bit of effort; nonetheless, this is the important truth buried in the overstated philosophical commonplace that 'ought' implies 'can'.) Compliance normally requires monitoring and feedback. So inference cannot far outrun the reach of people's ability to monitor it.

The monitoring that allows one to function as an effective inference engine must not normally require much in the way of a certain kind of overhead. In the first place, there is a regress problem. Suppose that you couldn't rely on your off-the-cuff views about what you were thinking: for

[14] David Velleman (1989; 2000) has claimed that agency, and so practical reasoning, is built around the desire to know what one is doing, and this much is right about Velleman's view: knowing what you're doing is very important to us. If I am right, however, it is important not because we *desire* it, but because it's *required*. The desire is present often enough—because people sometimes internalize demands—but amounts to no more than personal idiosyncrasy.

[15] In particular, we do not want to make the error I will discuss in our second Interlude, that of taking an account showing first-person authority to be conceptually entailed by rationality to *preclude* implemenations of it.

instance, suppose that, when your conclusion was challenged on the grounds that it was a non sequitur, you couldn't trust your memory of what you had just inferred it from. Then you would have to arrive at your view of what your premises had been inferentially. And if this further inference is itself challenged, and you can't rely on your impression of what its premises and conclusions had been, you'll need a further argument to establish what they were. . . and so on, up the line. Sooner or later, the regress must terminate in inferential steps of which you are confident, without *further* investigation, that they *were* your steps. There must be opinions you have about your own beliefs, desires, sensations, current course of action and so on that you can rely on implicitly, without further thought.[16]

Leaving regresses to one side, it is clear enough that if you cannot just state what your inferential steps are (or, a moment ago, were), without extensive investigation, you are not in a position to make effective use of the techniques of inference. Just as offshoring works when the corporation can (generally) rely on its monitoring procedures, and offshoring one's mind works when one can (generally) just take it for granted that, for example, what's in the datebook is what one put there, so the traditional sort of thinking inside one's nonextended mind works when you have first-person authority with regard to your thoughts. What the corporation needs overseas, it obviously needs at home: low-overhead information, information that tracks the situation on the ground, and that does not require extensive additional investigation before it is used. The very similar monitoring capabilities that are evidently a requirement on some types of extended cognition likewise have internal analogs: if you are going to engage in inference, you have to have low-overhead representations of (aspects of) your own cognition, representations that you can use, normally, without further examination or investigation. The sort of monitoring that a thinking, reasoning mind requires is just first-person authority—or, putting to one side the mysticism-laden controversy about what is entailed by first-person authority, properly so-called, the sort of low-overhead reliable access that first-person authority shares with its extended-cognition surrogates, and with its analogs in the world of transnational corporate activity.[17]

[16]For readers who are familiar with the notion of "deviant causal chains": you must also quickly arrive at mental transitions about which you are confident, without further investigation, that they are not "deviant." The idea that first-person authority is needed to terminate such regresses is not new; see Boghossian, 1989, p. 10, and the point is probably foreshadowed at Ryle, 1967, pp. 162f.

[17]Keep in mind that low-cost access is not always *fast* access: recall those letters sent by boat, from the colony to the British East India Company headquarters in London. If

2.5

In an intervention in the ongoing debate on extended cognition, Kim Sterelny reminds us that human beings and their immediate ancestors have been cooking for a long time: exactly how long is controversial, but long enough for our teeth and jaws and digestive tract to have ended up significantly smaller than those of our nearest primate relatives. We chop and grind and ferment and bake and boil our food outside our bodies, and if the fact that we write things down in datebooks and use calculators and look things up online is enough to make the case for extended cognition—for the thesis that some of our thinking goes on outside our bodies—then, by parity of reasoning, our culinary practices should support the thesis of extended digestion, the notion that we do a great deal of our digesting outside our bodies. Since that's an unreasonable view, Sterelny concludes, the extended cognition position is unreasonable as well.

And what exactly *is* wrong with the thesis that human digestion is extended, over and above what for some people might be the unappetizing sound of it? The flour mill which grinds your grain, and the bakery which converts it into the loaf of bread you finally eat are not *your* mill and bakery, but a shared resource, supplying perhaps thousands of people. Even most home kitchens are shared resources, serving families rather than individuals. Better to think of them as part of a constructed environment, as scaffolding for the activities they support, but not as an extrusion of *your* digestive system into the environment. Likewise, the university library, the search engine I rely on, the signage I use to navigate when I drive, and the satellite constellation that supports the GPS device I also use to navigate when I drive are shared resources; they are not *my* cognition, extruded into the outer world, but better thought of as aspects of niche construction, as shared scaffolding for the cognitive activities they support: something on a par with the pheremone trails laid down and followed by ant colonies as foraging ants are recruited to the task of retrieving food.[18]

Let's pursue Sterelny's train of thought past the point at which he left it, and consider under what circumstances it makes sense to implement extended cognition, that is, extended cognition on which you *have* a monopoly, as opposed to making available some sort of shared cognitive resource. Starting with the most central and dramatic case, you'll notice that the logical demands we impose are closely tied to the demarcation of persons

---

we allow my datebook to be offshore memory, we should also allow that I might keep it in a safe.

[18] Sterelny, 2010; for the ants, see Hölldobler and Wilson, 2009, pp. 61–63.

and their minds. If I believe that p, and I believe that $\sim p$, I'm exhibiting logical inconsistency, inconsistency for which I'm responsible, first, in that I'm required to drop one or another of those beliefs—this is something I'm supposed to *do*—and second, in that if I don't, I can be *censured*, as logically inconsistent or irrational. But if I believe that p, and *you* believe that $\sim p$, we are merely disagreeing; to be sure, at least one of us must be wrong, but there is, as far as that goes, no *logical* demand, on either of us, to change our minds. If I believe that either p or q, and I believe that $\sim p$, I am logically committed to inferring q (with the option of withdrawing one of those beliefs instead); if I believe that either p or q, and *you* believe that $\sim p$, no one is logically committed to inferring q.[19]

There is a deep question in philosophy of logic in this vicinity: we evidently treat logical and evidential consistency as though it were a matter of local informational hygiene; why don't we simply require that all information in a pool shared by all human beings be consistent?[20] Here's what I take to be the right frame for that question. Let *psychologism* be understood as the claim that, in order to figure out what correct inference consists in, you first need to figure out how minds work. Logical consistency, and more generally consistency of both the theoretical and practical varieties, is a standard with which individuals are to comply. We impose and enforce that standard, to the extent that we do, because it is *feasible* (enough, given how human minds work) to do so; because we cannot make anything like a mind out of humanity as a whole, it would not be feasible to enforce it on the pool of information collected by all humans.[21] Psychologism was no fallacy: because inferential demands are imposed only within the minds of persons (or on entities that, like some institutions, are structured to mimic persons), what inferential demands can reasonably be imposed depends on what those minds are like. This means that the following very practical design questions are two sides of the same coin: what kind of minds are we to have, and what logical requirements shall we take on?

A monopoly on cognitive resources—resources that, instead of being shared, are entirely yours—supports compliance with just the sort of stan-

[19] Millgram, 2009a, p. 71.

[20] Bernard Williams thought of information as a shared resource in this way; it is to be added to the pool only if it is true (2002). Since its truth would guarantee its consistency, users drawing on the pool should not have to take additional steps, steps for which they are held responsible, to assure that the small bodies of information they have copied from the shared pool are internally consistent.

[21] And why *can't* we function as that giant mind? See Cherniak, 1986, pp. 128f, for an elegant argument; I recap the idea below, in ch. 7, n. 8.

dard we've started to consider.[22] Useful cognitive resources typically need ongoing maintenance; quite often these resources amount to a system, and the sort of maintanence I have in mind can be characterized as assuring that the system is in compliance with a set of global constraints. For instance, my spending decisions shouldn't have me exceeding an overall budget (my financial resource allocation scheme); my schedule for the day has to be practically coherent, among other ways in not requiring me to be in two places at the same time; my theory of what's going on around me should be internally consistent and fit the evidence. So maintaining the cognitive resources in question means making local decisions that adjust one element or another of the system in a way that maintains its compliance—or brings it into compliance, or closer to compliance—with the global constraints. I decide not to buy a new pizza peel because it would put me over budget, or I stick to my budget by deleting some other purchase that costs more than the peel.

For the sorts of cognitive resources of interest just now, whether a local adjustment to the system brings it into (or closer to) compliance with the global constraints depends on the state of the rest of the system. For the case of external cognitive resources, this means that assuring compliance in the course of making a local change depends on whether other local changes are being made elsewhere in the system by other agents. (If your spouse is making spending decisions off where you can't see him or her, it's hard to know whether what you're about to buy will put the family over budget or not.) Allowing agents to access and alter local elements of a system independently, when it's required to obey this sort of global constraint, creates a resource management nightmare. A natural solution is to assign the resource to a sole administrator, who is the only agent empowered to make changes to it.

In the version of the solution I have in mind, the administrator is assigned *responsibility*, where that minimally entails that it treats the global constraint on the system as a prescription, that the administrator responds to the prescription by making local adjustments with a view to satisfying the global constraint, and that it is subject to sanctions if it fails to do so.[23]

[22] Of course, there are other reasons for privatizing cognitive resources; most obviously, sometimes people have different needs. For instance, if I'm spending tomorrow in Los Altos, I need a map of the South Bay, whereas you need a map of Salt Lake City. And people might need resources that are differently configured; for instance, even if a navigation app is a shared resource, the turn-by-turn directions it generates for me have to be different than the turn-by-turn directions it generates for you.

[23] So this is a much more full bodied sense of responsibility than that in which a pro-

Logical consistency is a global constraint of just this sort. Whether adding a proposition to a system of propositions maintains its consistency depends on what the other propositions in the system are. If I'm considering adding some proposition to a system of propositions, I can't do much in the way assuring consistency if someone else, elsewhere, is also adding and deleting propositions while I'm not looking. So systems of propositions are maintained by owners (the people who, we say, *believe* them), and who are responsible for keeping them in compliance with the requirements of logical consistency.[24]

I introduced the generic design solution as creating administrative monopolies over cognitive resources, and I said that as though the administrators already existed. But a self is brought into being by creating the monopoly: we are actually considering the metaphysics of selves, who to a first approximation are these administrative devices. The device is, roughly, a particular way that cognitive resources can be privatized, to address what

---

grammer might say that a particular subroutine is responsible for executing this or that operation: you don't sanction the subroutine if it screws up. This in turn means that I have to correct what might perhaps be a misimpression, to the effect that each and every account of first-person authority in the literature fits neatly under my own discussion of it. Davidson, 2001, explains an agent's awareness of his own mental states as a requirement on the correct interpretation of that agent. Here the locus of the requirement has been shifted away from what I was just calling the administrator: the onus of ensuring that you know what you are thinking, etc., falls on your interpreters, rather than on *you*.

To say that sole-use resources are a natural implementation solution isn't to say that they're the *only* solution. Wikipedia maintains a large body of information that any user can make changes to at any time, and those changes have tended to improve the overall consistency of that body of information.

[24]If you were interested in extended cognition in nonhuman organisms, you would evidently have to look at packages that didn't include overt awareness of the global constraints or sanctions consisting in the assignment of, as Robert Brandom would say it, 'normative statuses'. Perhaps we can get a sense of what such a privatized resource might look like by returning to the pheremone trails of ant colonies—adding on this time through the notion promoted by Hölldobler and Wilson, that we should think of them as "superorganisms," i.e., individuals. The pheremone trails leading to food sources, new nesting sites and so on are external to the location of the colony. They typically carry information used in allocating a colony's labor resources: as the trail is built up by recruited foragers, it becomes stronger and attracts more foragers; as foragers return empty-handed, the intensity of the trail signal weakens. And such cognitive resources are likely to be privatized, in the sense that they are controlled, modified and available only to the one colony. (Qualification: when the colony is monogynous than polygynous: many-queens colonies don't seem to be organized into individual 'superorganisms' in the same way.) Chemical signals are species- and often colony-specific; moreover, trails are more likely to be laid within territory from which conspecific competitors are excluded (2009, pp. 78, 177, 182–185, 202–205, 209f, 215–217, 230, 249f, 262, 271, 277, 279, 282).

should strike computer scientists as a database management issue: the route to understanding first selfhood, and eventually persons, proceeds through a hitherto unheralded division of our field, namely, the philosophy of filing.

Extended cognition is evidently a revisionist version of such a monopoly, one that makes sense when there is good reason to extend the zone of exclusive responsibility for the consistency (logical or otherwise) of a system of cognitive resources. And we are finally in a position to explain the philosophical importance of first-person authority: since such a monopoly will fail to maintain the consistency of the cognitive resource unless it possesses the sort of low-cost, reliable access to the state of the system that we have already characterized, first-person authority plays the role of a feasibility constraint on the design.

2.6

Since what we're talking about is implementing design solutions, the interesting question is not whether there is (*already* is) a lot of extended inference, but whether it's a good idea to require it: whether and why it *pays* to assign 'external' inferential resources to individuals, in roughly the way we assign 'internal' resources to individuals, and impose on those individuals the requirements of keeping them in compliance with logical and more generally inferential standards. (But notice that the design solution I'm gesturing at is a package; you could also ask whether it pays to implement other packages which include only some of its components.) The most important payoff of an extended mind is in my view that it might allow us to legislate and enforce more demanding forms of consistency: that is, in the central case, to impose new and more rigorous logical requirements. The obvious candidate for an illustration is Bayesianism, but let me give a another, lower-key, but still somewhat speculative example of the sort of thing I have in mind.[25]

[25]Spelling out that obvious candidate: a recently popular view is that all reasoning about matters of fact should proceed by assigning probabilities to them, and updating the probabilities in accord with Bayes's Theorem (e.g., Earman, 1992). Bayesians' enthusiasm notwithstanding, we do not normally engage in Bayesian inference. Because we are responsible—because we are *held* responsible—for our inferences, inference is a designation we reserve for cognition that we can monitor, and endorse or retract. But Bayesian inference requires belief-like states ('credences') with strengths taking precise numerical values in the interval [0,1]. Even if our beliefs do have precise strengths, we do not know what they are; because we can't give reasonably precise numerical estimates of the strengths of our beliefs, we are unable to monitor the way the strengths of our beliefs fluctuate, so as to determine whether they conform to the demands that Bayesians would like to impose. We can't *comply* with Bayesian requirements, and so there's no point in

Followers of Elizabeth Anscombe have identified what they take to be the central consistency constraint on both the construction of actions and on their justification. Actions are composed of subsidary actions, which are sequenced so as to bring the agent, step by step, to the action's termination point or 'end'; the embedded actions are justified by pointing to the larger action which they jointly compose. For example, I am removing the can opener from the kitchen drawer, because I am opening a can of cat food, because I am putting cat food in the cat's dish, because I am feeding the cat.[26] The 'calculative,' or means-end or instrumental organization of the subsidiary components of the top-level action is a condition (and, as far as the die-hard Anscombians are concerned, the *only* such condition) on the top-level action being well-formed; it is, so to speak, a practical consistency condition that we impose on the production of actions, and it gives rise to the (the *only*, for the die-hards) *logic* of action. If I am asked why I am opening the can, and I reply that I am feeding the cat, it counts as criticism—criticism of the practical rationality of the proceedings—that opening the can is not going to get the cat fed, because this particular can is somehow empty, or because I have just finished feeding the cat, or because opening the can is not the first stage of plating the cat food.

In view of our argument up to this point, it is perhaps not a coincidence that one focus of discussion within the Anscombian school has been the relevant form of first-person authority: that when you act, you know what you are doing, without looking or otherwise investigating.[27] After all, in order to assemble actions that conform to the Anscombian consistency

imposing them; and so, we *don't* impose them.

If we *were* able to tell what the strengths of our beliefs were, with sufficiently low overhead, then we could impose Bayesian inference as an inferential requirement, i.e., legislate that only Bayesian updating *count* as inference. (That is, if we put to one side the difficult-to-navigate obstacles posed by the computational intractability of Bayesian updating; I'll touch on these in sec. 4.6, below.) In doing so, we would be *adding to our minds*, making our minds perhaps more tightly wired, or our psychologies more inclusive (in that they would contain more in the way of mental states than they now do).

[26] I am here smoothing over differences between Vogler, 2002, and Thompson, 2008, Part II, the two most prominent contemporary representatives of the view; they are developing ideas from Anscombe, 1985. Just to be clear, my appropriation of the Anscombean position for the purposes of illustration does not mean that I endorse it; Bowman, 2012, Mosdell, 2012, pp. 116-119, Brewer, 2009, and Vardar, 2024 have each helped convince me that it is overidealized and unrealistic.

[27] Small, 2012, and Rödl, 2007, esp. ch. 2; Schwenkler, 2011, and Schwenkler, 2015, although in other respects doctrinaire, provide a more nuanced version of the credo, one which takes account of the role of perception in the feedback loops that control activity—that is, that most of the time, you need to see what you're doing to pull it off.

conditions, routinely and in a timely manner, agents must be able to expeditiously determine whether their subsidiary actions are lined up in the right order. To do so, they must be able to know what subsidiary actions they are executing, when they are executing them; they must also have the calculatively structured sequence that they have in progress at their fingertips; and since they must normally generate further stages of an ongoing action by selecting subsidiary actions on the fly, they must be able to determine whether a candidate action belongs in the calculatively structured organization. In short, they must be able to answer a series of Anscombe's 'Why?'-questions: "Why are you opening the can?" "I'm putting cat food into the dish." "Why are doing that?" "I'm feeding the cat."

The modality of that 'must,' however, is that of a *demand*. Thus Anscombian first-person authority, like the other forms of first-person authority we have been examining, is in the first place a requirement, one that travels along with the demand that agents live up to the Anscombian action-theoretic notion of consistency—and only in the sequel a *fact*. And because *ought* implies *can*—again, because there is not much to be gotten by insisting that people know what they are doing, if they are simply unable to—the demand is imposed only because we *are* able to implement ways of knowing what we are doing that are expeditious enough. As before, it really does not matter what these are, as long as people generally *can* answer questions both about the embedding actions, about subsidiary actions embedded in them, and about how they fit into the calculative order. (Though let's not forget that there are fairly severe limitations on how well we're able to keep track of what we're doing: when the course of action becomes even moderately complex, agents cease to be able to report on its structure, or where they are in it, which is how we came to have PERT charts, Gantt charts, and related project management software. And once you become expert at what you do, you often become unable to say how you do it: you stop being able to say what the steps are.[28])

But notice that there are a great many demands on the control of action that we don't make, evidently because we're poorly equipped to discern, in real time, whether we're living up to them. We aren't all that bad at reporting *what* we are doing, but we're terrible at reporting *how* we're doing it, that is, at getting the adverbs right. *That* we are walking, we know; *how* we are walking, we rarely know. That may change; we have already seen

[28]That is, you probably *will* report a series of steps, but these will be, not the steps you are actually taking, but the ones you were told to take as a novice. The phenomenon was noticed as early on as Mill, 1967–1991, VII:189f.

experimentation with in-shoe sensors that permit real-time feedback on gait and posture. There is a market for devices like these only because people lack adjectival first-person authority about their actions.

Let's consider an additional constraint we could impose that has the look and feel of a consistency condition. Returning to an earlier example, actions consume resources (money, but not only money), and these resources need to be budgeted. It's easy to imagine applications that track resource expenditures in real time, that assign them to action-specific buckets, that provide alerts about budget trends, and that automatically adjust the budget of one course of action to accommodate costs incurred by other actions. Once we have such devices in place—that is, once we have additional dimensions of practical first-person authority, where I mean the only sort of first-person authority that matters: routine, reliable, on-the-fly data on our activities—we can collectively opt to impose much more rigorous demands on how actions are planned and executed. (E.g., from here on out, we must be ready to report not only on what we are doing, and in what order, but on how we are staying, and going to come in, under budget.) I will leave for later the issue of when it will be appropriate to count such expanded requirements as an upgrade to the *logic* of action. But since, in my view, sometimes it will be, we can make this observation: the choice of logic and the choice how to extend the self that enables more ambitious logics proceed in tandem.[29]

2.7

Before coming in for a landing, let's ponder a couple of exotic-sounding conclusions that might seem to be in the offing.

First, what seems to matter in first-person authority is whether the access that one has to one's cognitive states makes it feasible to comply with the inferential requirements for which one is responsible. Different institutional and other demands might make it reasonable to impose different responsibilities. Does it follow that how extended one's mind is turns out to be context-sensitive? That is, that for some purposes, certain resources might count as parts of your mind, and for others, not? Worse: that since we have suggested that your logic and the reach of your mind should be

[29] Returning to our riff on Kim Sterelny's discussion of extended digestion, administering requirements generally involves a good deal of collective overhead. So we cannot afford different consistency requirements for each individual. Consequently, a psychologistic philosopher of logic should think of logic itself not as a bit of one's extended mind, but rather, as scaffolding shared by many minds.

chosen jointly, one set of consistency constraints might count as your logic for some purposes, and another for others?

For instance, the tax authorities might impose bookkeeping requirements that do not even occur to your romantic partner; the auditing process is slow-moving, and so it suffices for tax purposes that over the course of a week, you can retrieve your ledgers from your safety deposit box. (When you are asked why you went to Fiji, you check the ledger, which says that it was a business trip; so that is what you report to the IRS.) The IRS might even determine that the ledgers count as part of your extended mind, insisting that what you report and what is in the stored ledgers be consistent. Your romantic partner may demand faster answers; in this case, asked why you went to Fiji, the right answer is not, "Give me a week to check my safety deposit box," but: "Of course it was for *you*."

There are real pressures towards maintaining a single scope for the inferential requirements one has to meet—that is, for one's mind to have a fixed, rather than context-sensitive, boundary. In addition to the overhead involved in keeping diverging stories straight, inferences based on diverging mental states will produce diverging outputs, which our social norms penalize. (If you tell the IRS one thing about what you were doing in Fiji, and your romantic partner another, you must be *lying*.) And if there are conflicts between those outputs that must be resolved for practical purposes, *you* will have to do the resolving; the problem here is not the apparent regress of selves behind selves, but the emerging need for more complicated ways of managing more complicated forms of consistency.[30]

Evidently, persons are in certain respects analogous to currency regions. Just as we do not want more than one currency to be used in our day-to-day transactions, and normally prescribe one currency as legal tender in a given area, we do not want multiple inferential consistency regimes that overlap, but are nonetheless differently scoped. There are good reasons to settle on a single boundary for a person's mind.

But second, you might be wondering whether, on the story we have been assembling, in a rather different manner, there are more minds around than you would have thought. Persons and minds are the administrators of consistency regimes. Here is a consistency regime: when you pick up, say, a cup, the pressure exerted by the fingers on one side of the cup must be the same as the pressure exerted by the thumb on the other side.[31] Maybe

[30]Thanks to Rachel Bottema and Liam Perri-Reichert, respectively, for these last two points.

[31]I'm grateful to Rosa Cao for the suggestion, and for the example.

there is, buried in your nervous system, an administrator anxiously complying with this family of consistency requirements, whom you do not so much as notice, because it is unable to talk to you. Maybe: I have no in-principle objection to discovering additional subpersonal minds. The interesting philosopical question here is what it will take to count as managing a consistency regime, as opposed to merely satisfying a requirement.

2.8

We need a moment for methodological remarks. The discussion to this point has been a bit like one of those chess openings in which someone is pushing pieces out to the center of the board as quickly as he can. We have not just extended mind, and not just first-person authority, but what's coming to be called the "adoption problem"—that is, whether it makes sense to *choose* a logic.[32] It's a lot to keep track of; why do it this way?

Analytic philosophers today inherit a default methodology, which consists, more or less, in collecting things you happen to think about some topic—these opinions are dignified with the label 'intuitions'—and treating them as data points around which to build a theory. And that has been pretty much the way the discussions with which we began, about first-person authority and extended mind, have been conducted. However, when you do it this way, there simply isn't a lot of traction. People can think just anything, and anyway, who cares what someone *happens* to think?

Problems consist in part of constraints: if you don't have constraints, you don't have a problem. If we are thinking about what the right logic is, not as a brute matter of fact, which is somehow supposed to be represented in those opinions that philosophers have, but as an engineering response to a constellation of problems, then we can ask what logic we should be implementing. And when we do, one problem can serve as a constraint for another; this means that it can be easier to find solutions for problems—for instance, about first-person authority, or extended mind, or philosophy of logic—when you consider them *together*.

And we *are* starting to think about logic, alongside the self, but with a difference. Both the antipsychologistic approach to logic, which over the course of the twentieth century was the exclusively respectable approach, and the psychologistic approach, with which I'm sympathetic, have treated the question of what correct inference consists in as a discovery problem,

[32]See Kripke, 2024, and other articles in the same issue of the journal in which it was published.

rather than an invention or design problem. That is, logicians and philosophers of logic have been asking what correctly done reasoning *already*, as a matter of fact, and independently of our choices and decisions, consists in. I am dissenting from this widely shared presupposition. In my view, the more important question is: What *shall* the forms of inference and the patterns of correctly performed reasoning be?

The problem of inventing the logic we need (and deserve to have) shouldn't be misconstrued as an invitation to just make something up, anything at all. The design problem we are considering is highly constrained; after all, if it were not, it would not be much of a *problem* at all. It is constrained by many mathematical facts, *inter alia*, the facts of proof theory, and of model theory, and of recursion theory (that is, by what question-beggingly gets called 'mathematical logic').[33] It is constrained by deep features of the world; for instance, by what John Stuart Mill called the 'Law of Universal Causation', namely, the contingent fact that inductions work out—and still further constrained by the fact that the Law of Universal Causation is not quite as universal as Mill thought. And especially it is constrained by facts about organisms that determine what they can be asked to do, in particular, we have just seen, by what kinds of internal monitoring and monitoring of detachable subsystems they are capable of, and also by what kinds of conclusions it is important for them to arrive at.

That list is likely to make it sound as though the constraints that figure into the choice of a logic are fixed. But as we've already registered in passing, what it's doable to keep track of can shift. If you happen to have been watching the extended mind debate, you'll have noticed a science-fiction tinge to it: someday soon, we'll all be cyborgs! The enthusiasm for bad cinema is offputting, but we *do* live in a technologically fluid era. Accordingly, we are adding an item to our philosophical agenda, that of considering how adjustments to our current inferential regime might reasonably take shape, and I'll turn to that in the next chapter.

That the problem is highly constrained does not mean that there is no freedom of choice: this is perhaps clearest when there are too many different sorts of pattern that could be traced out in the world, when we

[33] Question-beggingly, because there is no decent explanation, in the tradition of twentieth-century philosophical logic, of why these bodies of mathematical theory count as *logic*. But I do not here want to take up the question of why we classify model theory as logic, and number theory as something else, or for that matter, why we classify some demands on inference as 'logical', and others as, say, merely prudential. I'll make some preliminary remarks about how to proceed with the question in the upcoming Interlude, in sec. 3.7.3.

are cognitively capable of doing one sort of tracing or the other, and not capable of both at once. But even when there is one right answer, given the constraints, the choice one makes is still a choice. (It's accepting the offer you can't refuse.) This is a problem calling for great ingenuity and deep originality. It is a *practical* problem, and it is, again, the problem, not of what logic *is*, but of what it is *to be*.

We do need to pause, and ask ourselves what sort of approach to take to such a task. The question of how to reason about practical problems is badly undertheorized, and the shortfall is not confined to the theoreticians: we are quite generally brought up on a very small menu of practical inference techniques, and tend to default to straight means-end arguments. But when it's a question of what it is to reason correctly, simply taking means to your end is likely to get you to something that's not reasoning at all. A number of decades back, in discussions of what was called the 'ethics of belief,' examples were bruited about in which someone would ignore or misinterpret evidence that it would too costly to acknowledge; that's self-deception, or wishful thinking, not proper argumentation.

So the very next thing to take up is the question, or if you like, the meta-question, of how to think about the complicated problem that this chapter has put on the table.

Chapter 3

Private Persons and Minimal Selves

It's a commonplace that privacy can now be abridged and abdicated in ways that weren't routinely possible until very recently. I want to draw attention to an alternative configuration of the mind that these techniques make available, which I will call the *minimal self*.

My explication of minimal selfhood is going to take the long way around. I will have to explain what the ethical and political concept of privacy has to do with the very different philosophers' notion of logical privacy; this part of the discussion will connect the recent debates over extended cognition and first-person authority to one another. To get into a position where I can do that, I'll have to explain how selfhood and the laws of logic are also related topics. And to do *that*, I'll start out with an exercise in what Paul Grice and, following him, Michael Bratman have called 'creature construction.'[1]

Philosophers have a long history of treating selves and persons as an occasion for old-school metaphysics. Within the practice of old-school metaphysics, that only you can think your own thoughts is a remarkable fact requiring an equally remarkable explanation. I will be exploring the lower-key proposal that selves are administrative devices, that logical privacy is a bureaucratic requirement rather than a remarkable fact, and that the privilege of keeping your thoughts to yourself—a large part of privacy, in the layman's understanding of it—is an aspect of the information management regime to which traditionally organized selves belong. In the more minimal

[1] Grice, 1975; Bratman, 2007, *passim* and esp. 49f; Millgram, 1997, ch. 5, is a creature-construction argument on a related topic.

alternative version of the self which I will describe, not everything you think needs to be thought by you.

3.1

Let's turn first to motivating the design of the administrative device. Creature construction arguments proceed by describing a series of progressively more ambitious organisms or robots, each of which handles a performance shortfall identified in its predecessor; the immediate point of these descriptions is to isolate the features that support the incremental improvements in performance, and thereby to justify incorporating those features into the design of an agent faced with a particular range of challenges. After we have the upcoming creature construction argument in place, I'll return to the question of what we are to make of the conclusions of such exercises.

I'm going to opt for imaginary robots over imaginary organisms, and the primary dimension along which I'm going to arrange my robots is how much the robotics design team knows about the environments in which they're intended to operate.[2] Accordingly, at the beginning of the series, we can place the Mark I, which is constructed for a thoroughly understood and fully predictable environment: perhaps an assembly line on which regularly spaced, identically positioned widgets undergo a mechanically identical modification. For this task, the sequence of motions of the robot's arm and gripper can be hard-coded into silicon, and consequently the Mark I needs scarcely anything in the way of sensory inputs; it does not, for instance, need to see the oncoming widgets. And it needs scarcely anything in the way of a representation of its environment. ("Scarcely anything": the designers might want to build in a series of checkpoints, e.g., to verify that the gripper did close down on the part.) Because robots are thought of as autonomous (at least to some extent), this is a trivial and limiting case of an industrial robot.

But of course the Mark I is usable only in an extremely controlled environment. If a human being further up the line is putting the half-assembled widgets onto the conveyor belt any old which way, the Mark I must be junked, and the factory will upgrade to the Mark II: a substantially more sophisticated device equipped with a camera, capable of determining the positioning of a widget, and capable of computing the movements of its manipulator. The Mark II will construct and update representations of the

[2]Thus this particular creature construction argument is very much in the spirit of Ismael, 2007.

positions of the oncoming widgets, of its own effector, and perhaps other elements of its environment. However, those quite limited representations will not yet amount to a map—say, of the factory floor. The important step that has been taken here is that of delegating to the device part of the control that, in the Mark I, could be exercised directly by the robot's designer: only a small part, however, since in this case the designer is able to anticipate a great deal of what the robot is and is not to do. For instance, it is normal in robots of this type to manually specify a volume outside of which the robot's arm may not move.

The design team might decide that it wants a robot, the Mark III, that can be repositioned at different points in an assembly process. The Mark III, unlike its predecessors, is able to move around the factory floor, and so now a map is necessary. Perhaps the team can have the robot load a floor plan they provide during its initialization procedure, but because equipment is expected to move, the Mark III will need to update its map; moreover, because avoiding industrial accidents is a priority, it's a good idea to allow for updating in real time, on the basis of sensor inputs. (If the robot bumps into something that its map says is not there, it should be able to overwrite the map entry.) And because the programmer does not know where the robot will be at any particular time, he will have to delegate to the robot still more of the task of determining what it is going to do: now the robot must compute how it is to execute its tasks, given an initial location, and a map of the factory floor which it's responsible for updating.

There are many such incremental improvements to be walked through, at each of which the design team hands off more control to the robot it's designing. In the interest of brevity, I'm going to skip ahead a few models to a technique which I'm mentioning primarily because of its familiarity to philosophers. In one way of controlling the activities of a robot, let's say it's the Mark N, it derives a series of operations to execute via backward chaining from a specified set of goal states.[3] There are a great many philosophers who seem to think that such means-end derivations are the only form of problem solving through which the solution to an action control problem can be computed.[4] In fact, programming methods and control techniques vary, and here we can see this one to be just another step in such a sequence of designs: the step at which the engineer reserves specification of goals as his own prerogative, but delegates computation of a sequence of operations that will attain those goals to the device. The point of doing it this way

[3]SHRDLU was a famous early softbot of this kind. See Winograd, 1972, pp. 117–123.
[4]Smith, 1987, is a well-known example.

is that the robot will be functioning in an environment that the designer understands well enough to assign objectives to his robot, but does not understand well enough to be able to specify, ahead of time, how those objectives are to be achieved.

I want to emphasize that there is nothing special about this stage of the sequence, and just to make sure the point sticks, let's also sketch the Mark N+1, designed for a somewhat more demanding environment, in this case, a robotics competition with many, many rounds. At each round, the robots are scored by a panel of judges on a number of dimensions—but while the dimensions are labeled ("versatility," "intelligence," etc.), the design team can no longer predict what implementable objectives will guide the robot successfully through the contest. Their solution is to allow the scoring, at each stage of the competition, to modify both their robot's objectives, and the priority ordering among them. Each objective is given its own 'bank account,' and whenever it is successfully achieved, its account is augmented by the panel's score; objectives bid, in a simulated auction, to control the robot's behavior during the next stage of the competition; at every round, some percentage of the objectives mutate. Over the course of the competition, the robot's mix of objectives evolves to match the contours of the judges' implicit standards.[5] Here the design team understands the environment well enough to identify elements of a behavior-guiding feedback loop, but not well enough to be able to assign objectives to the robot ahead of time, and so they delegate to the robot the task of computing its own goals on the basis of values taken by suitable variables in the loop.

Let's move forward a number of additional stages, to robots intended for still more challenging environments (and now we are perhaps pushing past the frontiers of even contemporary, machine-learning based robotics). The designers will have to delegate the task of computing goals or objectives to the robot in even less structured environments than that of the Mark N+1. To make this vivid, let's imagine that, just as NASA's Jet Propulsion Laboratory has been sending Mars Rovers to a nearby planet to see what's out there, the space agency on some other planet is working on a mobile robot that they are going to use to explore the Earth. Let's call it the *Earth Rover*. What should the Earth Rover look like?

Because the point of such a robot is, precisely, to explore its environment, its designers know very little about what it's like where it's going. So the Earth Rover cannot come equipped with a map it loads from an

[5]The approach is adapted from the induction algorithm described at Holland *et al.*, 1986, pp. 47–50, 116–134, 146–150.

initialization file; instead, it will have to be its own cartographer. It will also have to develop representations of its environment that don't function in a map-like way; for example, when it records that the larger inhabitants of the Earth seem to be bipeds and quadrupeds, that observation is not associated with any particular map coordinates. Let's refer to this class of data as the Earth Rover's theory of its surroundings. The task of map and theory construction is continuous with the activities of earlier robots in our series, but, obviously, much more demanding;

How will the Earth Rover make decisions? The Earth is (from the point of view of the design team: is likely to be) a patchwork of very varied landscapes, and a well designed Earth Rover ought to do very different things in the different landscapes. For instance, if it lands in the Sahara, it should take soil and rock samples; but if it lands in the rain forest, it should take foliage samples; and if it lands on the Upper West Side, it should take sushi samples. Maybe its designers know about rocks, but they don't know about rain forests or Manhattan sushi bars ahead of time, and because the robot's home planet is so far away from the Earth, the robot cannot be built around the assumption that it will get timely guidance from its mission control when it first encounters them. Apparently, such a robot must be able to set its own objectives, as it develops its map and accompanying theory of where it has landed.[6] But how can the design team equip a robot to make headway

[6]Philosophers too often have entrenched reflexes that need to be turned off before philosophical headway can be made, and the reflex most salient in this case is insisting on a high-level objective or end, one which can be set by the programmer and not delegated. ('Exploring' would be a likely proposal, in this case; 'scoring high' is a likely suggestion for the Mark N+1.) But such broad, cover-all-bases objectives do not support backward chaining: in more traditional philosopher's jargon, when your end is something on the order of 'exploring,' you're not yet in a position to find a means to your end; you must first engage in some form of deliberation of ends. In Aristotle's version of the problem, everyone wants a happy, well-lived life—but what *is* a well-lived life? Before proceeding to try to live one, you have to do something else: perhaps specify, in a much more concrete way, what living well (and likewise: exploring, or performing well by the judges' lights) comes to—and perhaps also what it comes to in *these* circumstances. What this entails is that such ends are formally very different from the goals that do anchor means-end reasons. As a class, they are not well-understood: we don't know how to build robots that act on such ends, and we aren't able to say clearly how we use them ourselves. (For an overview of the specificationism literature, see Millgram, 2008.)

In any case, it is a mistake simply to assume that such engineering problems are to be solved by introducing a high-level or a 'final' end as a reference point, as opposed, say, to constructing feedback loops. E.g., if we define a global variable whose value depends on how much has just been accomplished in the way of additions or changes to a map; if we make the ways in which the robot sets itself mapping tasks sensitive to the value of the global variable; if we do all of that right, we could tell our funding agency that we

on the parallel mapping and planning problems they are envisaging?

Here is an approach that is a likely and promising part of the architecture for such a robot. (I am not suggesting that it will suffice on its own.) The development team already knows that the robot will be constructing a map of the environment it is traversing. They know it will be developing a theory to accompany its map. They also know that the robot will maintain data structures that guide its activities: perhaps task schedules or prioritized lists of objectives or variables that play designated roles in feedback loops. To have some shorthand for this, let's say that the robot is maintaining a field of representations. At any point in its mission, the Earth Rover will have available to it an already constructed field of representations.

Now, the design team can be reasonably sure that the Earth Rover's field of representations will contain errors: after all, even the best professional cartographers make mistakes, and who has not found mistakes in his to-do list? Again, the robot will be far, far away on Planet Earth; its mission control will be unable to fix those mistakes, and the robot will have to correct them itself. But those very mistakes can be exploited; in fact, turned into the fuel on which the robot's cognitive processing runs. To mine its mistakes, the design team can give the Earth Rover a set of rules that enforce informational hygiene on its field of representations. Hygiene violations are defined to catch what the design team anticipates are likely to be errors in the evolving field of representations; for the most part, they trigger one or another procedure meant to identify and rectify the problem. (But hygiene rules need not always target such errors directly: for example, the high-payoff hygiene violations are easier to identify when encoding conventions are uniformly followed, and so the team might determine inconsistent encoding to be itself a hygiene violation; we impose this sort of requirement on ourselves, a trivial example being our insistence on standardized spelling.)

What sort of triggers can the design team anticipate are likely to indicate mistakes in the map under construction? Perhaps they know enough about the surface of the Earth to expect that if the robot has visited the same map coordinates twice, and recorded a different altitude each time, or very dissimilar imaging, something has gone wrong. (Mostly they will be right, but not always: maybe the first time around, the Earth Rover has found its way to the top of an overhanging cliff, and the second time it is

---

were building a robotic model of curiosity: and the robot's exploration of its environment would not be, formally, goal- or end-directed. (For discussion of such feedback loops in human beings, see Millgram, 2005, ch. 1, Millgram, 2004.)

at the bottom, under the overhang.) So the robot's informational hygiene will require that each point on the map be tagged with no more than one altitude. Objectives in the task list that are not paired up with plans for accomplishing them might count as hygiene violations. And—I'm providing this example to alert the philosophers—the robot might have in its theory a pair of representations with the respective forms 'p' and '$\sim p$'; the design team might also designate that as a hygiene violation. They're pretty sure that, when it happens, something has gone wrong, although, from so far away, they won't necessarily know what.

Let's consider that last example a little further. There are evidently many responses the Earth Rover design team might provide to such a violation of representational hygiene. The crudest way of resolving the issue is to delete one of the conflicting representations at random; but the chances of guessing right are no better than fifty-fifty, and in any case, this procedure fails to leverage the hygiene violation to correct less obvious errors in the field of representations. If it is a matter of what the Earth Rover has marked on its map, the robot could revisit the relevant map coordinates and take another look: this will involve adding an objective to the task list, planning a route, and executing the plan. Because such investigations are expensive, however, the design team might see if the error can be diagnosed and fixed without spending valuable surveying time. If memory is cheap enough, the robot can be built to save backups of its representational field, to try to locate the source of the error in its history, and to revise its current informational state on the basis of those diagnostics.

In this approach, computing new objectives can be just a special case of rectifying hygiene violations. For instance, certain kinds of gap in the robot's map might be designated as violations. Sometimes the robot will be permitted to resolve them through interpolation; however, when the data for the surrounding regions are insufficiently smooth, the Earth Rover can be required to assign itself a new objective: that of conducting on-site observations. (When it's a gap in the theory rather than the map, we philosophers will call it induction rather than interpolation, and a very sophisticated version of such a robot may conduct observations elsewhere, meant to support better interpolations or inductions.) Again, when there are too many objectives on its task list to be integrated into an executable plan, that may count as a hygiene violation; one way to resolve it is for the Earth Rover to generate plans that include different objectives from the list, to rank the plans using an appropriate coherence metric, and to delete the objectives that are not included within the most coherent plan. Notice that the most coherent plan, on some coherence metrics, may be one that adds new objectives to

the list: the most coherent plan that unifies the already available objectives of surveying region A, and of surveying region C, may also involve surveying a region, B, between A and C, even though surveying B is not a means to either prior objective.[7] So cleaning up violations of representational hygiene can amount to doing (anyway some of) what we have already decided the Earth Rover needs to do: update and recompute its goals.

3.2

Creature construction arguments resemble the more familiar category of state of nature arguments: in both sorts of argument, a streamlined version of a familiar phenomenon is exhibited as solving a simplified version of a pressing problem. The two types of argument differ primarily in that creature construction arguments do not require the robots or organisms being discussed to appreciate the merits of the solution themselves, whereas state of nature arguments turn on the agreement of the population in the state of nature that the solution is satisfactory. In both sorts of argument, the point of exhibiting the artificial solution is to elicit a distinctive response, a recognition that has more or less this form: oh, so *that's* why we do it this way, and *that's* a way to see familiar, ordinary so-and-sos! We've reached the point in our own creature construction argument where I need to invite that moment of recognition.

We have described an administrative device used to manage the activities of agents that have to navigate fluid and hard-to-anticipate environments. The device is constructed by designating a field of representations that is subject to a set of hygiene constraints, and then by implementing mechanisms that do a sufficiently good job of ensuring ongoing conformity of the representational field with the hygiene regime. We can now introduce (reintroduce, if you've read the previous chapter) two technical philosophical terms: *selves* are those administrative devices.[8] The *logic* for a self is the

[7] For some experimentation with this approach, see Millgram and Thagard, 1996.

[8] Why the choice of term? I retained "person" in the evocative phrase, "private persons," even though in ch. 7 I'm going to give "person" a technical sense, one that ties it to the ongoing philosophical discussion of personal identity. (In an earlier published version of this chapter, I did use "person" where I'm now saying "self".) I'll do some sorting out of our terminology in the first Interlude, below.

Philosophers have also developed the habit of using "agent" not only when the production of action is intended; these days, the term picks out any freestanding locus of epistemic activity as well, and often functions as a synonym for "person." Whatever the merits of that way of talking generally, in the case we're discussing, the term "agent" would be appropriate: in its original meaning, an agent is someone you send somewhere

system of informational hygiene constraints governing the designated field of representations. As we saw, the design solution involves the delegation of a particular level of responsibility: selves are *responsible for* getting their field of representations to conform to their logic. When it's people, rather than robots, these assignments of responsibility have a social dimension; we demand conformity to and sanction deviations from the hygiene regime.[9] And this last claim about how responsibility is assigned in the design approach we're now discussing matches our own practice. If I believe that p, and I believe that $\sim p$, I am logically inconsistent: because I am violating the principle of noncontradiction, I am required (or anyhow, this is the standard view of the matter) to alter my beliefs and remove the contradiction. However, if I believe that p, and *you* believe that $\sim p$, we merely disagree; there is no inconsistency on anyone's part, and no requirement on either of our parts to revise our beliefs. For now, our present way of doing things enforces the logical norms only within the boundaries of a self (or an institution, such as a corporation, that is modeled on a self).[10]

If we are right to think of a self as this sort of administrative device, we have to think of logic as an administrative technique as well. Thus, what selves are and what the laws of logic are turn out to be are two problems to be solved jointly. Now, each problem on its own has given rise to a great deal of the wrong kind of philosophy: I mean, philosophizing that proceeds by table-thumping and insisting on whatever opinion one happens to already have.[11] But the coexplication of selfhood and of logic—the idea that types of selves and corresponding logics come as packages—promises to give us much more traction on each of them: given what the field of representations is going to have to look like, it is reasonable to impose certain hygienic constraints, and not others; given what constraints need to be imposed, the field of representations must be demarcated in such and such a manner.

Two points are best registered before proceeding. First, the view of logic I am advancing is avowedly psychologistic, by which I mean that logic is understood to address the question of what the prescriptive rules of inference and reasoning are, and to do that by considering, among other things, what

to do something for you, and in ch. 4 we'll again bring this side of creature construction arguments to the fore.

[9] Brandom, 1994, ch. 1, provides a philosophically useful caricature of the system of penalties and demands.

[10] But keep in mind that we're examining a member of a family of related techniques. For instance, Bratman, 2014, takes up occasions on which, when we're doing things together, we're required to have consistent beliefs to support interlocking plans.

[11] Millgram, 2009b, sec. 7, briefly describes a representative example from mid-twentieth-century philosophy of logic.

the mental activity of reasoning is like. For most of the twentieth century, accusing a philosopher of psychologism was a bit like accusing someone of communism during America's McCarthy era; it has been a long time since anyone has given level-headed consideration to the merits of psychologism. Allow me to recommend reviewing the arguments against psychologism, in both the analytic tradition (Frege) and the continental tradition (Husserl). You will find that they are fallacy-ridden, and that there were never any good (or even halfway-decent) arguments against psychologistic philosophy of logic.[12] I mean the present discussion to invite reconsideration of this long-neglected approach to logic.

Second, if selves are the administrative devices I have just described, selves are a design solution that will normally be only approximately implemented. Over the coming two sections, I will flag some of the ways in which actual people deviate from the streamlined and idealized functional characterization I have given. For now, bear in mind that no one is likely to be more than an approximation to selfhood. Selves are idealizations that are enormously important both for understanding what we are and for guiding what we do, but which we should not expect to find in nature.

Are we selves, even approximately? (I.e., is this the right idealization?) We have to get around in a world that is, by turns, familiar and deeply novel; even once we become familiar with some region or aspect of our environment, we are often forced to leave it and find our feet in a very different region of the world. We cannot get by on the goals we grow up with, or find in our animal repertoire; we have to map out our world, bit by bit, and change our plans as we see more of the map; when our plans change, our objectives change with them.

Naval vessels are described using lead ships that are of the same approximate design, as when we say that *Fletcher*-class destroyers were used in the Pacific theater. Adopting this convention, if the terse description of our capabilities that I gave a moment ago was correct, then we are Earth Rover-class agents. Earth Rovers can be expected to operate as selves, and we recognize the administrative procedure in our own activities. I think it is safe to presume that we are, anyway approximately, selves.

[12] Kusch, 1995, and Ringer, 1990, pp. 295–98, provide historical and sociological background; Millgram, 2009a, pp. 223–25, discusses one of the arguments against psychologism.

3.3

I need to pause to speak to an objection that many philosophers are likely not just to have but to feel very strongly about. I have described the hygiene rules introduced at the last stage of our creature construction argument as a robot's logic, and I have presented this as a way of understanding *our* logic: an informational hygiene regime that can be justified as part of a particular design solution. But, the objection will run, whatever this is, it isn't *logic*. The hygiene constraints I described were negotiable; the design team might have settled on other rules than they did for this robot, or on very different rules for a somewhat different robot; they might decide at some point to upgrade the robot's list of constraints. Whereas, the objection continues, logic is a matter of fact, not something you can *decide* upon, and it is the same for everyone. (There cannot be one logic for Earth Rovers, another for people, and yet another for who knows who else.) The science of logic (what follows will be a bit of a caricature, but not much of one) is the investigation of the consequence relation; it is the investigation of the structural features of a realm of abstract objects; logic is a branch of mathematics. Logic can be studied, described and understood, but not *chosen* or *altered*.

Doing justice to the complaint requires the most dramatic available illustration. The principle of noncontradiction was introduced above as an example of a choice the design team *might* make. For all I have said, they might have instead determined that cleaning up contradictions was not likely to be the best way of getting the robot through the challenges posed by its environment, and simply left noncontradiction off the list of constraints the robot is responsible for enforcing on its field of representations. On the view I am developing, even the law of logic that says that, for any p, $\sim(p \wedge \sim p)$, is a practical choice, made on the basis of performance considerations.[13]

[13]It may even be a choice that was, in our own case, made incorrectly. In imposing these sorts of constraints, feasibility is normally a relevant consideration; your robot should be able to live up to its logic, more or less. Ancient and medieval logicians were unaware that truth-functional consistency is a computationally intractable problem; it is not a demand we can live up to, and in real life the constraint is only enforced locally, and never globally. (For an accessible recap of Cook's Theorem, see Garey and Johnson, 1979, sec. 2.6.)

Notice also that in our own practice, contradictions are resolved not by simply deleting one of the conflicting conjuncts. Importantly, we often correctly emerge from the investigation triggered by the contradiction with a qualification to the truth of one or both sides of it: p was, we realize, approximately true; or it was only roughly true; or it was technically but not substantively true; or it was true allowing for the idealizations it incorporated... The enormous and subtle vocabulary we have for registering these variations on partial truth amounts to an elaborate repertoire for grading the accuracy of representations (Millgram, 2009a), and one beneficial effect of enforcing an inferential hygiene

On the traditional way of thinking, the bottom line is that both of p and $\sim p$ cannot be true, and that is all there is to it. But let's return to our Earth Rover mapping its environment, which may be the Sahara, or the rain forests of New Zealand, or Manhattan. If you think about it, it is obvious that different kinds of environments call for different kinds of maps. If the robot is on the city streets, it should construct a street map, but if it makes its way down into the subway system, it should construct a subway map. Its designers cannot know in advance what sort of map will be called for, and so, if they can, they will also delegate to the robot the task of determining what sort of maps to construct. Let's call the improved version that does this the *Earth Rover Advanced*; we can reserve the name "Earth Rover" for the simpler version that comes with a fixed set of map conventions.

It should be clear that because different types of maps will differ formally, different hygiene rules will be appropriate for them. For instance, a street map represents the distances from one point to another, and so testing for consistency of those distances with one another is appropriate. A subway map only needs to be topologically correct, and the distances from one point on the map to another will be properly determined by layout, readability and typographical requirements.[14] So the Earth Rover Advanced will also have to compute its own hygiene rules. That is, from the point of view of the creature construction argument we are developing, there is no answer to the question: is logic fixed and unalterable? Rather, the question is how much responsibility has been delegated to the robot, and how much control is being reserved by the design team for the operators. At the Earth Rover stage, the design team delegates responsibility for constructing and maintaining a field of representations, while reserving to the programmers the specification of the constraints the field is required to satisfy; at the subsequent stage, the design team delegates responsibility for adjusting the constraints governing the field to the robot as well.[15] As before, the motivation in the former case

regime is to increase the level of nuance in those grades.

And here is a recent observation in this vein: "tradeoffs may be regarded as a twentieth century analog of paradoxes in Ancient Greek philosophy... Paradoxes *problematize* entrenched commitments or stances, indirectly pushing us to re-investigate what we have been taking for granted. Pragmatic tradeoffs... problematize entrenched dichotomies, pushing us towards seeing spectrums of choices instead" (Haber, 2025, p. 296).

[14]Some subway maps try to have it both ways; it's instructive to compare a current New York City subway map with the much more usable (and incidentally aesthetically superior) 1970s Metropolitan Transit Authority (MTA) map.

[15]Will the robot's logic be *formal*? As MacFarlane, 2000, has documented, that question has been taken to mean at least three different things over its history. But we can say this much: Constraints that are maximally general, rather than subject-specific, are suitable

is that the robot's designers know enough about the challenges posed by the robot's environment to specify its logic, but not enough to anticipate the concrete results to be obtained by enforcing the logic; in the Advanced case, the motivation is that they do not know enough about those challenges even to specify the robot's logic ahead of time. And so the question for us, as philosophers of logic, is: when we consider ourselves from the design stance, how much delegation of control should we take there to be?

In *Principia Ethica*, G. E. Moore pointed out that questions like, "It's pleasant, but is it *good*?" are always intelligible, in that they might, in principle, be answered one way or the other: they are, as he put it, open questions.[16] He concluded that being good could not be, say, being pleasant (and more generally, that 'good' could not name any natural property); if it were, the question would not be 'open'. Today, Moore's Open Question Argument is remembered by metaethicists as historically important—it kicked off twentieth-century metaethics—and also regarded as thoroughly confused. It *was* thoroughly confused, but Moore had stumbled on an important philosophical tool, one whose proper use he never understood. If we are right, asking such questions can serve, both in metaethics and philosophy of logic, as a probe that registers where in the creature construction sequence one can find *oneself*.

That we are able to ask, "It's pleasant, but is it *good*?" tells us that we are not well modeled by a robot whose designers understood its environment thoroughly enough to hard-code an evaluation of any particular sensation—this being Moore's naive model for pleasure—into the silicon. Instead, the

---

for guiding their robot through environments that are expected to contain subject matter that is unfamiliar to the design team. And constraints that can be applied simply by checking the forms of the representations, and which do not require judgment calls, are much easier to monitor, and so much more straightforward to impose. We can expect the design team for an Earth Rover-class device to specify a formal logic for it, if they at all can.

Can they? The design team may also anticipate a *messy* environment, one in which its robot's inferences will need to be driven by forced matches: for instance, by descriptions that are not fully true, but rather true *enough*. (See Millgram, 2009a, chs. 4–6, for a step-by-step exposition of this train of thought.) In the vocabulary of our creature construction argument, that is to say that the robot must count some configurations of its field of representations as hygiene violations when they are not strictly ruled out, but come *close enough* to being ruled out. It is quite implausible that such a 'close enough' can be specified formally: for one thing, it is normally content- and task-sensitive. So a still more advanced robot, designed for an environment anticipated to be messy, will have a logic that may have a formal component, but which will in large part not be formal. We should think of an exclusively formal logic as one step of a creature construction sequence, and in subsequent chapters, I will take up defeasible constraints on inference.

[16] Moore, 1903/1960.

right design, for creatures that have to cope with environments like our own, involves delegating the evaluation of sensations (and more generally, features of the 'natural' world) to the agent.[17] And if we are *un*able to see questions like, "It's good, but does that give me a reason to do it?" as sensible, that tells us that the further step to a particular more-ambitious delegation of control and responsibility was not necessary or feasible for *our* as-if designers. If we have these responses to this pair of questions, we are in a position to see ourselves as a design solution tailored to an environment fluid enough for the designer to need to delegate the task of evaluating one sort of item or another to his creation, but sufficiently well-understood for the designer to specify what the ensuing evaluations are to mean for the robot's decision-making. It might be necessary, when a robot has to operate in a still more fluid environment than that one, for its designer to allow it to recompute what its reasons are, *given* a set of evaluations; if we can see that last question as open, we should proceed on the assumption that we are better modeled by the robot at *that* stage of the relevant creature construction. Let's call this use of such questions *Moore's Probe*.

Return now to philosophy of logic. Philosophers who find it unintelligible to ask, "It's a contradiction, but is that a reason to revise my beliefs?" may think that their intuitions about the principle of noncontradiction are revealing to them a matter of fact belonging to a mind-independent science of logic. What they are really doing is locating themselves in the creature construction sequence: they are identifying themselves as Earth-Rover-class, as opposed to Earth-Rover-Advanced-class agents.[18] The responses elicited by Moore's Probe are valuable, but not as limning a consequence relation to be found in a realm of Platonic entities; instead, they help us identify our own capacities and limitations.

3.4

If you look at a directory listing on your computer, you will notice toggles on each file or subdirectory for read, write and execute privileges. The Earth Rover follows an administrative procedure which enforces informational hygiene constraints on a field of representations. How will the design team set

[17] Why the scare quotes? Just try to get a philosopher to explain to you what he means by "natural".

[18] Philosophers whose use of Moore's Probe returns this result sometimes get huffy at the suggestion that there is another alternative, and here is a typical instance: "no attempt will be made to argue with those who think it acceptable to contradict oneself" (Williamson, 1994, p. 189; compare p. 136).

the privileges on the items in that field?

The Earth Rover is assuming responsibility for correcting violations of the hygiene regime. It cannot do that unless it can identify violations. And it cannot identify violations unless it can check whether the entries in the field conform to the constraints. So the designers will assign it read privileges to the field of representations it is responsible for maintaining; that's just part and parcel of the implementation of this administrative device.

That the robot *just knows*—via what is likely a primitive but in any case a low-cost, routine and highly reliable operation—what the entries in its field of representations are is an implementation requirement: one way or another, the design team has to make sure this is a feature of their device. The philosophers' labels for this requirement being met are 'first-person authority,' and, more recently, 'self-knowledge'; that you *just* know what you think, in that very special way that others do not, is treated as a remarkable feature of the mind, and usually taken to require a very special sort of explanation, in the style of old-school metaphysics. Accordingly, a philosophical subspecialization has emerged whose members compete to provide such explanations. Just for instance, according to some of them, you have what might as well be a camera inside your head that tells you what you think. Again for instance, according to others, your pronouncements about what you think are to be construed as analogous to cries of pain, and are not properly reports at all.[19] But we can now see that it scarcely matters

[19]For sample 'detectivist' views, see Armstrong and Malcolm, 1984, pp. 108–137, Nichols and Stich, 2003, ch. 4, esp. pp. 160–164. The label is due to Finkelstein, 2003, which is itself one of several developments of expressivism. Unsurprisingly, we also find denials that there is any such feature of the mind to be explained; Gopnik, 1993, and Carruthers, 2011, are examples.

Sydney Shoemaker devoted a substantial part of his efforts, over an academic career, to accounting for first-person authority. (See especially Shoemaker, 1996, but also Shoemaker, 1963, ch. 6.) With a handful of exceptions, the arguments fall into two groups: those that show that when an agent falls short of self-knowledge, its rationality will be impaired, and those that show how a rational agent can exercise first-person authority even when it cannot do the job by directly retrieving items stored in memory. I have found these very helpful in thinking about the functionality of an Earth Rover, although, as we'll see in our second Interlude, I have objections to some of them.

From the point of view of the design analysis we are developing, arguments in the first group show why first-person authority is an implementation constraint on Earth Rover-class agents. And arguments in the second group show that first-person authority does not need to be implemented across the board as hard-coded memory access operations. However, that latter group of arguments do have to be taken with a grain of salt. As we'll see, in our second Interlude, some of the proposed workarounds impose high computational overhead (because they depend on the agent running through overly involved trains of ratiocination). When the overhead is too high, these will be poor or even unworkable

how the implementation requirement is satisfied, as long as it *is* satisfied. And unless we imagine our robot's design team to be obsessed with elegant engineering, we can expect it to be satisfied in a number of different ways: there should be no *one* such explanation for first-person authority. The appearance of an occasion for old-school metaphysics very often arises from confusing a *demand* with a *fact*; for philosophical purposes, all that matters here is to explain the demand.

That the Earth Rover has write privileges for its representational field is also an implementation requirement. The Earth Rover is assuming responsibility for correcting violations of the hygiene regime. It cannot do that unless it can remove violations. And that entails either deleting, overwriting, or adding representations. A largish proportion of the recent discussion of self-knowledge turns on the observation that you can know what you think by making up your mind about it.[20] This describes a side-effect of what is, once again, not a fact but a requirement: the Earth Rover has to be able to write out changes to its data; if it cannot revise entries in its field of representations, it cannot maintain its hygiene regime.

The requirement that the Earth Rover coordinate its representations as specified by the hygiene constraints will not normally be met if other parties are also able to modify representations within the field. (E.g., if the

---

implementation choices.

[20]Among such 'constitutivist' views, McGeer, 1996, argues that you know what you think because, once you have decided and announced what you think, you can live up to your announcement, by ensuring that your subsequent behavior conforms to it. (See also McGeer, 2007, and, for a related view, Bilgrami, 2006.) Moran, 2001, argues that you know what you think because, whenever the question arises, you can reconsider your opinion about the topic at hand, on its merits. The accounts are very close indeed, but we can distinguish them this way: Moran is focused on upstream inputs to your decision, and McGeer on downstream followons. The question of philosophical interest is raised indirectly by Lawlor, 2003, which points out that they *do* sometimes come apart: you make up your mind, but don't follow through. So what brings it about that the phenomena of interest to McGeer and Moran so often travel together? How is it that once we have made up our minds what we think, we frequently follow through, and live up to what we have said we think? How is it that when we follow through, it is often because we have made up our minds? In my view, this is best addressed as an implementation issue, via the sort of creature construction argument we are now developing.

Moran seems to take his 'transparency' view to compete with detectivism, but this is evidently an error. If you find out what you think about some question by reconsidering it, the bases for your reconsideration—without loss of generality, we can take these to be the premises of an argument—are either available to you in the very same way, or in some different way. There is only so much reconsideration you can afford on any occasion, and no amount of it will staunch the regress. So sometimes you must know what you think, not by way of exercising your write privileges, but by exercising your read privileges.

robot tries to eliminate the conflict between representations with the forms p, $\sim p$ by overwriting the p-representation with a q, that will not ensure conformity with the hygiene constraints if, at the same time, another party is likely to be busily adding representations to the field that perhaps conflict with the q-representation.[21]) If you look once again at the long directory listing on your computer, you will see that it allows the access toggles to be flipped different ways for different users: for instance, a file might be marked as readable and writable by the system administrator, as entirely off-limits to another group of users, and as read-only for you. We can anticipate the design team will enable write privileges (and, if the robot is going to be operating in a hostile environment, read privileges) for the robot alone.[22]

Persons, evidently, are private, and because I will presently introduce a related but contrasting administrative device, let's call instances of the traditional design solution *private persons*. Because human beings who implement such a design solution will end up shaping their intellectual apparatus around its features, truisms like, 'Only you can make up your own mind,' and 'Only you can think your own thoughts' will end up conceptual truths, and will seem to invite superlative explanations. They do not. Logical privacy is a side effect of adopting efficient descriptive conventions against the background of the stably implemented administrative device of private persons: that is, against the background of a good deal of *de facto* privacy.

Third, recall that the Earth Rover will sometimes resolve violations of its hygiene regime by developing and executing a plan for revisiting a site about which it seems to possess contradictory information; more generally, sometimes its method for correcting violations involves adding new objectives to its task list and taking action. There is no point in having the robot write subroutines for itself unless it can run them; accordingly, it is also part and parcel of the design solution that the Earth Rover have execute privileges to its representational field. But the Earth Rover will not be able to complete the tasks it sets for itself if other parties are able to interfere with its activities by running their own programs on it, and so the design solution will also reserve execute privileges solely to the robot.

We can now turn to making a distinction and adding a needed qualification. I introduced selves as zones of logical responsibility, and I pointed out that read-write-execute privileges are necessary if the administrative de-

[21] That said, there may be other workable approaches. As remarked above, Wikipedia is much more consistent than the rationale I have just given would lead you to expect.

[22] For a biological version of the point about restricted read and write privileges, see Sterelny, 2004.

vice is to be effective. However, it is important to acknowledge that we are not merely selves: we do engage in the administrative activities I have been describing, but this is not everything we do, nor everything we are. For instance, your muscles are not elements of your field of representations, but they are nonetheless part of you. And the Earth Rover's computational engine may process a great many representations on which it does not enforce the sort of hygiene rules I have been describing. Let's call the computational activities of the Earth Rover *cognition*, and the representations that belong to the field on which it enforces its hygiene constraints its *mental states*—collectively, its *mind*. Mental states are, in this usage, a subclass of cognitive states; we should not expect the Earth Rover to have across-the-board read-write-execute access to all of its cognitive states. After all, we have no reason to think that, from the point of view of the design team, it needs those across-the-board privileges.

Let's explore this distinction. First, consider whether we should think of perceptions as mental or merely cognitive states. If they *are* mental states, they might seem to be an objection to the account I have been developing; we cannot change what we see and feel in the way that we can change our mind about other things, which is to say that we do not seem to have write privileges to our sensations. Now, that observation has to be complicated a bit, because we are told that the computational states that implement our sensations are in fact undergoing constant revision. What matters for our purposes is that the revision is not transparent to us, we do not control it, and we are accordingly not held responsible for it—anyway in the way we are held responsible for logical inconsistency. We can model perceptions as subject to an asymmetrical informational hygiene rule: the Earth Rover is required to rectify mismatches between its perceptions and elements of the field of representations we have already described, but not by altering the perceptions to match other data in the field. As before, this is reflected in our own practice: expectation may determine perception, as the catchphrase has it, but we are not held *responsible* for adjusting our perceptions to our expectations, and we do not merit labels like 'inconsistent' and 'irrational' when we fail to do so—on the contrary.[23] I mentioned earlier that we are

[23]In fact, the situation is trickier than that. Although Dennett, 1991, is now considered somewhat dated, it is deserving of continued attention for developing a series of arguments whose conclusions, however, the author seems to have misinterpreted. These arguments establish that a perceptual experience consists of a constantly updated stream of representations, no element of which has the sort of administrative status shared by other elements of the central field of representations in a robot of the Earth Rover's design. That is, no item in the stream *counts*, for purposes of the hygiene regime, as the final, designated sen-

best thought of as approximations to the administrative idealization of selfhood; now that we have noted that perceptions are only partially covered by the sort of hygiene regime I was describing, we can choose to classify perceptions as cognitive rather than mental states, or to describe the way we handle sensory inputs as one way in which we are messier than the clean administrative device I was using as a model.

Memories are also liable to strike philosophers as another potential counterexample: a class of mental states for which we have read privileges but not write privileges.[24] We are not supposed to change what we remember when it proves to be inconsistent with what we believe on other grounds, and as with perception, we can accommodate this in either of two ways. We can classify memories as cognitive states rather than full-fledged mental states, or we can register that the highly simplified, idealized model that we arrived at in our creature construction abstracts away from much that goes to make up actual human beings: by acknowledging, that is, that we are only approximately selves.[25]

---

sation which other representations are required to match. Accordingly, the requirements of fit between sensory inputs and the designated field of representations in minds such as our own are much harder to formulate, in that the sensory inputs are moving targets.

[24]Perception and memory are not the only such states. Let's suppose that a member of the Catholic church is supposed to believe what the Pope believes on matters of doctrine. The member will not be able to resolve hygiene violations by changing the Pope's mind, not, anyway, in the way that the member can change his own. Once again, we are only approximations to selves.

[25]However, as with perception, the facts on the ground are more complicated. We regularly replace our memories with synopses, and those synopses with synopses of *them*... and there is evidently a creature construction explanation for that fact. (For a somewhat dated discussion of memory construction, see Conway and Playdell-Pearce, 2000.) Tersely: recall the suggestion that memories are added to a robot initially to support debugging; when a violation of the informational hygiene regime is identified, it will often be cheaper to look at the backtrace to find the source of the error than to revisit the sites of earlier observations. (No doubt memories will have other uses as well.) Now, even if memory is inexpensive, searching it is not. To keep the ever-increasing bulk of records usable, it will have to be continually redigested into short, simpler and easier-to-navigate versions; and it may be reasonable to do so on the assumption that, roughly, the further back in the past the event is, the less it matters to be able to retrieve details. The upshot will be a robot, the Mark M, whose memories *are* being revised on a regular basis, but—and this is what allows us to classify them as cognitive rather than mental states—these revisions are not subsumed under the set of responsibilites for maintaining the robot's field of representations, as I described it above.

Philosophers worry about personal identity—about what makes you the same person you were yesterday—and you may expect the addition of an archive to the Earth Rover to be an entry point into a theory of what it is for Earth Rovers to persist over time. I will explore a creature construction approach to diachronic personal identity in ch. 7.

3.5

Now that we know what selves are, I want to speak briefly to a recent debate over their borders at a time. The argument has to do with whether minds can be 'extended'—that is, with whether or not all of your thinking goes on inside your head. The alternative is that some of what is properly your own mental or cognitive activity might take place in intellectual prostheses such as calculators or datebooks, or even in nonartifactual parts of your environment.[26] Having taken up the discussion in ch. 2, just now I will do the very minimium in the way of straightening up needed to get to the next stage of the argument at hand.

First, in this debate, the participants treat the phrases 'extended mind' and 'extended cognition' as equivalent—although some of them lean heavily on the more scientific *sound* of 'cognition'.[27] Recall, however, that we have distinguished these terms: your mental states are subject to a hygiene regime which you are responsible for enforcing; cognitive states are representational states that may or may not be subject to those hygiene rules. Second, the participants take themselves to be arguing over a question of fact: whether there *are* (or might be) mental or cognitive states outside people's heads. Taking the existence (or possibility) of cognitive states outside the head to be a question of fact is a mistake, but not one that concerns us here.[28] However, it should be clear at this point that what

[26] For an exposition and defense of the position, see Clark, 2011.

[27] In somewhat the same vein, you find the latinate 'intracranial' replacing the ordinary English 'in the head'.

[28] "Cognitive science" was used as a rallying point for a research program derived from a defunct philosophical theory of mind, Hilary Putnam's functionalism (1975c). Deep background: Most of the argument about extended cognition, which I'm trying to sidestep, arises from a deep problem with that view. The idea at the center of the research program is to treat minds as computational systems; the problem is that computation is a purely formal notion. So where one draws the line around the physical region that is being represented computationally is left entirely to the discretion of whoever is constructing the computational representation. It is a question of notation and convenience, not of fact. So when Clark and Chalmers, 1998, announced that cognition could be extended, they were not making a discovery, but stumbling upon the vacuousness of computational functionalism.

(Briefly, for those to whom this is an unfamiliar observation: Any physical object can be represented as a Turing machine—in fact, can be represented as implementing *any* Turing machine—and here's a trivial example to give you the idea. Let the state of your body this minute be a simple Turing machine's initial state, and let the state of your body the subsequent minute be the machine's halt state. Let the surface of your body this minute be one of two symbols from the machine's alphabet, and let the surface of your body the next minute be the other symbol. Set the transition function of the machine to map the

count as someone's *mental* states is a design choice: in our own case, it is something that, subject to feasibility constraints, we can *legislate*.

Nowadays, most people carry wirelessly networked computers with them, pretty much at all times. (They think of these computers as telephones.) They store a great deal of information on those devices, and they do not, at least yet, censor the flow of information between their brains and their phones. As things stand, if there is a representation of the form p 'in your head,' and a representation of the form $\sim p$ on your phone, you may be likely to miss an appointment, but you are not thereby inconsistent. We could change that, by requiring you to enforce the hygiene rules not only on what is now your mind, but on the information stored on your mobile telephone as well; we could mandate that a violation of the extended hygiene regime renders you criticizable as inconsistent and irrational. And if we did that, your phone's memory, or part of it, would *count* as part of the field of representations for which you are responsible: it *would* be part of your extended mind. However, I do not find it profitable or instructive to argue over whether to make this change, anyhow on its own.[29]

3.6

Here, on the other hand, is a modification to the administrative device we have been discussing that it is now realistically possible to implement, and which I will try to convince you is of philosophical and practical interest.

---

initial state and the first symbol onto the halt state and the second symbol. Presto: for two minutes, you're a Turing machine!)

Some attempts to resist the thesis of extended cognition have appealed to the practice of cognitive science (e.g., Adams and Aizawa, 2008, Rupert, 2009). But you shouldn't take the possibility of extended cognition to be settled by appeal to what cognitive scientists happen to do these days, when they're figuring out human or animal cognition. The leading idea of the program, again, is that cognition is computation, but computers are *devices*, and computation is an artifactual category. Devices are items we can alter and improve; you can no more argue that extended cognition isn't really cognition, because it looks different than what cognitive scientists have been looking at so far, than you could have reasonably argued in 1975 that cars can't come with fuel injection, because 'automotive science' had to that point proceeded on the methodological assumption that cars come with carburetors.

[29]Looking back on his earlier engagement with extended minds, Chalmers takes it to have been attempting to renegotiate the concept of belief, or perhaps words used to describe things in the neighborhood of belief (2025, p. 2909). And since words are only words, and Chalmers reasonably thinks that "you shouldn't get to hung up on words," there's a felt need to explain why the proposal mattered: "words," he adds, "matter at least for practical purposes."

In human beings, the read privileges to the contents of someone's brain are restricted to that someone, and there's a feasibility consideration that dictates doing it this way: we do not know how to support read privileges for third parties. But third parties (these days, primarily corporations and state agencies) *can* access the contents of a networked portable computer, such as a mobile telephone, or of cloud storage, in real time. Not only can a third party search for violations of specified hygiene rules, it can correct them. We already do rely on this way of handling errors in trivial cases; when a driver is following turn-by-turn directions, and he misses a turn, the navigation device recalculates the route and gives corrected directions on its own. It is not hard to imagine progressively more ambitious forms of outsourced hygiene control. Starting very close to where we are now: when the device determines that there is too much traffic for the two of you to meet at the restaurant you had planned, the scheduling application negotiates a different restaurant with your lunch date's scheduling application, makes reservations, gives you the appropriate turn-by-turn directions, and both of you find yourselves getting out of your cars, not where you had originally agreed, but in time for lunch. However, if something goes wrong, although this sort of delegation may provide you with an *excuse* ("there was a glitch in the scheduling software"), today the responsibility for getting things right (say, for turning up at the right place and time) ultimately rests with you. If you do not actually turn up, you apologize, where that consists in acknowledging that you are at fault.

That could change. Consider the following reassignment of responsibility for informational hygiene. You (but I will qualify the use of that pronoun in just a moment) are responsible for enforcing informational hygiene requirements over a brain-based field of representations. You are responsible for enforcing the coordination of the brain-based field with a specified field of representations on your mobile telephone; that is, you are responsible for making sure their contents match. But the responsibility for enforcing informational hygiene requirements on the networked device is outsourced to third parties. (For instance, if what you think you are doing today and what the phone-based schedule says differ, you are thereby inconsistent, and required to make them match; but a service checks your on-line schedule for consistency.)

In this way of assigning responsibility, when the error was one the third party was supposed to correct, and you seem to apologize for not turning up on time, 'I'm sorry' no longer acknowledges that you are at fault; your excuse is no longer an attempt to mitigate blame. Rather, saying that you are sorry merely expresses polite sympathy, in something like the way that

an administrator says that he is sorry when it turns out that accounting has not cleared your paycheck: he is acknowledging that it is unfortunate, but what accounting does or doesn't do is just not his responsibility.

Selves are administrative devices. Private persons are the traditional version of this device, in which a single central field of representations is demarcated by coextensive read, write and execute privileges; these support the allocation of responsibility for enforcing representational hygiene to a single locus. The alternative allocation of responsibility we have just described is no longer a private person (and because 'private persons' was our label for the only model of selfhood we have had, the alternative no longer counts by traditional lights as a self at all). For reasons I will get to momentarily, let's call the alternative the *minimal self*.

When I was describing the alternative assignment of hygienic responsibilities, I promised to qualify my use of "you". Recall that people structure their collective intellectual apparatus around features of their lives that they come to take for granted: pronouns like "I," "you," "he," and "she" have built into them the presupposition that a *self*—a traditional, *private person*—is being picked out. If minimal selves come to replace private persons, we will have to adjust our intellectual apparatus, and not just our conceptual and analytic truths, but our pronouns, to accommodate the new state of affairs.[30] In order to avoid awkward circumlocutions, and the even more awkward attempt to introduce terminology suited to arrangements that are not yet in place, I stuck with the current but not-fully-appropriate pronouns. You can imagine scare quotes around them as you make the necessary adjustments.

Minimal selfhood requires restricting logical privacy, which in turn requires surrendering a good deal of old-fashioned privacy. Electronic telepathy is an available off-the-shelf feature for digital computers, but not for brains. Consequently, the self, reconfigured as I have described, is minimal *in virtue of* being extended; minimality is not a matter of how large or small the fields of representations being adminstered are, but of whether there is an administrative monopoly.

Sometimes a one-person business becomes a large, publicly-owed firm; once upon a time, a city built on seven hills became a Mediterranean empire. It would be a mistake to think that the founder and CEO *is* (really!) the firm, in something like the way it was a mistake when the Roman Empire

[30]The general point is nicely made by Rovane, 1990; she is contemplating a society in which Parfit-style fissioning—people splitting into two—is an ordinary and anticipated occurence.

was taken to be really the city of Rome. And it would be roughly that mistake to observe the human being at the center of the minimal self, to observe that he still retains administrative responsibility for his brain-based representations, and to conclude that the human being *is* (really, still) the self. The human being has become a component, albeit an essential one, of a more complex administrative device, and that is reflected in his hygienic responsibilities: because the brain-based representations violate the hygiene regime when they do not match the computer-based representations, third parties can make up the minimal self's mind—can *come to believe*—on his behalf.

3.7

Minimal selves are to private persons as the libertarians' minimal state is to the swiss-army-knife state we inhabitants of the developed world have become used to living in. In the libertarian utopia, a pared back organizational core is retained in order to enforce the drastically pared back rules of the game, and to prevent armed invasion; other functions, even those that have come to be regarded as standard responsibilities of government, are outsourced or simply left to take care of themselves. A libertarian minimal state might outsource peripheral military functions to contractors, privatize the post office, or eliminate government food safety inspections. In the minimal self, the outsourced responsibilities differ, but the arrangement is structually similar: the brain and the body it is in play the role of an organizing anchor for an extended region of personal sovereignty, in which the exercise of responsibilities formerly assigned to selves is turned over to external agencies.[31]

How should we think about the suddenly possible future in which private persons have been supplanted by minimal selves? That is, since private persons are what we have meant so far in talking of selves, how should we think of a world without selves?

As we take up this question, we need to remain aware of the obstacles to addressing it intelligently. First, as recent history teaches us, it is hard to

[31]The large-scale social organization composed of minimal selves is evidently a novel form of feudalism, one in which vassalage penetrates the boundaries of the self. Plato thought that the form of justice was the same in the *polis* and the individual, and he took that as a constraint on his own political theory. Recent political theory rarely endorses or satisfies this constraint; you will notice, for instance, that liberal political theory presupposes citizens who do not themselves have the internal organization of a liberal state.

tell what the uptake of a technical innovation is going to look like in real life. For instance, in the mid-1960s, it was clear that networked computers were on the way; Janet Abbate recounts how the internet arose out of a series of mistaken expert judgments as to what users would do with the connectivity: email, amazingly in retrospect, was believed to be an unimportant application. The ARPAnet became the kernel of the internet as we now have it because its architecture was flexible enough to support functionality unintended by its designers.[32] We should assume that the picture we have to work with is wrong in all of its details.

And we should not assume that the picture we are considering *is* the future, details to one side. One alternative possibility is that the emerging technologies will reinforce rather than undermine the private person. As it is, the responsibilities we assign to selves are only sporadically enforced; we honor them to a great extent in the breach, and when we do live up to them, it is often enough from habit, rather than because we will be called on it if we do not. But social networking sites make the record of one's commitments public, and may over the long term make it very hard to walk away from them. Perhaps we have never really been selves, or even close approximations to them, and perhaps we will finally be forced to become the traditional, private persons that we imagined we were.[33]

Second, it is hard for us selves to see the choice between private personhood and minimal selfhood as one that *we* can make, and so, as a choice at all.[34] And third, the track record suggests that the collective choice will not be made for the reasons we philosophers are likely to see as the right ones. Think of what we might call *Google panopticism*, the startling emergence of new norms of privacy.[35] In the new way of doing things, the

[32] Abbate, 1999.

[33] For this last suggestion, I'm grateful to Aubrey Spivey; Vogler, 1998, suggests that, if this happens, we will not like it one bit.

[34] Nussbaum, 2001, highlights the perplexities of a formally similar choice considered by Plato.

[35] The phrase is adapted from Michel Foucault's adaptation of Jeremy Bentham's name for a prison he planned but never built. In Bentham's panopticon, the interiors of all cells would be visible from a central observation post. Foucauldian panopticism is an institutional rather than architectural innovation: in former times, punishments were draconian, but laws were sparse; in many modern institutions, rules cover even small details of behavior, violations incur relatively minor punishments, and checkpoints are frequent. Foucault discusses prisons, insane asylums, hospitals and so on (1995; 1988; 1975), but for an academic audience, universities are probably the most helpful example. Students turn in homework assignments, are quizzed on the reading, graded on their classroom discussion and so on. All of those grades are entered into a spreadsheet, and eventually merged into a grade that is sent to the registrar and compiled into a transcript,

details of individuals' lives are monitored and recorded with an exacting thoroughness formerly reserved for celebrities and spies—but the records are anonymized before use. Google panopticism was accepted by the public largely because it was the price of a free, high-performance search engine, of the minor convenience of being able to pay with a bank card, and of satisfying the yearning, appropriate to high school but not thereafter, to be seen to be a member of a popular clique.[36] If we become minimal selves, it will probably be for equally bad reasons. Nonetheless, we can consider the merits of such a move. Here I will restrict myself to one type of payoff.

Informational and representational hygiene regimes in Earth Rover-class agents must be selected in view of feasibility constraints; one does not solve a problem by delegating a task to a device that it is unable to perform. Human beings are quite limited in their native abilities, and these limitations no doubt go a good deal of the way towards explaining what those hygiene requirements have been. For instance, Kantians like to think that our intentions are allowable only if they can be certified by the so-called CI-procedure, which they think of as a test of practical consistency; the most obvious reason that we do not in fact impose the requirement is that in most cases we are unable to say what would happen if, as lay people put it, everybody did that.[37] Or again for instance, Bayesian epistemologists like the idea of having 'credences,' that is, of assigning probabilities to propositions, rather than believing them outright; they want us to ensure that our probability assignments are consistent, and they want us to update them on the basis of Bayes' Theorem. The most obvious reason that we do not in fact force these demands on people is that no one could actually live up to them.[38]

---

used to determine the student's honors status, his eligibility for privileges and so on.

Foucault made a big deal of the oppressiveness of panopticism, and no doubt my students feel that way, but in fact the reach of the technique is limited by its administrative overhead. *Someone* has to update all those entries in the file; if anything is going to be done on the basis of that file, *someone* has to do it. And every institution has personnel shortages.

The remarkable innovation of Google panopticism is that monitoring and followup no longer require human intervention. Many very small decisions, each tailored to the details of an individual's track record, can be taken without spending the time of a human administrator.

[36] *Was* it accepted? (That is, did the public understand the bargain they were making?) That's become a politicized question, but I can vouch for the sophisticates—I mean, computer scientists, programmers and electrical engineers, who *did* understand the bargain—having gone along with it in just this way.

[37] For a more precise rendering of the CI-procedure, see Millgram, 2005, pp. 90f, 141f.

[38] See above, ch. 2, n. 25, and below, sec. 4.6.

One final example: Lewisians believe that we should understand our conversations to involve a shared but invisible scoreboard; the semantics of our utterances are shaped by the constantly changing score.[39] (Just to convey the general idea, the scoreboard is supposed to include a vagueness parameter; if I say something that's *pretty* vague, that parameter resets to a value in the 'pretty vague' range, which in turn determines how a subsequent utterance, by me or by my conversation partner, is assessed: the subsequent utterance comes out true if it's true when it's also understood to be pretty vague.) We do engage in this sort of conversational practice on occasion, but we are not prudent to do so when anything important is at stake. Because we cannot keep track of the scoreboards we cannot see, conversation partners cannot verify that they understand the course of the conversation the same way; when the back-and-forth is not just friendly patter, conversational scorekeeping is a discourse management method beloved of shysters.[40] The most obvious reason we do not in fact require Lewis-style conversational scorekeeping is that we just can't do it properly.

However, if third-party contractors can perform monitoring and correction that unassisted humans could not, we are in a position to consider imposing more ambitious hygiene regimes. I have mentioned that we are not very good at anticipating the uses that will be made of technological opportunities, and so I won't try to say what I think those will be. Instead, by way of conveying a sense of what such changes might look like, let me gesture first at a recently imposed hygiene regime of this generic type that is familiar; after that, I will rehearse Bernard Williams's description of another much older shift in administrative requirements.

We could require that all information kept on your personal computer be associated with some node of a singly-linked graph, and also that each node in the graph bear a mnemonic label. That is just the organization of data in a directory tree; rather than leaving compliance up to users, it has long been enforced by their devices' operating systems. Much of that task has been recently assumed by applications and rendered invisible to users, who often do not know how their files are organized. Demands of the kind I am envisioning might well have this look and feel to their users, but need not be restricted to the superficial organization of information.

As a illustration of just how deep such changes to our collective hygiene regime can go, I am going to borrow Williams's account of how what

[39] Lewis, 1983.

[40] Millgram, 2009a, sec. 7.7, explains why; for now, when your cognitive capacities are swamped by the demands of a task, predatory interlocutors are bound to take advantage of that fact, and of you if they succeed.

we might as well call the logic of time underwent a dramatic revision between Herodotus and Thucydides.[41] Before the transition, one could make claims about past events without committing oneself to there being a determinate amount of time that had elapsed between the past event and the present: when you said that once upon a time, Minos ruled the waves, you had not yet said anything that entailed that there existed a n such that, n years ago, Minos ruled the waves. (Yes, there was a time when he did, but no *particular* time.) Because the way of thinking is so alien to contemporary sensibilities, it is worth emphasizing that what Williams is exhuming is not the thought that you might not *know* how long ago a past event happened; *we* have *that* thought. It is, rather, a logic of time less demanding than our own.

By the time we reach Thucydides, we are on familiar conceptual ground. To say that there was no time in particular at which an allegedly past event happened meant that it had not really happened at all. "Once upon a time" came to mean that it happened, not in the past, but in a fairy tale, and Williams nicely recovers the sense of bewildered outrage on the part of an older generation unable to fathom why the youth of that today were so dismissive of what were suddenly merely 'myths'.

Now notice that this change would not have had much in the way of a practical point (and probably would not even have been feasible) if not for devices like calendars, that is, devices that allow one to record and calculate with temporal location. (If that sounds too hifalutin: calendars allow you to keep track of and to count the days between events.) When such devices become routinely available (and I want to stress the 'routinely'), what we can think of as the countback requirement for claims about the past became a social option.

That last example was meant to remind you that these hygiene rules are our *logic*. No doubt there are costs to surrendering much of our privacy—both what ordinary people have meant by that term, and the logical privacy of interest to philosophers—but I will not consider them here. What we stand to gain is the opportunity to reconsider what the laws of logic are, to ratchet up their demands, and so—if those demands are well-chosen—to equip ourselves, or rather, our minimal successors, to cope with more challenging environments than our own.

[41] Williams, 2002, ch. 7. I'm not going to weigh in myself on whether Williams has the history right.

Interlude: Emerging Methodological Questions

It's time to pause, collect our thoughts, and look back at what we've covered up to this point. And there is another and perhaps more important task, one which necessitates a temporary shift in format. Both the theses I've been advancing and the stances I've adopted to move them forward bring further issues—and larger questions—into view. These need to be fully acknowledged, if a thoughtful reader is going to be kept on board, even where aiming for the sort of argumentation and closure characteristic of a self-contained essay would be out of order. Following out the train of thought I'm developing here will entail at any rate a willingness to press and hold the mute button on a number of presuppositions of some of the more familiar philosophical methods. You may or may not be ready to sign on to the alternatives, and again, this isn't the place to make the case for them, but here I'll be doing my best to be up front with you about what they are.

3.7.1

Let's take stock. I've asked you to suppose that metaphysics is intellectual ergonomics: that is, to a first approximation, those mysterious invisible objects which make up the apparent subject matter of the philosophical subdiscipline are to be thought of as virtual overlays, that is, accessories facilitating one or another technique of inference or reasoning. I know it's a lot to take on board, all at once and at the outset; it's not the usual way of thinking about metaphysics. What I've been calling old-school metaphysics is a pursuit typically treated by its practitioners as attempting to tell you what the world is *really* composed of, in some way that physics won't, and if, like me, you balk at that, the methodological replacement is gradually getting filled in and motivated; I'm hoping you'll allow it in the meantime.

And allow also that selves belong to the subject matter of metaphysics. The self is certainly one of its traditional topics, and if we count older discussion of the soul, it turns up in philosophy as early as Plato and Aristotle. What's more, it has that familiar look and feel: you can't see a self, even though everybody is supposed to have one (*exactly* one: selves come only in integer quantities), even though selves are supposed to be where the thoughts and concerns reside, even though the moment you no longer have one is when you have died.

If both of those suppositions are active, we should be trying to make sense of the self as intellectual ergonomics: that is, to account for it as apparatus that is there to support an inferential technique, or a family of them. Our first task was to say what technique (or sheaf of them) that would be, and what role selves would play in it. There's a slightly antique way of talking, on which logic articulates the Laws of Thought; on this conception of logic, its job is to lay down guidelines for conducting an activity, namely, reasoning, correctly. That prompted us to ask what selves and logic have to do with each other, and we started to explore this connection: that we can, and to some extent do, select the Laws of Thought in view of the selves we have, and we can, and to some extent do, select the selves to have, keeping in mind the Laws of Thought they enable.

How can that be? Selves, I proposed, are in the first place scopes for the application of consistency regimes. Let's add in a new illustration or two. If I am trying to make sure our candidate wins a seat on the city council, and I am also trying to sabotage his election, something has gone wrong; I am, as the Kantian moral philosophers tell us, practically inconsistent; I am irrational; I should revise my plans so that they are no longer mutually frustrating.[1] But if I am trying to make sure our candidate wins a seat, and *you* are trying to prevent that, no one is inconsistent, anyway as far as that goes, and no one is thereby irrational; we are merely political opponents. If you think that you shall never see a poem as lovely as a tree, but you also think that Larkin's gesture at Donne is just that lovely, you are contradicting yourself, and violating the demands of consistency; you should change your mind about one or the other of them, and maybe both. Whereas if *I* think that poems never live up to trees, and *you* think that "Poem about Oxford" does, that is merely a difference of opinion: perhaps one of us must be wrong, but neither of us, so far, needs to change his mind to avoid being charged with irrationality. What selves are *for* is to serve as domains of enforcement of inferential consistency requirements.

[1] For the New Kantian account of practical consistency, see O'Neill, 1989.

Consistency regimes mandate inferences: If you count as inconsistent when you agree that p, and that if p, then q, but so far don't agree that q, then from what you've agreed to, you're required to infer q, or drop your insistence on what now look like q's premises. Or at least that's what you have to do, once you get called on it. Evidently, one of the big payoffs of imposing consistency regimes is that they make you draw conclusions, and so, help you figure things out. And it's pretty clear that consistency regimes require scopes: if what they cover hasn't been somehow demarcated, there's no answer to the question of what needs to be consistent with what.[2] So it looks like we have selves in the first place to promote figuring things out: that's what accounts for this ergonomic device.[3]

3.7.2

Can we be considering what a self (or even a person) really is? I've introduced the self as a virtual object, one that in the first place serves as a scope for various consistency requirements. *You* are supposed to administer and comply with those requirements, by adjusting your beliefs, but also your plans, maybe your values, perhaps much else. And by living up to the demands on being a self, you become, and sustain yourself, *as* one; if you fail badly enough at complying with those demands, others will simply stop making them.

---

[2] Pretty clear, or so it seems to me, but you ought to be aware that there are other ways to see the issue: e.g., Newton, 2014.

[3] Do consistency regimes *require* what Aristotle called *separability*—there being edges, places where selves stop—or are there other ways of implementing them? We can imagine a boundary-free technique, in which a conception of distance (just for instance, it might be how closely two people are related) is introduced to play an analogous role: the 'closer' two representations are, the more pressure there is to rectify them, if they violate a consistency constraint. (So, in the example, I would be under a great deal of pressure to resolve disagreements with a sibling, less when it's a second cousin, and scarcely any, for passersby on the street.) And you might be wondering whether something along these lines characterizes our diachronic consistency requirements, at least in part. If yesterday I asserted that p, and today I insist that $\sim p$, you may reasonably ask why I changed my mind; but if, when I was five, I asserted that p, and today I insist that $\sim p$, I will respond to your query with a shrug: "I was *five*." I'll return to this suggestion in ch. 7.

While we're at it, we should register that items other than selves or persons serve as scopes for consistency regimes; we often enough need separability in the objects we want to reason about, as a precondition for requiring consistency of description, and the demarcation of selves appears to be a special case of a deeper and broader feature of consistency regimes, that one be able to say where and how they give out. For a scope demarcation that illustrates this point, see the discussion of Aristotle's substances in sec. 8.3.

But how can this make sense? *Who* is supposed to be getting help in figuring things out, if not the selves that are there already? If *I* am the one who is responsible for knowing what I think, for making changes when I violate a consistency constraint, and so on, the self has not been discharged. Have we lapsed into what Nietzsche calls out as 'redoubling', where we imagine an invisible copy of you hovering behind your mind and operating it?[4] But if we haven't, isn't the account viciously circular?

Here an analogy may help us get past the worry. Consider a relevantly similar account of the state, familiar in political science, that explains the state as the organization that, within the territory it controls, has a monopoly on violence.[5] Now, that's the first-pass version of the idea, as you'll hear it recited by students; but of course it cannot be correct as it stands. In any actual state, over and above drunk men picking fistfights in bars and the like, you will find, just for instance, criminal organizations resorting to violence as part of their stock in trade. And so its originator had already corrected the abbreviated slogan to this: that the state has a monopoly on *legitimate* violence. And in the world of modern states, "legitimate" in this connection *means*: authorized or approved by the state.

Thus in this account of statehood, the state hasn't been discharged, and that in more than one respect. (E.g., in which territory does the state have a monopoly on violence? The one claimed and controlled by that state.) This is a circularity that has very much the shape of the circularity we saw built into the self. So neither account can serve a reduction, and reductions do seem to be the mode of explication that many of us philosophers prefer. Nonetheless, Weber's account is certainly informative and even insightful—which is why it is so familiar. Likewise, this explanation of selves can hope to be on target, informative, and even revealing.

A political scientist who has adopted Weber's perspective is no doubt committed to there being an account of how a state so conceived can operate, such that the conceptual circularity does not unravel its decisionmaking processes. This account of selves likewise is committed to there being a theory of self-governing systems in the offing, on which systems can represent themselves in the course of making decisions, and those decisions will determine what such a system then does. We have been seeing treatments of this kind at regularly increasing levels of granularity.[6] Because they are now

[4] *Genealogy of Morals*, Essay I, sec. 13, in Nietzsche, 2000, at pp. 480–82; in an unpublished paper, Candace Vogler once caricatured the mistake as thinking of yourself as an 'angel in disguise'.

[5] Weber, 2004, pp. 33, 38.

[6] I have in mind especially Dennett, 1991, Dennett, 2004, Ismael, 2007, Ismael, 2016,

available, and because a recap would take more time than it's reasonable to ask of a reader, I've kept to what we need for the argument at hand.

There is, however, a subsidiary issue, that of sorting out legacy terminology: we're in the course of making technical vocabulary out of a handful of familiar words, and let's compile the conventions we've introduced so far. A self is a virtual object, overlaid on a human being, whose bottom-level task is to serve as a scope for consistency requirements. The type of requirements I have in mind have to be met *at a time*: if you believe that p, *now*, and also believe that $\sim p$, *now*, you're in violation of one particularly prominent such requirement. (And the five-dollar-word way of saying it: these requirements are *synchronic*.) I'll save discussion of how we are to manage selves over time for ch. 7, and when I get there, I'll reserve the term "person" for discussing selves seen as extended over time (that is, looked at *diachronically*).

In the meantime, I've reallocated the ordinary terms "mind" and "mental," restricting them to just what can be covered by the sort of consistency requirements we're considering. Because, as we've seen, consistency requirements are, as a rule of thumb, usefully enforced only when you know what it is you think, want and so on, if you like, we're retrieving what was right about the Cartesian idea that your mind is transparent to you: in the way we're using these terms, unconscious mental states will be anomalous. And we're making "cognition," "cognitive," and so on the go-tos for all the computing and data processing you do, whether or not you're aware of it, and whether or not we ask you to ensure consistency for it.

You might already be incredulous. The self is a device, and you might have registered an apparent implicature: that I'm telling you that *you* are a device. But: my self is *me*, and surely *I'm* not just an artefact of some technique or methodology! Isn't there much more to me than the aspects of my intellectual life that are subject to consistency requirements? What about my hopes and dreams? What about the sea of experience in which I swim? If death were just the cessation of those persistent demands for consistency, wouldn't I look forward to it with relief, rather than, as seems to be the norm, dread or, at best, grim readiness?

Now we can assuage that worry, anyway somewhat. There *is* a great deal of cognition for which we don't insist on consistency: your daydreams, for instance. (We've also mentioned perception and memory, as categories of cognition where we have one-way consistency requirements; you're supposed to keep your beliefs consistent with what you see and what you recall, but we

---

Part I, and Huebner, 2014, esp. pp. 154–60.

don't ask you to change what you see and remember to match your opinions, and charge you with inconsistency if you don't.) And there is in any case a great deal more to you than the parts of you that do information processing. The cost to adopting this manner of speaking is that we have to be careful about saying that you are yourself, even though "yourself" is, grammatically, a reflexive pronoun. It's a slightly awkward terminological regimentation, but you can keep track of it this way: the relationship between you and your self is rather like that between France and the French state: it is the State *of* France.

The state is a social device. Not every country or nation is clothed in a state, and states as we know them aren't found at all times or in all places. If the self is a device, is it going to follow that some people don't have selves? That selfhood is a prerogative of some cultures and not others? Isn't that a reprehensible thing to find oneself saying? Call this the *purity test* worry.

Precisely because selves *are* devices, using the locution this way is not to disparage those who haven't adopted it. A technique—*any* technique—will be found suitable for some purposes and not others, and by some people and not others. We've already started to consider the possibility of exchanging that device for another. We do owe it to ourselves to consider the connections we have made between the deployment of the self and our practices of assigning moral standing: that is, roughly, of who we take moral requirements to be addressed to, and who is protected by the moral requirements we're bound by.[7] For right now, the purity test concern can serve as a reminder that the consistency requirements we impose take many different forms. Sometimes the requirement is that the various things you think be kept jointly consistent; but at other times, it is that you remain consistent with one or another publicly designated position.

3.7.3

In both of the preceding chapters, your responsibility for maintaining various sorts of consistency has come up, and you may quite reasonably be wanting more in the way of explanation. I'm going to demur when it comes to one sort of direction I might be invited to take, that is, to provide an account of what gets called—but not by me—"normativity".[8] Saying what it is to be held

[7] That's necessarily left for a subsequent treatment of practical, moral, and social followons, but for a first pass over the topic, see Millgram, 2009c.

[8] That's a word I don't use in *propria persona*, for reasons covered in Millgram, 2015, ch. 5.

responsible requires shifting the discussion both to the social environment in which selves call other selves out for violating such demands, and to a straight-on discussion of practical reasoning; as I indicated early on, those topics are to be taken up in a sequel.

But I do want to speak to a concern about a special case of responsibility—*logical* responsibility—that has already made its appearance. We have inferential consistency requirements, where, for instance, if you believe that p, and that if p then q, but deny q, you're being inconsistent. And we also have, just for instance, the sort of consistency I demand of my students, when they format their bibliographies. (Pick one bibliography style, I say, and stick to it consistently.) We distinguish between them, and treat logical consistency as having a very special status. In the present approach, how should we be thinking about that distinction?

The first thing to notice is that we're quite comfortable with it. Students in my philosophy department, as in many others, have to satisfy a logic requirement, but they can petition out of it, by documenting that they've taken a suitable course elsewhere. If the course they've taken, maybe from a mathematics department, covers model theory and recursion theory, we grant the petition: that's logic. If it's number theory, we don't: that's not logic. And if it's theory of computation, we ask to see the syllabus: maybe there's enough logic for an approval? But *why* is number theory not logic? What distinction are we making?

It's not just a matter of tacit professional norms that have somehow been internalized; you can elicit the distinction we're after from just about anyone you might stop on the sidewalk, and do so, normally, in under five minutes. It's the difference between believing something because you can see, maybe on the basis of evidence, that it's true, and believing it because it's convenient: because it lets you fit in, or because otherwise it would be too painful to think about what you really do in your job, or because otherwise you would have to have a difficult conversation with your spouse.[9] That is, you can believe, as we might say it tersely, for the reasons that logic sets you—or the reasons that real or misguided prudence sets.

And I have to say that, surprising as it is, I've never come across an account of the distinction that wasn't somehow circular, but also seemed to get the extension of the logical more or less right. (Often the attempts turn

[9] And these sorts of cases have come in for philosophical discussion in a couple of rounds of the literature, which you can find via their respective keyphases: "the ethics of belief," and "the wrong kind of reasons". Williams, 1973b, ch. 9, is a good entry point into the problem space. See also Meiland, 1980, Millgram, 1997, ch. 2, Nguyen, 2019, pp. 451-55, Hieronymi, 2005.

on the notion that logic takes you from truths to truths, but a great deal of inference takes you rather from objectives to choices, or from vaguely or kind-of true claims to other vaguely or kind-of true claims; and often the attempts invoke the exceptionlessness of specifically deductive logic, but as we are about to notice in coming chapters, a great deal of inference is defeasible.[10]) I confess that I'm in the same boat as everyone else: I don't know how it is that I differentiate logical and nonlogical consistency requirements. So instead I'll say something about how, on the methodological approach we're taking, you would go about trying to figure it out.

Herbert Simon once discussed

> a simplified... organism [which in a footnote we're invited to imagine as a "mechanical turtle"] that has a single need—food—and is capable of three kinds of activity: resting, exploration, and food getting... The organism's life space may be described as a surface over which it can locomote. Most of the surface is perfectly bare, but at isolated, widely scattered points there are little heaps of food [for the mechanical turtle, I suppose these are batteries], each adequate for a meal.
>
> The organism's vision permits it to see, at any moment, a circular portion of the surface about the point in which it is standing. It is able to move at some fixed maximum rate over the surface. It metabolizes at a given average rate and is able to store a certain amount of food energy, so that it needs to eat a meal at certain average intervals. It has the capacity, once it sees a food heap, to proceed toward it at the maximum rate of locomotion. The problem of rational choice is to choose its path in such a way that it will not starve. (Simon, 1957, p. 262)

Simon goes on to show how to solve the problem he has set, but notice how he has described it: as his turtle's "problem of rational choice". When an 'organism' is as simple as this, the distinction between what is demanded by

[10]Sometimes logic is identified by way of what are meant to be necessary rather than sufficient conditions; most notably, that logic is formal. That's probably not a good frame of mind in which to approach defeasibility, but in any case John MacFarlane's doctoral dissertation (2000) has made the pitfalls of this sort of analysis apparent. He observes that the key concept in that explication, "formal," has meant different things over the history of the "logic is formal" claim. Different conceptions of "formal" have different assumptions built into them; not all of them are still intellectually available. For instance, you can mean by it what Kant meant, but then you have to be willing to sign on to his transcendental idealism, and that's no longer a live option.

reason or logic and what is demanded by prudence slides off its back; when it's a mechanical turtle, we just *give* Simon his turn of phrase. There is no use, in a 'life' as stripped down as this one, in asking what the turtle's logic is, and distinguishing that from what it's convenient, from the point of view of its designer, to have it do.

Now imagine embarking on a series of creature constructions. And imagine that somewhere, in the long journey from the mechanical turtle, through ever more complex organisms or robots, and on our way to human beings, we come upon a first creature for which such a distinction is functionally required. That step is where we ought to look, to find out what we meant by it.

3.7.4

By this point in laying out the present metaphysics of the self, readers with a traditional philosophical background may have experienced an upwelling of resistance, both to the methodology and to the picture being put in place. Your mind—in an older way of speaking, your very *soul*—is a reservoir of the most intimate certainties and of uniquely personal value; its workings are a mystery so deep that Descartes found himself needing to imagine an immaterial substance in which to house them. And here it is being presented as something on the order of an information storage construct: a designation that determines which entries are going to crosschecked against which others. The problems we are taking up seem to be centered on the practicalities of crosschecking procedures and database management. It's as though we're being asked to believe that the answers to questions that have moved two and half millennia of metaphysicians are going to be found in the *philosophy of filing*.

Properly disarming that response has to be a piece-by-piece matter of arguments and analyses, and that's something we're in the middle of doing. But right now I want to speak directly to the *impulse*, by invoking a half-forgotten work of literary criticism.

Widely read in the third quarter of the twentieth century, Erich Auerbach's *Mimesis* was a reconstruction of the trajectory of high literature in Europe, from the Greeks to *Madame Bovary*. At outset of that history, subject matter for serious treatment is exhibited as narrowly confined, in the first place to the doings of royalty and divinities. Gradually, however—and Auerbach shows us why the series of literary accomplishments should not be underestimated—writers learned to give the cold-sober, honest attention that Sophoclean tragedy had commanded to ever more mundane scenes.

Although the narrative wraps up with Flaubert's bored small-town housewife, and the troubles she lands herself in, there doesn't seem ever to be a stopping point to the progress of realism; as far as anyone can tell, there is always room to take on yet another aspect of life hitherto deemed too lowly to deserve or tolerate careful, respectful, and unflinching artistic treatment.

While I don't know of a work that does for philosophy what Auerbach did for literature, we can see easily enough how philosophy has been traversing a parallel arc. Recall how Plato dismissed the real world as mere becoming, clearly enough because it seemed intellectually intractable to him; it was only the Forms, abstract objects such as Beauty Itself, that could be objects of genuine philosophical understanding. We read Aristotle, coming one generation later, for his treatment of excellence in the engaged civic life, and for his conceptualization of biological organisms: that is, for philosophizing whose subject matter was one large step closer to real life. (For him, however, all of that was second-best: the ethically proper aspiration was completely disengaged contemplation, and Stephen Menn has been making a plausible case that the parts of his metaphysics that mattered most to him were the ones that we nowadays largely ignore: his unmoved mover—that is, God—and the planets trying to imitate Him.[11])

As in serious literature, we have come a long way, step by arduous step. There was Descartes's attention to the inanimate mechanical world, and then Hume's turn to human psychology and what can and can't be pictured. There was Kant asking how much of the apparent structure of the world comes from us, and where our ability to reason about it gives out. There were the pragmatists, wanting to know what difference a claim would make if it were true. There was Wittgenstein, asking what we have to be able to take for granted about our yardsticks and other measuring devices if we are not going to give up on using them. And without enumerating the many steps in between and since, we can nonetheless say that, as in literature, there is no last step. Here I mean to be taking the next of them, and if the philosophy of filing seems too lowly to be *philosophy*, that is a sign that we are still withholding our philosophical attention from some of the matter of human life. I am making one more of those attempts to direct our vision to what is right there in front of us, and to sandblast away a bit more of the residuum of metaphysical euphemism. But, to reiterate, there is no final step.

[11] On the former observation, see Richardson Lear, 2004.

3.7.5

Can this first use of consistency regimes—prodding people to figure things out, by requiring inferences of them—possibly work as advertised? Our illustrations of inconsistencies for which people can get called out have so far been cut-and-dried: the inferences in play were deductive, or deductive for all practical purposes. When inference is deductive, the premises guarantee the conclusion. But most of our reasoning is defeasible, that is to say, even when the inference is perfectly satisfactory, and the conclusion follows, you could encounter *defeaters*—learn some new fact, or accept an additional assessment or evaluation—that would leave the premises intact, but require you to withdraw the conclusion. Defeasible inferences reflect softer consistency regimes, and they underwrite inferential mandates that seem to have slack in them; it's not obvious how they are to be administered.

To get a sense of the problem, imagine that the legal advisor to the building committee is tasked with enforcing a particular consistency regime for the municipality's inspectors. The building code says that permanent structures may not be erected in the lot's designated green area, but the inspector, who has been bribed, does not report a violation: he claims that the added structure is a disability accommodation, which overrides. When the neighbors document that no one in the first-floor unit is disabled, the inspector invokes a variance; when, after a lawsuit, the variance is disallowed, the outbuilding is said to be covered by a religious-use exemption... and it gradually becomes apparent, as is characteristic in defeasible inference, that the list of issues to which one could appeal never, ever runs out. The legal advisor's task is endless, and she herself all too easily rendered ineffectual.

The problem isn't confined to matters of choice and action; it turns up when we're simply trying to determine the facts. The alarm has gone off, my phone informs me. If the alarm has gone off, an intruder is in the house. So an intruder is in the house. (And so—here's a further and practical inference to be drawn from that conclusion—an alarm service representative, and if the matter isn't resolved, a police officer, had better check on the situation. From factual conclusions, we draw practical consequences.) That was the right inference rule to work with, but there are indefinitely many circumstances under which we'd have to take back the conclusion: not if the battery in the key fob that disarms the system is dead; and not if there are dust bunnies inside the circuitry; and not if the houseguest has forgotten the code; and not if a branch has bumped the loose window; and not if hackers are harassing residents by triggering alarms remotely... It's evident, but you might want to pause and try it out, that with a little imagination you could

extend this list of exceptions as far out as you like. (I asked a policewoman how they respond to these alarm notifications, and she answered, perhaps unsurprisingly, "We don't.") Almost all of our inferences go through only *ceteris paribus*, or 'other things equal,' and when you look inside that other-things-equal clause, to see what's hidden there, you don't come to a place where it gives out: there are always more exceptions-in-waiting.

If the possible exceptions to consistency requirements are inexhaustible, how do you determine that no one of them is pertinent—even when all the players are acting in good faith, or even when it is just you trying to get things right, all on your own? It is not as though you can simply page through the checklist; no matter how far down you get, there are always more issues you have perhaps overlooked. If selves—and also, getting ahead of ourselves, *persons*—are administrative units that are there to underwrite consistency regimes, how are they supposed to *work*?

It's time to turn to this problem, to spell it out carefully, and to get to the bottom of it. That task will occupy us over the coming two chapters.

Chapter 4

TBA

Most inference is *defeasible*: that is, even when the reasoning is fine as it is, the conclusion follows only other things equal (*ceteris paribus*, in philosophers' Latin): there could always be further facts or assessments which, added to the list of the inference's premises, would defeat it.

For various reasons, I'm unhappy with attempts to make sense of the logic of defeasible inference using treatments that model themselves on the more familiar formalizations of deductive logic.[1] And so here I want to take a step back, and ask *why* so much of our reasoning is (and here's the term that's used to mark defeasibility in some other fields) nonmonotonic. With any luck, understanding what defeasibility is doing there will prepare us to

[1] The term was imported into the philosophical vocabulary by Hart, 1968, who, however, took defeasibility to characterize ascriptions of responsibility and the like, rather than straightforward, vanilla inference; Pollock, 1987, describes an early attempt at implementing a defeasible reasoning system. For recent and representative discussions, see Hlobil, 2018, and Horty, 2012; Reutlinger *et al.*, 2017, surveys the topic as it is approached by philosophers of science and causation. For an older approach, see Adams, 1966.

The awkwardness of fit between the usual approach and the subject matter has been remarked elsewhere:

> in deductive reasoning, showing that a conclusion follows from a set of premises involves checking that the conclusion is true in all possible models in which the premises are true. However, in the case of non-monotonic reasoning it will be possible, by definition, for the conclusion to be false while the premises are all true. After all, in such reasoning, inferences are provisional, and conclusions may have to be retracted in the light of further information. Thus an exhaustive search for counterexamples for any non-deductive inference will inevitably be successful and no inferences will be licensed. Accordingly, it appears that, far from being readily extendible to commonsense inference, semantic methods of proof are fundamentally incompatible with it. (Oaksford and Chater, 1998, pp. 136f)

investigate it in a different and more productive way. To foreshadow, I will explain how it is that we are the sort of creatures for whom the issues with which they have to cope are, very, very often, TBA.

I'll proceed by first reintroducing the phenomenon of defeasibility somewhat less tersely, with the intent of conveying how pervasive it is. Then I'll turn to a description of human beings as boundedly rational agents, forced to live in a world that is larger and messier than they can accurately track within the systems of representations that they construct. I'll describe a cut-down, more limited version of humanity: creatures who in some ways resemble us, but only apply inference rules mechanically. I'll make some observations about how such creatures—creatures to whom all inference seems to be deductive—will cope with a world that outruns the repertoire of their inference rules. And that will permit me to introduce defeasible reasoning as a way of improving those simpler proto-humans. At that point we still won't know *how* defeasible inference works—that will remain an open question—but we will have a clearer picture of what it is doing there.

4.1

Let's begin with a survey of some of the many patterns of defeasible inference; this will help to bring our explanandum into focus.

Inductive inference is traditionally characterized as proceeding from empirical premises either to a future, unobserved particular fact, or to an empirical generalization.[2] Simple-minded inductions-by-enumeration are something of a caricature of the category, but as a representative instance, this one will do:

1. Over the 1980s, the card catalog—a paper resource provided by the library—was replaced with an electronic catalog.

2. In the decade following the turn of the millennium, our journal holdings—another paper resource provided by the library—were replaced with electronic subscriptions.

3. Over this past decade, many books—yet another paper resource provided by the library—have been replaced by electronic access.

   ⋮

   Therefore:

[2]The classic treatment is John Stuart Mill's *A System of Logic* (Mill, 1967–1991, vols. VII–VIII); for a reconstruction of Mill's view, with discussion, see Millgram, 2009b.

4. For any given paper resource held by the library, sooner or later, it will be replaced by an electronic version.

   (Or alternatively:)

   In due course, all of the library's holdings will be electronic.

   (And, sure enough, I have actually heard the entirely virtual library touted as our utopian future.)

Put that way, while it's probably not that strong an argument, it's alright to draw the conclusion; anyway, let's allow it for the sake of our own argument. We're interested in the way the inference can nevertheless be defeated by additional facts, should they prove to obtain:

- The library has Book Arts and Rare Manuscript collections; ownership of the physical items matters.
- Hackers will come to routinely alter the content of a library's electronic holdings, and in that event, only paper holdings will be regarded as secure and reliable.
- The admissions office will determine that prospective students are favorably impressed by seeing shelves upon shelves of traditional, physical books.
  ⋮

It's not just induction; a great deal of theoretical reasoning—that is, reasoning to conclusions about the facts—turns on claims that aren't 100% true but, rather, true *enough*: kind of true, sort of true. (Idealizations and approximations are important classes of description that fall under this heading.) When an inference form (even one that is apparently deductive) invokes partial truths of this kind, the reasoning ends up defeasible. For instance, our library administration reasons roughly as follows:

1. If a library resource is there to be read, it can be read in either print or electronic format.
2. Library resources are there to be read.

   So:

3. Library resources can be read in either print or electronic format.

You'd think that that was a deductively valid argument, plain and simple; but what do we make of it, given that neither of those premises are strictly true? For instance, it's no longer news that libraries acquire audio recordings and films. People sometimes forget that you go to libraries not just to read, but to browse the stacks. I once went to a library in England to look at a scribble on an envelope, and since the aim was determining whose scribble it was, I needed to see the envelope itself, rather than an image of it. The immense piles of source materials that it is the job of historians to read—old newpapers and magazines, century-old tomes and so on—are only readable-in-bulk when they're not microform or online: after enough time spent looking at scans, your eyes and head start to hurt, and you have no choice but to stop. But we know what those library administrators are thinking: that those claims are pretty much true, preponderantly true, true enough to proceed on the basis of that argument.

Because the steps of their argument are only pretty much true, the apparently deductively entailed conclusion only follows other things equal. It doesn't follow when the *way* you read it on-screen doesn't count as carefully reading it, and careful matters.[3] It doesn't follow when features of the printed version—e.g., the fold-out, multicolored chart in William Whewell's *Philosophy of the Inductive Sciences*, displaying the history of 'colligations' that brought science forward to his day—can't be reproduced in a widely available digital format. It doesn't follow when the reason you're inspecting the book is to determine how it would have struck a nineteenth-century reader first picking it up. And not when the reason you tracked it down was to find out whether the pages had ever been cut.[4]

But the reach of defeasibility extends further still, since not all reasoning is about how the facts stand. Practical reasoning is about what to do; means-end inferences are usually taken to be the most central, and absolutely indispensable, building blocks of practical arguments.[5] And means-end inference is defeasible, too; let's build out our example-in-progress a little more.

1. We need to create an enormous, plush Welcome Center, in the most imposing building on campus—that is, the library—in order to attract prospective students.

[3] Cf., e.g., Mangen *et al.*, 2013.

[4] The claims about partial truths and defeasibility are drawn from Millgram, 2009a; the fold-out diagram can be found in Whewell, 1847, adjacent to p. 118.

[5] Vogler, 2002, develops the strongest arguments for the indispensability of means-end reasons of which I am aware. Millgram, 2015, sec. 6.2, briefly surveys other forms of defeasible practical reasoning.

2. In order to do that, we need to free up space by deaccessioning books and other paper resources.

3. Converting paper holdings to electronic form will allow us to deaccession the paper versions.

   Therefore:

4. We will convert the library's paper holdings to electronic form.

Once again, the inference can be defeated, both by additional facts, and by additional evaluative considerations:

- The electronic resources are almost all subscription products; if we have to cut the library budget, say, during a severe recession, the all-electronic library will suddenly be *empty*.
- The electronic resources are almost all hosted elsewhere; if the vendors go out of business, the books will vanish.
- Attempting to boost enrollments is a mistaken response to our current funding and budgeting model, which should be revised.
- Books can be stored off-site, which allows space to be freed up without getting rid of them.
  ⋮

Even theoretical and practical reasoning taken together don't exhaust the scope of defeasibility. Over and above inferences proper, a great many judgments that encode or underwrite inferences are also defeasible. For instance, consider a counterfactual like this one:

> If we had converted library resources from paper to electronic a generation ago, by now some large proportion of them would no longer be readable.[6]

That's probably true—if you're of that generation, can *you* still open the WordStar and WordPerfect documents on your ancient five-inch floppies?—but defeasible. After all, had there been a major and well-funded initiative

[6] Nicely illustrated in xkcd.com/1909—and if you're reading this a few years down the road, now would be a good time to check: can you follow the link and see the graphic? If you can, can you also see the mouseover text?

to ensure forward and backward compatibility in library assets, 1980s-era files would likely still be usable.

As the metaphysics cognoscenti are aware (but if this is too exotic to make sense to you, don't worry about it), semantics for counterfactuals try to accommodate this sort of defeater by appealing to how similar one way things might have been (one 'possible world') is to another. In this case, a possible world in which there's no compatibility initiative is more similar—'closer'—to the way things actually are than a world in which there is, and that's why we can ignore such initiatives, when we decide whether the counterfactual is true. So notice that the similarity judgments themselves are defeasible also: if we add in the information that there was no compatibility initiative because it was vetoed by our space-alien overlords, we're suddenly considering a possibility that's much *less* similar to the way things actually are.[7]

The long and short of it is that defeasibility looks like it's just about *everywhere*: in our inductive reasoning, when we deploy approximations or idealizations, when we reason about means to our ends, and even when we ask what would have happened, if...

And notice, looking back at our examples, that they share a striking and apparently logical feature, namely, that defeaters don't run out; no matter how many ifs, ands, and buts we've put on our checklist, it's always possible (and even straightforward) to think of more. If you're not sure about that, let me ask you to pause and experiment: try making up a few additions to each of the lists we've started. See? There's always an ellipsis—that "..."—at the end.

Defeasible inference is contrasted with deductive inference, where the truth of the premises guarantees the truth of the conclusion, no ifs, ands, or buts. Maybe we *sometimes* have occasion to reason deductively, but it's evidently unusual enough for those circumstances to amount to a special case, one that requires a special explanation. The default is defeasibility.[8]

[7] The points about the defeasibility of counterfactuals and similarity relations are condensed from Millgram, 2009a, ch. 11, and Millgram, 2015, ch. 7.

[8] Millgram, 2009a, ch. 2, takes a shot at demarcating that special case. However, for our purposes, perhaps there's even less to the special case than meets the eye. First of all, the home territory of deductive reasoning is mathematics, which seems to have as its subject matter abstract objects that we use as models for real-world phenomena. When we do use them that way, the defeasibility generally reappears as we consider whether the mathematical structure is a satisfactory model: whether we can deploy it to draw conclusions about the real world. For examples, see Cartwright, 1983, or Millgram, 2015, ch. 8.

Secondly, we have ways of folding inferential defeasibility into concepts, in ways that

But why do we think this way? You might well suppose that there would be fairly weighty reasons not to do so. How do you know whether to go ahead and draw a conclusion, when you have to check for defeaters, but there are indefinitely many potential defeaters, and so you can't? Don't *ceteris paribus* clauses trivialize your claims, giving them the form, "Such and such is the case, except when it's not"? And doesn't the openendedness of the list of defeaters that unpacks one of those *ceteris paribus* clauses mean that there actually isn't a standard of correctness for any claim or inference that includes them?[9]

Maybe those problems are surmountable—or maybe not. However, since we don't want to surmount obstacles unnecessarily, we wouldn't reason this way unless we had to. *Why* do we have to?

4.2

Human beings encounter and know about only a very small patch of the world; unlike the theologians' omniscient God, we have, so to speak, a drastically restricted field of vision. Moreover, human beings are boundedly rational: because we have limited computational capacities, small-finite memories and so on, we're not able to figure out all the things it should be possible in principle to figure out, given what we already know. When we can't, we rely on heuristics, that is, rule-of-thumb shortcuts that give us a good-enough answer (we hope) enough of the time, but which will perform suboptimally in some classes of cases (hopefully, less important ones).[10]

---

make the inferences that deploy them seem indefeasible. For instance, murder is absolutely prohibited—because if there's an acceptable reason to kill someone (self-defense, legal executions, a military conflict...), it's not counted as murder. Or again, in the world of Newtonian mechanics, $F = ma$ is indefeasible. But when you apply a specific force, one that you control, to a given mass, there are indefinitely many reasons it may not accelerate in the way the equation predicts; all that the First Law requires is that those defeaters for your inference be represented as forces as well. In cases like these, the in-practice defeasibility is hidden away within the application of the concept underwriting the apparently indefeasible inferences.

[9] I take these to be the issues moving Searle, 2005, esp. secs. 1–4, although he understands himself, mistakenly in my view, to be making out a contrast between theoretical inference and the ways we deliberate about what to do.

[10] The bounded-rationality tradition was launched by Herbert Simon (1957); see Conlisk, 1996, for an overview of the first generation of work in the field; the remarks bridging Bendor, 2010, pp. 29–30, tersely convey how deep the difference between the bounded rationality approach and its decision-theoretic older sibling goes. A look at Morton, 1990, Wimsatt, 2007, and Gigerenzer *et al.*, 2000, will provide something of a sense for how wide the range of forms that work in the field can take is.

We need to pause for a pair of preliminary methodological observations. What approaches are workable, when we're trying to make sense of bounded rationality?

First, these restrictions on our abilities to know what's going on around us, and to make up our minds intelligently as to what to do, are themselves extremely complex. (i) There are always further cognitive capacities we could opt to assess. Just for instance, you can test someone's IQ, or their mechanical reasoning abilities, or their talent for versification, or (as in an exam famously administered to applicants at All Souls) their knack for instantly composing an elegant essay around a one-word prompt. It's clear enough that you could extend this list indefinitely, and consequently there are indefinitely many dimensions along which a somewhat rational agent can be limited. (ii) Moreover, for a given capacity, the ways in which it is restricted are normally hard for someone who is boundedly rational along that very dimension to understand. For example, when someone is bad at mathematical reasoning, it's both hard to make clear to him what he's bad at, and precisely *how* he's bad at it; when someone is a poor writer and bad stylist, usually one part of the problem is that he is unable to recognize bad writing, and fails to understand his own limitations. Analogizing, when we have some cognitive capacity in a stunted version, we shouldn't expect to be able to understand it very well—and in what ways it *is* limited.

It follows that, precisely because we are boundedly rational, we will have to substitute a simplified representation of bounded rationality for the extremely complex reality, if we're going to make any headway on the topic. So the first methodological point is that when we're investigating bounded rationality, we inevitably make do with idealized or simplified or approximate characterizations of boundedly rational agents, rather than a complete and accurate theory. Accordingly, we will have to *choose* our highly simplified representation of a boundedly rational creature's intellect.

Second, the very notion of bounded rationality has to do, on the face of it, with what it takes for an agent with limited cognitive resources to function in one or another environment. For that sort of problem, what Daniel Dennett once called the 'design stance' is appropriate. So I will be asking you to imagine someone trying to design a boundedly rational agent—you can picture this taking place in a robotics lab—and to consider what the tradeoffs look like from his point of view. Now, "agent" has come to be used by philosophers as a synonym for "person," but we can hark back to its older meaning (still present in phrases like "secret agent"): an agent is someone whom you send somewhere else, to do things on your behalf. We've become used to communication that is for all practical purposes instantaneous; in

former times, however, the VOC's agent in Batavia couldn't call up headquarters for instructions; he would have to decide what to do, on his own and on the spot.[11] The problem I want you to be imagining our designer trying to solve is how to equip his robotic agent, or maybe his Frankenstein monster, to do *that*; as you might by now expect, the trajectory of the discussion will take the form of a creature construction argument.[12]

To briefly recapitulate those preliminaries: we are going to be considering a ruthlessly stripped down model of one type of bounded rationality of interest to us, and we are going to be treating it as presenting us with a design exercise.

By way of motivating the simplification I'm going to be working with, we know from experience (and from the laboratory) that people have a hard time wrapping their minds around too many cognitive elements at once.[13] So instead of thinking of the cognitive field of the agent-in-progress as a single, holistically unified field of representation, let's imagine the agent as maintaining a system of representations, which are relatively simple, discrete units; only a few-to-several of these are the focus of attention at any given time. Some of them serve to, more or less, map the world around it; others are lists of less-localized facts, as well as components of more ambitious theories about what is and isn't the case. But we shouldn't construe all of these representations as purporting to 'mirror nature', as Richard Rorty once would have put it; some specify goals or objectives; some encode evaluations or assessments; others may lay down rules. And we can already make a few observations about how this system of representations will have to look.

Obviously, it can't represent more than a tiny (actually, infinitesimally tiny) fraction of what there is to think about; most of the world will be left uncharted, and on the practical side, only the smallest part of the creature's option space will be visible to it. And even what it represents will have to be ruthlessly simplified—with *mis*representation almost certainly the price.

Let's introduce what we can call the agent's *inferences*: moves that add a representation to the system it maintains, on the basis of represen-

[11] The Verenigde Oost-Indische Compagnie, or Dutch East India Company; Batavia was today's Jakarta.

[12] For discussion and models, see the references in ch. 3, n. 1. And for the design stance, see Dennett, 1989, e.g., pp. 16f.

Because adaptationist accounts of evolutionary outcomes are so common, we want to be clear that this is not going to be one more of them. I am not asking how it came about that natural selection produced human beings who reason defeasibly, but rather what it will take to get a certain type of boundedly rational agent to function successfully. I'll return to this point in our second Interlude.

[13] See, e.g. and famously, Miller, 1956.

tations already present in the system; they may also delete or replace a representation already in the system. (Remember, this is a ruthlessly simplified model, so please hold on to your views about what inference *really* is.) Right now, we're going to imagine the agent doing so much more clunkily than we do: these additions, deletions or replacements are driven by inference rules, which are applied mechanically. Just for example, *modus ponens* is executed this way: when the agent registers that it has representations of the form p and $p \supset q$ in its system of representations, it adds q to the system; if $\sim q$ is already in the system, it is deleted.[14]

Insert yourself into the point of view of a designer, who now has to specify rules to drive this agent's inference engine. You will no doubt start out hoping for an in-principle-best repertoire of inference rules, and let's mark two variant approaches. One would be to deploy rules triggered by whatever subset of representations within the system of representations produces what is most likely to be the correct answer. (Call this the *internalist* approach.) Alternatively, you might look for rules triggered by premises that your agent doesn't necessarily have available, but which would most likely produce the correct answer. (That would be the *externalist* approach; it requires the agent to go hunting for missing premises.)

Hopes notwithstanding, you are quickly going to discover that your agent isn't able to use that in-principle-best repertoire of inference rules. On the externalist approach, there's no guarantee that missing premises for an inference rule can be rounded up in a timely or cost-effective manner; remember that agents see only a very small patch of a very large and complex world. And even confining ourselves to the internalist approach, it's very plausible that many of those in-principle-best inference rules would have to be triggered by a sizable number of representations. (I used *modus ponens* as my illustration of an inference rule, but suddenly it looks highly atypical; won't we expect the typical member of the repertoire to require a conclusion drawn from perhaps thousands or even millions of premises?) As we've already noticed, agents that are cognitively limited in the way we are can't be expected to grasp those complex arguments. And in any case, a boundedly rational agent can't be expected to successfully survey its own system of representations, in a way that will capture those inferences: if it first surveys all the individual representations, to match them to inference rules with one premise; next, all pairs of representations, to match them to

[14]Under the heading of simplifications, for now we're excising an aspect of reasoning emphasized by, e.g., Harman, 1986, pp. 5, 11f: that even when applying deductive inference rules, we decide whether to accept the conclusion or reject a premise. Our baseline agents don't, but we'll reintroduce such choices below.

inference rules with two premises; next, all triples... all in all, for n representations, the search it is executing takes $O(2^n)$ operations—that is, it is not computationally tractable.[15]

A smart designer will gracefully acknowledge the problem, and cancel those plans for that in-principle-best repertoire of inference rules; instead, the aim will be to provide a repertoire of *heuristic* inference rules. As with any heuristic, these can't be expected to provide the right answer in any and all circumstances. But if the designer has a good grasp of a circumscribed environment in which his agent will act (let's say, its *niche*), he can cut down those in-principle-best inference rules, deleting the triggers that are not likely to matter in the agent's niche—or more generally, which won't degrade performance too badly when they are ignored.

At this point, the inference rules that the creature deploys have started to look more familiar. Instead of drawing a conclusion from perhaps tens of thousands of premises at a time, the agent is tasked with drawing a conclusion from a small handful of them. (That is, the search it has to perform is now feasible.) On the externalist specification of its procedure, by and large the designer will have specified rules whose missing premises can often enough be ascertained in a timely manner, and without incurring unreasonable costs.

4.3

I laid down earlier on that the agent we're considering applies its inference rules mechanically, which was a way of excluding defeasibility from its logical toolkit. When a heuristic inference rule is triggered by a cluster of representations, the agent simply *does* add the conclusion to the system (or it overwrites some element of the system of representations)—no second thoughts.

But now let's consider how such an agent is going to perform when it's released into a world roughly like our own, and try to envisage its performance from both outside and inside.

When our designer selected the agent's heuristic inference rules, he was optimizing for performance within a particular niche; recall that the point was to pare down the number of premises that a rule invokes, by discarding triggers that likely wouldn't matter in the (controlled, well-under-

[15]The observation is adapted from the intriguing argument in Cherniak, 1986, esp. pp. 49–52, 57, 61–70, 127–129, to the effect that you can't build individual agents that are arbitrarily intelligent; see ch. 7, note 8.

stood) circumstances that the agent is anticipated to encounter. So we should presume that the agent will perform reasonably well within the tightly constrained corner of the ecosystem (or society) which it inhabits. However, even here we can't expect heuristic rules to produce the right results 100% of the time. And in the contemporary social world, niches are disrupted; they vanish; agents are forced out of them. When this happens, the agent will draw its rule-driven conclusions *anyway*; so we can expect the agent's performance to degrade rapidly.

I'm now going to take the liberty of discussing the inner life of what we had been imagining as a robot. What is the phenomenology of heuristic inference rules, deployed without the logical apparatus of defeasibility? The agent understands its rules as exceptionless, and it applies the rules mechanically. Especially when the designer has hardwired the inference rule, or otherwise made it wear the aspect of necessity, the agent is liable to have the experience of encountering the inconceivable, of coming face to face with the impossible, and of living in a world that is *broken*.[16] So suppose the designer allows the rules to be applied, still indefeasibly, but now allowing the agent this much discretion: it may opt to reject a premise of the inference, in place of drawing the conclusion. Then when applying one of those heuristic inference rules produces a visibly incorrect answer, such an agent will be prone to a Popperian response: the exception has *falsified* the rule, which must be discarded.[17]

To apprehend the failure of a heuristic inference rule in any of these ways is not conducive to graceful recovery. But because these are *heuristic* inference rules, sometimes they *will* fail. Thus the agent we have been contemplating proves to be extremely brittle.

4.4

The heuristic inference rules of our boundedly rational agent leave it prepared only for a narrow range of circumstances, generally the social and technological analog of an ecological niche. Even such a niche is not generally a fully controlled, drastically simplified environment, and in a world like ours, which contains many such niches, and a good deal of unstructured wild

[16] 'Hardwired': Nozick, 2001, pp. 120–155; 'otherwise': see Millgram, 2015, ch. 8.

[17] Popper, 1977; in the choice-directed Kantian variant defended by Christine Korsgaard (2009, sec. 4.4.2), an agent in a special class of such circumstances is to modify his 'provisionally universal' maxim, incorporating the exception into its antecedent; for discussion, see Millgram, 2011.

terrain, heuristic inference rules are bound to fail more or less frequently, and those failures can sometimes be catastrophic.

The fix that is bound to occur to the designer we are imagining will be to prepare the agent for failure. This particular coin has two sides to it. First, the agent is to recognize the failure of a heuristic inference rule as not necessarily invalidating the rule itself: this is to understand such rules as being fine as they are, even if on occasion, some additional factor makes them give the wrong answer. Here it matters that there's a retrospective explanation for why, in a particular such case, the rule didn't do its job; the availability or otherwise of such explanations allows the agent to distinguish cases that motivate reconsidering the rule (and perhaps upgrading or replacing it) from those in which the rule is to be left as was.

Second, while the sort of systematic survey that would guarantee error-free conclusions remains out of reach, it will nonetheless pay to add a layer of prechecking, especially for high-priority defeaters, to the execution of an inference rule. For instance, as the library considers its move to replace paper with electronic resources, its administrators ought to keep in mind that its clients are drawn from the many disciplines supported by the university; each discipline has its own methods, and consequently is likely to have different needs. They should at any rate be asking, ahead of time, how the substitution will affect the workflow of historians, biologists, nursing students, etc. Further facts about those usage patterns may amount to good reasons to abort or modify the policy—that is, the conclusion of the practical argument which we put into the mouth of some library administrator.

At this point, we recognize what we have in front of us: we foresee that the designer will fold defeasibility into his agent's inferential equipment. And with that, we can see what problem defeasibility is there to solve. Defeasibility is a way of equipping a boundedly rational agent, one which operates with inference rules that have been sized to its computational limitations and observational restrictions, to cope with environments in which those rules will misfire—because the circumstances an agent encounters raise issues which are, as far as those rules are concerned, TBA. (In our creature construction narrative, it is a fix for a design optimized to solve restricted versions of a general problem, once those restrictions can no longer be assumed to hold.)

Here we aren't going to take up the question of how to implement this logical device, but we can look ahead and do a bit of guesswork about what shape it will likely take. Suppose that what we are considering is quite possibly a series of unprincipled kludges. (You might well: in the last iteration of our example, we had learned from experience that it pays to check in

with the differently specialized academic clientele. That's not anything like a procedure that is bound to catch the highest-priority problems that can arise in arbitrary applications of the means-end inference pattern). Then we shouldn't assume that anything like a formal logic is the best theory of it.

4.5

Perhaps the recognition I invited a moment ago was premature; we haven't yet accounted for the most important distinguishing logical mark of defeasibility, that is, the openendedness of those lists of potential defeaters.

By way of making that feature more vivid and more puzzling, notice that on the one hand, we allow that you can always think up more defeaters; they don't run out. But most of us find it unreasonable to insist that there are *infinitely* many defeaters. It is very easy, when worrying about *ceteris paribus* clauses, to find yourself talking as through there were a cardinality that was larger than any finite integer, but smaller than $\aleph_0$—a slightly crazy supposition that we could call "the sub-continuum hypothesis".[18] Putting this logical illusion, and the right way to undo it, to one side for a moment, keep in mind how very distinctive the logical appearance of the phenomenon we are trying to capture is.

The worry we are trying to face up to now is that from the point of view of the agent-under-construction, who doesn't know which of the triggers of those best-in-principle inferences were pared away, there might *seem* to be indefinitely many further defeaters. But in the story we told, we were assuming that each of those best-in-principle inference rules required only a finite, albeit sometimes large, set of premises. Those unpacked *ceteris paribus* clauses seem to the agent to stretch away into the indefinite distance only because it is too myopic to see all the way to their finish lines.

While we're addressing this worry, let's also take up an objection to our earlier train of thought. If the designer we were imagining is sufficiently smart and well-informed, won't he develop a repertoire of heuristics that *won't* fail his agent—or anyway, that will fail so rarely that the glitches can be treated as noise, and ignored?[19] Surely we don't want our creature

[18]But an unreasonable supposition not because we're unable to explore the possibility formally; see Mycielski, 1986.

[19]If he did, would we want to stop calling them 'heuristics'? After all, we introduced heuristics as rules of thumb that only work most of the time.

I won't try to adjudicate the terminological choice, but keep in mind that even when a rule is infallible, there's a distinction we don't want to lose track of: I may tell a child to play only with other children whose names start with a vowel, because I know, though she

construction argument to turn on the deliberations of an inept, incompetent, badly informed designer.

But a creature construction argument is meant to legitimize an aspect of agency by exhibiting it as a solution to a problem.[20] For the deliberations of our designer to do that job, the designer's deliberations have to make sense to us: they are meant to allow us to see why *we* would do it that way. That means that the considerations appealed to by the imagined designer's arguments have to be considerations we are able to understand. Now, we are boundedly rational ourselves. That means that we must imagine the designer of our agent-in-progress to be boundedly rational, in pretty much just the way we ourselves are. (Not that we have a choice: it's not like we have any idea how design problems would look to someone who wasn't boundedly rational.

For every *modus ponens* there's a corresponding *modus tollens*, and let's consider how that would look. Allow for a moment an alternative: a designer who is infinitely intelligent, completely knowledgeable, and who is out to design rules we can apply deductively and infallibly. This is to try to imagine *God*—in whatever way people used to work at imagining the inconceivable—guaranteeing our reason, in roughly the manner of Descartes. (So the direction is not unprecedented.) Since we can't track this designer's deliberations, however, and He is made out to be the source of our own inferential equipment, that would put an understanding of our logic beyond reach. But now, we *think* with our logic, and if we can't see why our logic makes sense, *we* turn out to be unintelligible to *ourselves*. This is, in my own view, a reason to prefer bounded-rationality creature construction argumentation to the more familiar posture that takes logic to be philosophically rock-bottom.

If the agent's designer is boundedly rational, that means that his own

---

does not, that in her classroom this condition efficiently excludes the bullies. The (perhaps oxymoronic) infallible heuristics we are momentarily entertaining should be thought of as like the advice the child receives, as opposed to the reasoning that justifies providing it.

[20]It's worth noting that the term is sometimes applied otherwise. In Bratman, 2014, what is described as a creature construction is a reduction of group or shared agency to the interlocking plans and policies of individual agents, the point being ontological parsimony, rather than legitimization. Bratman's reduction of group agency proceeds by showing how it can be implemented. This contrasts dramatically with Bratman's earlier work, which did not try to show that intentions and plans could be reduced to beliefs and desires; on the contrary, intentions, plans, and policies were presented as an additional, special piece of one's mental and practical life, the objective being to show why we need plans and intentions—that is, how they are important. What we are doing here is the latter sort of exercise: we're aiming to legitimize defeasibility in inference, without considering how to implement it.

arguments—the ones that account for the design of his agent—are themselves defeasible: he overlooks things, too. Defeasibility in the designer's reasoning filters down to the agent's: normally, whatever the designer has overlooked is going to be one more thing to go wrong in the agent's inferences, in ways that the agent doesn't anticipate. And that means, first, that we aren't in a position to insist on the sort of omniscience in the designer that would bullet-proof his agent's heuristics; moreover, it emerges that there is no perspective—none we can make sense of, and none that we can appeal to in an argument like this one—from which the openendedness of those lists of potential defeaters is merely illusory.

'Emerges' how? Evidently by a regress argument: a designer's bounded rationality is to be explained by his own designer's bounded rationality. . . and so on, up the line. Each layer of explanation introduces the possibility that more has been overlooked. Summing up all those possibly-overlooked issues, we have indefinitely many matters TBA, all of which can filter down to that first-level agent. Thus perfection in the design of the agent cannot be invoked to eliminate the possibility of failure from its repertoire of heuristic inference rules. And so even if the agent could be sure that he was working with heuristic inferences that had been cut down from the very best inference rules the designer could find, and also that those source rules had finite lists of premises, they still could not be assumed to reliably handle issues that the designer had overlooked, but which the agent might nonetheless encounter.

That observation allows us to put in place a further methodological point. Creature construction arguments are defeasible—as is all nondeductive argumentation. You may be starting to think of objections to the argument we have underway; many of those objections will field candidate defeaters for it. Since the list of potential defeaters is openended, we can't check all of them, which is to say that the disposition we philosophers are trained into in graduate school, to consider every possible objection, is misplaced. As always, when we are assessing a defeasible argument, we have to prioritize the defeaters we take under assessment; just now, we will want to select the ones that deepen our understanding of how the argument works, and of what's at stake in it.

4.6

One way to cast the conclusion at which we've arrived is that, once you've reached the stage in the design of a boundedly rational agent where it's

equipped with inference rules, there are good reasons to equip your agent as well with the logical apparatus of defeasibility. But these days, rule-based systems might seem like one of two options, the other being to construct a probabilistic reasoner.[21] (However, if you're not into probability-driven argument, feel free to skip this section.) If the agent is going to be a Bayesian inference engine (that is, if it will be assigning probabilities to nodes in a Pearl network, and updating them in accord with Bayes's Theorem), perhaps defeasibility won't be an extension we need to provide. And since it is something of a mystery how defeasible inferences are to be implemented, that might even speak in favor of choosing probabilistic inference over the more traditional systems of rules that we are used to representing with the tools of truth-functional and quantificational logic.

The probability formalism requires a definite sample space.[22] Here's an easy way to see that: the probabilities have to sum to 1, and if there is no definite range of possibilities to distribute them over, that requirement isn't well-defined. If a list of potential defeaters is openended, then it's a mistake—a *logical* mistake—to try to tag them with probabilities.[23]

Suppose our designer has opted to construct his agent around a Bayesian network, and is initializing its nodes and edges as we speak. We've argued that our designer has to be thought of as himself boundedly rational; that means that he will overlook things, and there's no perspective which we know how to adopt from which we can enumerate the things he might be overlooking. So when the agent's Bayesian network gets built, indefinitely many nodes, representing those overlooked possibilities, will end up missing. The construction of the network will thus involve an idealization, that the indefinitely many overlooked possibilities are simply not there. (That's what gives us that definite sample space.) And there is a further optimization we will see in such an agent. Bayesian updating is computationally intractable; Bayesian networks were invented as a way of replacing the general and intractable problem with an easier one, but updating Bayesian networks is also NP-complete.[24] That means that, if you think that Bayesian inference is the in-principle-correct way to reason, you should also think that there will

[21] For the ideology, see Earman, 1992; for the mechanics, Pearl, 1988.

[22] See, e. g., von Mises, 1957, pp. 11f, 15, 28. In a typical axiomatization (e.g., Kolmogorov, 1956, sec. I.1), the requirement is built into that very first designation of the set of elementary events: a set is individuated by the elements that it contains. (If you add another element to a set, you have *another* set.)

[23] As Oaksford and Chater, 1998, p. 282, observe in passing, "it is not clear how to encode knowledge in terms of probabilities... when there are options that have simply not been considered at all."

[24] Cooper, 1990; for orientation around that concept, see ch. 5, n. 12.

be problems, representable even with moderately sized networks, for which it will not be computationally possible to arrive at the in-principle-correct answers. (The usual way of making the constraint vivid is to observe that no matter how large or fast your computer is, the universe will be over long before your program returns the answer to your problem.) That means in turn that, instead of building a Bayesian inference engine, our imaginary designer will be selecting approximation algorithms (i.e., heuristics); and once again, the agent had better be prepared for those occasions on which an approximation algorithm produces results that are simply not good enough.[25]

If that's right, we can observe, first, that even a Bayesian agent needs to be able to cope with those exceptions. Perhaps we train the Bayesian autonomous vehicle to make driving decisions on the basis of what cars normally encounter by way of roads, other cars and so on. But we turn out to have omitted occasions on which the vehicle is being strafed by hostile crop dusters: the possibility just didn't occur to us. What should an autonomous vehicle do, when the *North by Northwest* crop duster turns up above that country road? Whatever the apparatus of defeasibility is, Bayesians need it, too—and we have just registered that it is, in conception and formally, deeply discontinuous with the probability calculus.[26]

And second, it is suddenly obvious what probabilistic inference must be, from our point of view: a heuristic, suitable for use by a boundedly rational agent, in circumstances where the openendedness of those *ceteris paribus* clauses can be idealized away. To think of your sample space as definite is a bounded-rationality technique; you have to be prepared for it to break down. Even if, as some people in the business seem to think, God has endorsed the theory of probability as the doxastic gold standard, to human beings as we have rendered them, probabilistic inference is just one more special-purpose, unevenly reliable shortcut.[27]

---

[25] And there will be occasions on which they aren't; see Dagum and Luby, 1993.

[26] Here's another way to put that point. Carnap, 1963, pp. 211–13, introduces his "requirement of total evidence": you're not licenced to infer a probability (or a degree of confirmation) from only *some* of what you know, if other information you have is also relevant. That is to say, Carnap is insisting that you take into account—as we would say it—defeaters for a probabilistic inference, of which you are aware. Surely, if you were able to step back and see a bigger picture, you would need to take into account evidence—defeaters—of which you are now unaware. So even when you do your reasoning with probabilities, you ought to ensure that you are logically prepared for things you haven't anticipated, coming at you from left field.

[27] The mismatch between the workings of that unremarked logical operator, the '. . . ' in those *ceteris paribus* clauses, and the probability formalism has, it seems to me, a further source. Now, as Auyang, 1998, points out, the probability formalism has various and quite different uses, but *one* of them is to mediate the transmission of guidance across

4.7

Around the turn of the twentieth century, both the emerging analytic tradition and its continental competition decisively rejected what came to be called the "psychologistic fallacy". In the formerly widely accepted psychologistic approach to logic, the objective was to explain what reasoning done properly was—that is, it was a guide to an activity. And since the reasoning was done by minds, the contents of such a guide were to depend on how minds worked. (You wouldn't dream of writing a guide to riding bicycles without knowing how bicycles work, and a how-to manual for your brain is *just like that*.) Abandoned in favor of a view on which logic is the science of the consequence relation, and a subfield of mathematics, psychologistic philosophy of logic has been consigned to a hostile neglect.[28]

But notice where the argument once again has taken us. We wanted to make sense of defeasibility in inference (and in related matters, such as similarity judgments). We explained it as a feature of the design of a boundedly rational agent, and in particular, as a way of compensating for the inevitable failures of the only sort of inference rules such an agent could deploy. But to do that is to invoke the limitations of a mind, on the basis of an understanding of how it works. That is, as we noticed in previous chapters, we have been engaged in psychologistic philosophy of logic.[29]

---

disciplinary barriers. The archetypal challenge consists in a specialist who, in virtue of literally years of training, understands a complex and nuanced set of issues, and who needs to convey information or advice to a client in a different specialization (or for that matter to a layperson). What the providing specialist would do if he could is explain the inferences supporting his conclusions—along with their defeasibility conditions, so that the client could assess for himself what might go wrong. But without those years of training, the client simply cannot understand any of that; his attitude will tend to amount to: "Just give me a number!" The providing specialist will often be fully aware that the probability he provides is false to what he knows, but it will be the best he can do, and he will be frequently forced to do it. That is, probabilities are too often introduced as a way of ranking defeasibility conditions because they are a way of *dumbing down* (and being genuinely misleading about) defeasibility conditions.

[28]The canonical texts in the respective traditions are Frege, 1978, and Husserl, 2001. For a sociology-of-knowledge perspective on the psychologism debate, see Kusch, 1995, and Ringer, 1990, esp. pp. 295–298. The phrase used to characterize the more current conception of logic is lifted from Sorensen, 2001, p. 10.

[29]I had better preempt a frequent response. It used to be routine to accuse psychologistic logicians of confusing empirical psychological generalizations and prescriptive laws of logic, so notice that that is not going on here. (A typical and recent expression of the posture, from Safranski, 2001, pp. 59f, a popularizing biography: "Psychologismus. . . den Versuch also, das Logische aus dem Psychischen zu erklären." And then the author goes on to quote Heidegger objecting that the logical "ist ein 'statisches' Phänomenon, das jenseits

Perhaps putting psychologism to one side for as long as we did was not a mistake; we learned an impressive amount by exploring the avenues opened by Frege, Russell, and others, and it was important to see how far they would get us. Nevertheless, we now find that, in our efforts to make sense of nondeductive reasoning, we have been stalled for long enough to make it likely that we have been working with the wrong tools. And if it proves time for a shift of key, and that thinking about defeasibility using the conceptual toolbox of bounded rationality is what allows us to make headway, then we owe it to ourselves as philosophers to reconsider the methodological givens over which we have been riding roughshod here. Psychologism turns out to be unavoidably an agenda item on the bounded-rationality program, and for all of us philosophers, it's time to reopen the question of whether psychologism *was* a fallacy, after all.

---

jeder Entwicklung und Veränderung steht." That's surprising common ground with the analytic mainstream, and as you'll have realized, it's being contested here.) Creature construction arguments establish the legitimacy of something or other—in this case, a logical device—and not by identifying the device with a psychological generalization. More generally, I can advise you not to believe the mischaracterizations of psychologistic logic by its unrestrainedly polemical opponents. Psychologistic logicians (such as J. S. Mill) were trying to lay down guidelines for how *to* think, which they understood very well was not, for the most part, how people *do* think.

Chapter 5

• • •

Defeasible inference is usually introduced by contrasting it with inference or reasoning that is deductive: if the premises of a deductively valid argument are true, then (and necessarily) its conclusion is true, which is to say that no matter what *else* you learn later—as long as it's not that you made a mistake about one of those premises—you won't have to take the conclusion back. Whereas a defeasible argument can be perfectly fine, premises and all, even though additional information might require you to withdraw its conclusion. The distinction is not new: John Stuart Mill contrasted subject areas suited to the 'Geometrical Method', where "what is once proved true is true in all cases, whatever supposition may be made in regard to any other matter," with domains in which "mutual interference" may bring it about that "concurrent causes annihilate each other's effects."[1]

Vagueness has become a technical notion in recent decades, and let me pause to introduce it as well. A sorites (to laypeople, a slippery-slope) argument has the following shape: If not a single flake of snow is coming down, it's not snowing; however, just one flake of snow won't make a difference to whether it's snowing or not; so if just one flake is coming down, it's not snowing either; and if two flakes are coming down, it's not snowing; ditto three... and the ultimate conclusion will be that it's not snowing, even as the ski resorts send out their avalanche control units and your friends pile into the 4WD vehicle and head up the canyons for a day of winter sports. A concept (in this case, "is snowing") is vague in this technical sense if it lends itself to constructing sorites arguments.[2]

[1] Mill, 1967–1991, VIII:888; VII:373.

[2] The points I will want to make should hold, *mutatis mutandis*, not just for predicates but referring terms such as names, as well as for scopes of quantification.

Not all of what is informally called vagueness looks like this. Concepts can exhibit open texture, that is, there are situations, usually situations that have never come up before and that no one has stopped to think about, in which it was never settled whether the concept applies or not. Rabbinical law interprets its prohibition on working on the sabbath to forbid operating electronic equipment; what does it say about sending someone a fax, when in your time zone it is not yet the sabbath, but in the time zone to which the fax is being sent, the sabbath has begun? Perhaps there is a principled way to resolve the problem, but when the prohibition was formulated, fax machines and time zones were far, far in the future.[3]

Up to this point, the discussions of these two topics—defeasibility and vagueness—have largely been kept separate. (Actually, the former has had *three* separate discussions: in more traditional treatments of reasoning and rationality, the keyword is *defeasibility*, because those further facts that interrupt inferences are "defeaters"; in philosophy of science, generalizations are said to hold only *ceteris paribus*, or "other things equal"; in AI, such reasoning is *nonmonotonic*—and just to be clear, these are approximate synonyms, not labels for distinct but perhaps related topics.) Moreover, the discussion up to this point has considered these phenomena almost entirely as they appear in the wild, so to speak: turning to snowflakes for an illustration, as above, is typical in this regard.

Here I want to consider defeasibility and vagueness together, and not in the wild, but as they are induced by a large-scale strategic choice that we have collectively made, in order to address a world that is too challenging for any one person to handle on his own. Division of labor is not a new feature of the human social world, but over the past few centuries, it has become both far more extensive and far more intensive; the difference in degree has become a difference in kind, and one with philosophically important consequences.[4] More and more, we are highly specialized, and when division of labor goes sufficiently deep, it entails division of intellectual and evaluative labor.

The flip side of division of intellectual labor is ignorance, of almost all those things it is not your job to know; the flip side of division of evaluative

---

For a no longer current but still useful introduction to the problem space, see Keefe and Smith, 1996. For a representative sampling of work done in it at the high point of the recent wave, see Dietz and Moruzzi, 2010.

[3]For discussion of open texture, see Shapiro, 2006; the example is due to the talmudist Dr. Yael Fisch.

[4]Indeed, as my colleague Matt Haber adds, each year we are reminded at commencement ceremonies that differences in degree *make* differences in kind.

labor is not being competent to assess most of what lies outside your area of expertise. Defeasible inference and vagueness are ways of preemptively making room for all those things you don't know and can't assess. In our social world, we have to understand them as aspects of a heuristic approach that has been adopted by boundedly rational agents.

One last bit of foreshadowing: When defeasibility and vagueness are introduced as this sort of strategic choice, treating a class of inferences as deductive becomes a strategic choice also; this raises the question of how such a choice could be motivated. So I will wrap up by reopening the question of how to characterize the contrast with which we began, between defeasible and deductive inference.

5.1

Let's add in a little more granularity. When we represent the workings of defeasible reasoning, we find ourselves saying things along the lines of: the inference goes through, *ceteris paribus*, or 'other things equal'. And if we are asked *what* other things have to be equal for the conclusion of the inference to stick, we find ourselves starting up an openended list, one that trails off into an ellipsis. That '. . . ' is the logical marker of defeasible inference that most demands philosophical analysis.[5] Quite generally, I think we do best to account for it as an artefact of our bounded rationality: the design of the human cognitive engine inevitably involves tradeoffs, in which guarantees of accuracy and correctness are sacrificed for speed and economical use of resources. Engineering analysis, rather than a focus on what can be proved within and about a syntactically characterized formal system, is the right way to understand its performance.

Zooming in on just one of the drivers of the design: As logicians and as philosophers of science, we tend to interpret ourselves as reasoning on the basis of universal generalizations. (We know how to represent reasoning that applies exceptionless generalizations, and doesn't science aspire to laws that hold everywhere?) But agents produced by natural selection are boundedly rational and come with narrowly circumscribed ranges of what they're able to observe; it is hard to believe that they would be built to make these sorts of generalizations the linchpins of most inference. After all, and just for starters, the history of science quite convincingly demonstrates that if universal generalizations are to be any good, they will be expensive. And so

[5]Brandom, 2001, sec. 4, is one of the better preliminary attempts to construe *ceteris paribus* clauses as logical markers, with due attention to that openendedness.

instead we should expect local sketches, generated cheaply and on the fly, to handle relatively small problem spaces, and whose inferential deployment accordingly requires the logical apparatus of defeasibility.

But one would think that the highly specialized disciplines that produce our highest-quality intellectual resources would work differently: surely *they* aspire to, and do, provide generalizations that hold across the board. On the contrary, or so I will argue: defeasibility is going to characterize our inferential practice, *especially* when we are contributors to or clients of the sciences.

5.2

Examples of defeasible reasoning are everywhere, and let's take the first one that comes to hand. Some time back, my administration, like many others, determined that our hiring processes are conducted electronically: applications are submitted online, read on-screen, and at each stage the status of each application is updated in the system. The reasoning underlying that determination presumably ran roughly as follows:

1. Any process that is basically paperwork can just as well be made paperless.
2. Hiring faculty is a process that is basically paperwork.
3. So it can just as well be made paperless.

On the face of it, that conclusion ought to follow from the premises. What's more, and to anticipate our ultimate destination, it's easy to construe the argument as either deductively valid or as defeasible (and in a few moments, I'll indicate how). For now, let me ask you to hold that thought, and to read it as defeasible, so that we can start up an in-principle-open-ended list of defeaters.

It would be a good idea to withdraw the conclusion if it turned out that the paperless version of the process shifts a significant amount of work from the secretarial staff, who formerly organized and managed the paper files, to the tenured faculty, who now have to perform the electronic versions of these tasks online. (Secretaries are less expensive than faculty, and tenure-track faculty time is a limiting resource in the academic ecosystem.[6])

[6]The dynamics of the shift roughly recapitulate the proletarianization of housework studied by Cowan, 1983, in which, as one appliance after another came on the market, the middle-class housewives, who had started out managing households, came to perform the menial labor they had formerly supervised.

It would be a good idea to take the conclusion back if it turned out that the software introduces glitches and bottlenecks that regularly require the intervention of the technical support staff, and which can't be resolved within the very tight timeline of a search. It would be a good idea to take it back if it turned out that system upgrades will be scheduled at the same time as application deadlines. (Yes, I've actually encountered this.) It would be a good idea to take it back if it removes the flexibility required in handling letters of recommendation for senior applicants—and more generally, in negotiating the social norms that dictate how such things are managed within the field. It would be a good idea to take the conclusion back if it turned out that people read more poorly on-screen, and produce lower-quality assessments. It would be a good idea to take the conclusion back if the process opens up the university's systems to malicious attackers. It would be a good idea to take it back if it turned out that the electronic workflow undercuts the university's diversity and equity priorities, or if the software package is sharing the applicants' private information with third parties, or if the interface, which is after all most applicants' only point of contact with the university, is experienced by them as frustrating, and contributes to a lasting dislike of the institution, or if it becomes one more factor reshaping the agendas of the broader applicant pool away from real research, or (now just as a reminder that potential defeaters can be far-fetched, as well) if the software activates the long-dormant but blasphemous Curse of Cthulu... And it should be clear enough that, as the '...' indicates, this list could be extended indefinitely.

Already we can observe that not only doesn't the well of potential defeaters ever run dry; there's nothing in particular that they have to be *about*. The inference itself is motivated by efficiency, and while some of the defeaters in our illustration are matters of efficiency also, others have to do with IT security, with gender equity, with privacy, with reputational capital, with the commitment to research on the part of a properly academic institution, and—although I won't pretend to deep knowledge of H. P. Lovecraft—I suppose with the salvation of your immortal soul.

I remarked that you can read that sample inference either as deductive or as defeasible, and now that we have representative defeaters pending, we can make those options concrete. Take the concern that a badly timed upgrade could prevent applicants from uploading documents, or lose data already in the system. This could be read as an objection to one premise or the other. (Perhaps it shows that some processes that are basically paperwork *can't* just as well be made paperless, or alternatively, that hiring faculty *isn't* a process that is basically paperwork: one or the other of those

claims needs to be circumscribed.) Alternatively, those premises can be allowed as they are, and the objection seen as marking one of those occasions on which perhaps you oughtn't to go ahead and draw the imminent conclusion. Which way to render the argument is a *choice*: it's not something settled by what's there on the page, but turns on matters of policy. Adding in a further logical point: before opting for the deductive reading, take into account that deductive arguments are primarily useful when they have true premises. To get flat-out-true premises in an argument like this one means contouring them to sidestep the live defeaters, but as we've seen, the well of potential defeaters is bottomless, and it shouldn't be clear that that's possible—or even when possible in principle, feasible.

We can also observe that the illustration straddles the distinction between theoretical and practical reasoning—that is, between figuring out how the facts stand, and deciding what to do. Theoretical inferences take you from truths to truths, or anyway that's the usual way of explaining what it is for such an inference to be valid: that it does so. There isn't an uncontroversial account of what the relevant statuses of the steps of a practical inference are, but it's procrustean to construe them as truth and falsity. So the phenomenon of defeasibility is a good deal broader than it might have seemed at first. An inference is defeasible not just when you might have to allow the conclusion to be false after all, even though the premises stayed true, and even though the inference was satisfactory until you learned about some further facts; it can be defeasible in that you're still committed to the objectives that kicked it off, but now that you've realized how important the side effects of proceeding will be, you need a mulligan on your decision. Whatever we say about defeasibility, it should naturally cover the full extent of the apparently homogenous territory, rather than stopping short at the artificial border of one category of reasoning or another.

5.3

Our focus here is defeasibility and vagueness as they arise out of division of labor, but just by way of warming up, and to motivate the approach we want to take, consider what generates the defeasibility, in our toy-but-nonetheless-real-life example. The premises of the argument exude sloppiness; it is not as though the concepts they invoke—"basically paperwork," "paperless," and even "hiring faculty"—are carefully contoured for precision in deliberation, and it is not as though the claims themselves are the products of careful, thorough investigation and vetting. Even a moment's thought should be

enough to convince us that they are only, as the turn of phrase has it, good enough for government work. (The phrase is apt here: my home institution is an arm of a state government.) A hire involves often lengthy meetings on the part of real live people, on-campus interviews, many typically delicate behind-the-scenes conversations, phone calls to handlers, and much else; it's not just paperwork, and when it is, the results are terrible. Conversions of paper-based workflow to the updated electronic version modify how decisions are made, sometimes in subtle ways, and sometimes in not-at-all subtle ways. For instance, once notes on files and meetings are maintained electronically, where they can be instantly reviewed by any higher-level administrator, they can no longer be used for candid communication between committee members, and department meetings can no longer be a venue for deliberation. The process bifurcates, into a ritual meant to create a sort of Potemkin Village of reviewable notes and minutes, and informal chats in the hallways and offices where any useful back-and-forth now has to be conducted.[7]

Again, I laid out the illustration so as to make it read naturally either as a deductively valid inference or as a bit of defeasible reasoning. To read the argument as deductively valid but reject its conclusion means locating objections to it at its premises. That way of diagnosing the weaknesses of the toy argument presents us with an additional choice. The standard way of going on concludes that one or another of its premises must be simply false. However, we can also, and more realistically, understand the overt form of the argument as deductive, but take those premises to be kind of true, sort of true, but not flat-out, across-the-board true. But now, arguments with

[7]If you're not in the business, you may be wondering why that's true. After all, shouldn't properly taken institutional decisions be transparent, and isn't something crooked or corrupt if they can't be?

In any genuine choice having to do with, say, hiring or tenure decisions, there are always pros and cons, and normally those considerations don't come off sounding like exercises in hyperbole. Of a particular candidate, you might think that their research will be professionally respectable, but probably never agenda setting; that their graduate teaching will significantly strengthen the program, but that their undergraduate teaching will not necessarily be anything to write home about (or the other way around); that the curricular coverage they provide will be narrow, but that we do need those classes taught; that they will likely be outward-facing to the profession, but not our medical campus, and so on. One learns very rapidly that administrators treat the mention of a downside as a veto, and that hyperbole throughout is an absolute requirement. (Just for instance, they need the word "excellence" to be repeated constantly.) The upshot is that institutional deliberation has to happen, in one way or another, behind closed doors.

For more generally framed second thoughts about the rhetorical question that begins this note, see O'Neill, 2002, ch. 3.

premises that are partially true in this way have to be treated as defeasible.[8]

However, if we're unhappy with the premises of the argument, think what it would take to replace them. I can imagine a large, very motivated corporation (back in the day, Xerox) commissioning a study to find out how to think about paper and electronic workflows. A university's administrators are busy people; they live with budgets and deadlines, and are not about to commission or wait on the results of such a study. They need, and are going to make do with, rough and ready rules of thumb that are workable enough in the environments in which they operate. And this means that the driving premises of the arguments they are explicitly or implicitly fielding will be local generalizations, first in the sense that they are largely based on observations that administrators at a particular institution, or perhaps at a cluster of institutions, happen to have made, based on whatever it is they have encountered; and second, that they thus tacitly rely on background features of that domain.[9] The list of background features that figure this way is generally not articulated, and the track record suggests that it couldn't be: no matter how long your list is, there are always further features of a situation that might turn out to be what underwrote a local generalization of this kind.[10]

But now, in the relevant sense (allowing for the moment highly artificial exceptions), *everyone* lives with limited budgets and is constrained by deadlines: that's why rationality in the real world is always *bounded*, and let me quickly introduce the concept for those who haven't encountered it before, both in its old-school and cutting-edge versions.[11]

---

[8] Or so I have argued elsewhere: Millgram, 2009a.

[9] That is, recalling that many of these ideas have venerable antecedents, those generalizations will be what John Stuart Mill called "Empirical Laws": Mill, 1967–1991, VII:516–24.

[10] We see this most vividly in competitive and especially hostile environments; again and again, opponents identify hitherto unnoticed presuppositions of each others' strategies, along with ways to dislodge them. Military history is especially instructive in this regard, as is the current arms race between the actors who are attempting to keep our IT systems secure, and the actors who are trying to breach them.

[11] At this point, I'm putting to one side the other driver of defeasibility which I mentioned earlier on: that even ideally rational agents are never omniscient. Briefly motivating that choice, some philosophers—most famously, Descartes—have thought that an observationally limited agent could nonetheless do at least a great deal of science deploying only nondefeasible inferences. Moreover, the Bayesian picture of ideally rational but observationally limited investigation is popular enough to have been elaborated into its own subfield, 'formal epistemology'. An adequate discussion of either would complicate an argument that already has about as many moving parts as it's reasonable to ask a reader to track, but for a start on the latter, see sec. 4.6.

There are a great many problems that we know how to solve in *some* sense: for instance, there is a procedure that is guaranteed to find the solution, provided that our computational and other intellectual resources are unlimited. (You would need an infinitely fast computer, one that doesn't run out of memory, and so on.) However, and surprisingly often, what is possible in that sense is not feasible for actual problem solvers, sometimes only in practice, and sometimes in principle. Under the former heading: We're told to choose the option with the highest expected utility, but there are so many recipes in that shelf of cookbooks that finding the best combination of them would take much, much longer than we have before the dinner guests arrive. Under the latter: The options we're choosing between might be routes that take us to all of the cities on a road map, with the objective being to minimize the distance driven. (This is the famous Traveling Salesman Problem.) Here, finding the best option is computationally intractable; no matter how powerful a computer you have, even for not-all-that-large instances of the problem, the universe will be over and done, long before your algorithm returns an answer.[12] So we instead adopt heuristics, ways of solving a class of problems that trade off in-principle-correctness for speed or frugal usage of other resources: look for one recipe in each course category (soups, salads, mains, desserts), and take the first combination you find that seems good enough; assume that the network of roads can be embedded in a Euclidean plane.[13] A heuristic might get good-enough results rather than the best possible results, or get the correct result most but not all of the time, but there is no point in trying to list all of the ways those tradeoffs might go. (There are always more compromises to entertain.)

That was the old-school notion of bounded rationality; let's contrast it briefly with its cutting-edge successor. We introduced bounded rationality by contrasting it with ideal rationality, but we don't always have the anchor for that contrast. Often we don't know what would count as the in-principle-correct solution to a problem, but we need to consider accuracy-vs.-resource

[12]Intuitively, an NP-complete problem is a magic key to a class of problems (those that can be solved in polynomial time by a nondeterministic Turing machine), among which are many problems which are thought to be very hard: i.e., if you had a solution, you could convert it into quick-enough solutions for every problem in that class. So to show that a problem is NP-complete (as Traveling Salesman is) is to create a very strong presumption that, anyway for worst-case inputs, it's not realistically solvable.

For a now quite dated but still very user-friendly introduction to the concepts, along with an early survey of problems known to be NP-complete, see Garey and Johnson, 1979. For a discussion of the reasoning that underwrites the presumption, see Thagard, 1985.

[13]Klein and Mozes, unpublished. And of course giving up on a solving a problem is an option; see, e.g., the short list of criteria at Simon, 1967, p. 32.

tradeoffs anyway. For instance, you might take the various results descended from Condorcet and Kenneth Arrow to show that we don't know what ideally rational collective decision would *be*, and maybe even that there *couldn't* be such a thing.[14] Nonetheless, we select one voting procedure over another, perhaps because, although it cuts corners in one way, it's easier to implement in others. Cutting-edge work on bounded rationality acknowledges that sometimes it's *all* heuristics, and lets go of the assumption that we can take our understanding of ideal rationality as read.[15]

So when you look more closely, we're *all* like those hasty, sloppy university administrators in the way we do our reasoning; we have no choice but to be. Or perhaps just almost all: the highly artificial exceptions I mentioned a moment ago typically belong to the academic and scientific enterprise, where at least in theory, the constraints are removed or softened. (Tenure means: take as long as you need to understand whatever it is you're looking into; we still won't fire you.) So you might have been thinking that science is different; here we can do what is necessary to find high-quality generalizations that are unrestricted in scope. But I am about to suggest that locality in the naive sense is a model for and explanatory window into a comparatively new source of defeasibility, one that has the sciences as its home turf.

5.4

Division of labor is visibly a pervasive and very deep feature of contemporary human social organization. Philosophers have gradually come to recognize that once division of labor goes deep enough, it inevitably involves division of *intellectual* labor; it has perhaps not been sufficiently appreciated that it also involves division of *evaluative* labor. Let me quickly sketch the shape the both of them take, and some of the consequences.[16]

[14] Moulin, 1981, pp. 239–42; Shubik, 1985, pp. 120f.

[15] Wimsatt, 2007, is an example of this approach, which reminds us that what had seemed to be an in-principle-correct technique for arriving at results can turn out, on further inspection, to itself just be another heuristic.

How decisively has the transition from old-school to cutting-edge approaches been made? Bendor, 2020, which contains a very thoughtful discussion of the different 'reference points' that research in the bounded rationality tradition has taken within different fields, at one point remarks that "the fully rational agent is like the unicorn: well-defined but fictional"; but elsewhere also remarks that "sometimes 'optimal' is not well-defined for the problem at hand".

[16] Early and classic discussions include Putnam, 1975a, on the division of linguistic labor, and Hardwig, 1985, on the division of epistemic labor. Millgram, 2015, esp. chs. 4, 6, and

Division of intellectual and evaluative labor is a bounded-rationality strategy. Nowadays, there is simply too much to know, and to know how to do, for any one person to have mastered more than the tiniest bit of it. And so we demarcate disciplines—and, along with them, subject areas—into which people are trained, often in arduous apprenticeships lasting many years. If you are an academic, think back to your own training, but the phenomenon I am pointing to isn't confined to universities.

At the end of such an apprenticeship, a specialist does not simply know a great deal that other specialists don't: he *understands* things—he controls an intellectual vocabulary that is unavailable to people without that training. He has acquired standards, professional preferences, objectives and modes of evaluation that subserve assessment in his field, and which outsiders neither comprehend nor are competent to deploy. Extending our earlier example, if you are an academic philosopher playing a role in a faculty hiring process, you are making decisions on the basis of nuanced and delicate assessments of candidates' work that you are unable to explain to the administrators who have to approve the decision.

What a specialist knows, understands and deploys along with his assessments are in part generalizations; these are both theoretical (again, about matters of fact) and practical (directly governing courses of action).[17] So let's consider where they must for the most part come from. Within a highly specialized discipline or profession, the generalizations its members most centrally deploy, if they are worth much at all, must have been hypothesized, formulated, checked and debugged within the discipline itself—or perhaps in a closely-related field. Continuing to extend the example, a certain sort of insistent appeal, in the course of a potential hire's argument, to intuitions about what people do and must care about—for instance, everyone cares about their own autonomy or self-governance, or again, if something is a plausible object of concern, you have be a 'realist' about that sort of thing, and so on—are a warning sign, and have to be very carefully investigated: they might betray an unthinking allegiance to a parochial philosophical style, or crudeness in temperament, or that a job candidate is still very early on in the trajectory of their intellectual development. Whether or not you think that generalization is on the mark, who could plausibly come up with it, and who could plausibly push back against it, other than very experienced philosophers, people who have been fully enculturated into the

9, emphasizes division of evaluative labor.

[17] Certainly specialists don't *only* traffic in generalizations; just for instance, models have on occasion come in for a good deal of attention, as in Giere, 1999.

discipline? But now, the specialists in one field do not command the knowledge of other fields, or for that matter control their conceptual vocabulary or exhibit mastery of their assessment tools. That is, after all, why we have specialists—that is, intellectual and evaluative division of labor—in the first place.

Although we specialists sometimes lose our grip on this very basic fact of life, all of those disciplines share a single world, and in any practical context, the division into disciplinary subject matters is artificial. A generalization formulated by professionals of one kind is bound to be applied in cases where the course of action—and whether to accept the conclusion the generalization seems to entail—has to accommodate the understanding, knowledge, and assessments of other specialists: that is, considerations of which one is ignorant. As we put the point earlier on, the flip side of intellectual division of labor is massive ignorance, and the flip side of evaluative division of labor is far-reaching evaluative incompetence. One way or another, room has to be made for folding in overlooked considerations, when they are brought to one's attention after the fact.

Now, it's in the spirit of the view I'm developing to allow for exceptions to one's generalizations, and because some physics is often presented very much otherwise, I'm going to allow that perhaps it provides generalizations that really do have exceptionless, universal coverage.[18] Such exceptions noted, the people in the business of finding generalizations for a specialized group to which they belong—or for that matter for outsiders—find that task delegated to them as a stage in the division of intellectual or evaluative labor.[19] It's not just that their task is finding generalizations, rather than fixing faucets; their task has been narrowed to finding generalizations *about such and such*. That is why we have social epistemologists, condensed-matter physicists, invasion biologists, and snow reporters—and here what matters is not just that we could extend this list for a great many pages, but that it is always growing. As the intellectual and evaluative division of labor responds to the pressures of ever-increasing knowledge, there are constantly new fields and subfields into which it is articulated.[20]

[18]The concession is there in part to sidestep a controversy: see Cartwright, 1983, Cartwright, 1999, and for a very interesting alternative to the crude yes-or-no stances normally taken on the question, Wilson, 2017.

[19]Compare Bendor, 2020, accounting for the ways such tasks are farmed out: "Scientists too are boundedly rational" (emphasis deleted).

[20]People within such a system need a guiding conceptualization to navigate it. In this case, it seems to me, that has proved to be a metaphor adapted from the notion of locality we encountered in our first lap of discussion. There are *fields*; *areas* of specialization; *domains*... and much the same image is found in other languages are well (e.g., German:

The practical problem is not just that specialists have to formulate the generalizations that they themselves will use, or will export for use by others, despite being ignorant of a great deal that might prove relevant. From the point of view of anyone embedded in that division of labor, the ignorance is meta-ignorance; he can't survey the unknowns. Recall that highly specialized experts literally cannot understand a good deal of what is accepted, or under consideration, within other highly specialized fields; acquiring the conceptual vocabulary is too expensive for it to be feasible even to acquire interactional expertise with respect to more than a handful of them.[21] But without being able to ask other specialists what you might be overlooking, and to understand their answers, you can't have much of a grip on what it is you don't know.

Then if you are a specialist formulating a generalization, aware that for all you know, there are a great many issues which you are not in a position to identify, but that would require you to modify it, you face a stark choice: You can choose to adopt a posture of brash confidence, and insist that your generalization is exceptionless and holds across the board. (This is the approach recommended by Popperians.) Or you can opt to advance it more modestly, by making room up front for all those issues of which you are ignorant.

We've briefly considered which approach is better, but let's supplement the earlier conclusion with a different line of argument. Suppose you have decided to incorporate into your generalization everything there is to be learned from intensive interdisciplinary conferences. Once again, for the most part, specialists only understand the specialized vocabulary of their own field. The issues that will be flagged in conference by other, differently-trained experts will normally have to be couched in *their* various conceptual vocabularies. When those flags are folded into your generalization-in-progress, you will end up producing a generalization that no one can understand. It is a good rule of thumb that to use a generalization reliably, you have to understand it.[22] So the determination to incorporate everything

Gebiete).

[21] For the concept of interactional expertise, introduced with much more optimism than seems to me warranted, see Gorman, 2010.

[22] Of course, in part that is true because merely mechanical applications of a rule tend to go badly, and do so because they are insensitive to its defeating conditions; however, I don't mean to appeal to what would here be a question-begging consideration. But in part it is true because application of a rule presupposes the ability to *recognize* both its triggering conditions and its consequences: you're not going to have much use for a generalization about bicycles if you yourself can't identify something as a bicycle when you see it.

there is to know about the limits of a generalization into its formulation turns out to be the resolve to make it unusable.

5.5

We began by noticing the openendedness of those *ceteris paribus* clauses. We've sketched what is just one of several reasons that the ellipsis that we registered as its logical marker has to be there, but we can already say one or two things about what it signals. It is not just that your generalization is liable to have overlooked issues that might matter to the uses that will sooner or later be made of it; those issues can have to do with wildly variable subject matter, and so even saying what they are normally has to be couched in vocabulary invented by specialists to cope with their particular subject matters. This entails that defeasibility is unlikely to lend itself to being represented formally in the traditional manner. For one thing, formal treatments start off by assuming a uniform, nicely regimented, antecedently understood language. But the languages that specialists are forced to invent differ not just with respect to their concepts and referring terms; they can be arbitrarily different *formally*.[23]

That '. . . ' serves as a placeholder for all those things we might have overlooked, perhaps because, like our administrators, we were simply in a hurry, but certainly because, even when we're not in a hurry, our heads just aren't big enough for all the knowledge and all of the evaluations we have to control, and we divide up that task of knowing and evaluating. The division of intellectual and evaluative labor guarantees local ignorance, and local ignorance entails that your inferences have to be ready to accommodate assessments and facts that you didn't know about ahead of time. Generalizations will mostly be local, and the logical apparatus that exploits them will come with cognitive machine oil, to allow the gears coming from different specializations to mesh smoothly. (Think of defeasibility as that sort of grease: the key to making sense of it is that it has to do the *job* of grease.) So what we need instead of a formal logic is an account of the techniques for identifying and accommodating whatever it was you had overlooked, in a timely and cost-effective manner.

[23] For a fascinating guided tour of a handful of examples, see Wilson, 2006.

5.6

Turning to what I'm about to suggest is especially a working component of such techniques in the wild (that is, in the not-yet-specialized natural world), vagueness is usually introduced with illustrations that make it out to be, to put it crudely, quantitative. The question is *how many* snowflakes it takes for it to be snowing, or *how much* red there has to be in the paint for it to be red rather than merely reddish, asked in circumstances where it's apparent that no particular number or quantity could do. And perhaps the diet of examples, to borrow a turn of phrase from Wittgenstein, contributes to the assumption made by most work on the subject, that the route to a satisfactory philosophical understanding is, more or less, a formal logic of vagueness.

But these illustrations are misleading. Open texture and vagueness quite often, maybe normally, appear in tandem. For instance, many rules of the road, dating to the middle of the twentieth century, have as one of their trigger concepts *the driver of the vehicle*—and over the course of the previous century, that concept could be counted on to clearly apply, or to clearly not apply, in just about any particular case. As the technologies that will one day support fully autonomous vehicles gradually fall into place, not only is it obvious in retrospect that there was a great deal of open texture in the concept, but also that it will support sorites arguments. If cruise control doesn't make the difference between being the driver and merely a passenger along for the ride, then ABS won't, and skid correction won't, and self-parking capacities won't, and lane awareness and automatic braking won't... and before we know it, we've agreed that someone who sits behind the wheel of a Tesla, watching videos on his phone while his car slams into the side of a truck, was driving. Today's internet-connected cars can be controlled remotely; are you still the driver if a hacker has taken control of your windshield wipers? Partial control of your brakes? Of the steering?

When a sorites argument is underwritten by open texture—that is, by the many qualitatively distinct unanticipated adjustments to the cases that populate a concept's core, and that carry it out, step by tiny step, into the penumbra of borderline cases and beyond—working inferentially with a vague concept requires being alert to the indefinitely many ways that an instance can fall away from the central applications of a predicate, in ways that might interrupt one or another prospective inference. Inferences turning on vague predicates and names can well exhibit the same sort of openendedness in the list of things to which a reasoner has to attend. And more generally, the intellectual demands that vagueness makes, even when

the inference pattern it appears in is deductive in form, should strike one offhand as very similar to those imposed by defeasible inference. What, after all, is the presumed practical difference between a deductively valid argument and a perfectly good but defeasible one? If the deductive argument comes with a guarantee that allows you to ignore anything beyond the argument's premises, there's no need to keep an eye out for endless problems that are hard to anticipate, and so you can execute your inferences *mechanically*. Deduction spares you having to be *alert*; it removes occasions for judgment. Defeasible inference, evidently like inference in the face of vagueness, is not like this: it puts a premium on being alert, and it requires judgment throughout its execution, because you have to keep an eye out for problems that are all over the map, hard to anticipate, and no one of them can be counted upon to be the *very last*.[24]

Vagueness should be seen as one element of a technique for managing a subclass of defeasible inferences, a technique that is the default expedient in, as I was putting it, the wild. But defeasibility in the wild can make our circumstances seem desperate. If I don't know anything about what might come along to defeat my generalization, how can I use it? If the unknowns are not surveyable, how can just being aware that there are unknowns help, and how can defeasible inference be—or be a component of—a heuristic? Vagueness and open texture have been with us throughout our preindustrial history, and they fall into place as stopgap measures. We should take a bounded-rationality approach to this problem space as well: vagueness amounts to identifying a central stereotype as safe, for one sort of inference or another, while adding a warning that the more in the way of deviation from the central stereotype, the more inferential jeopardy one is in. Vagueness and open texture are suitable, *faute de mieux*, where you can expect users to make sense of that 'more' and 'less': in the usual examples of sorites vagueness, a more or less quantitative 'more' and 'less,' and when it's open texture otherwise, a more qualitative 'more' and 'less'.

Here we are focusing on these phenomena as they are induced by the extreme forms taken by division of labor that we encounter now, and are

[24]The social importance of having inferences that we can treat as mechanical, where you don't have to keep watching out of the corner of your eye, is hard to overstate. For discussion, see Millgram, 2009a, ch. 2.

Of course, deduction doesn't remove *all* occasions for judgment. For instance, per the Quine-Duhem thesis, when a falsified conclusion requires you to give up a premise, you still have to think about which premise it will be. Alertness is not an all-or-nothing matter; nonetheless, what deductive inference is in the first place for is to economize on it, by curtailing the zones in which a reasoner needs to expend his attention.

bound to encounter even more of down the road. When someone shaped by such a division of intellectual and evaluative labor looks for generalizations, there's bound to be an enormous amount of information (but not just information) to which he's oblivious. That means that concepts ought to come with open texture *built in*. After all, useful predicates are formulated to fit useful generalizations. So when one formulates a predicate, there are indefinitely many issues of which one can be quite sure in advance that one is ignorant. It would be foolish to give a predicate a fully definite shape without understanding these issues; intelligently formulated predicates have large stretches of their borders left open.[25]

5.7

Let's review the picture that we have emerging. Vagueness is an ingredient in certain approaches to defeasibility management; defeasibility shifts some of the burden of inference to reasoners, and vagueness is a way of signaling that it hasn't been fully decided ahead of time exactly which inferences are the good ones, while allowing users an intuitive sense of how urgently they need to watch out. When we're trying to get by in the wild, vagueness in the narrow sense seems to be both widely relied upon, and something of, as they say in computer science, a cheap hack. If your ignorance is meta-ignorance, if the unknowns aren't surveyable, can even a warning that the farther you stray from the straight and narrow, the more inferential jeopardy you are in—can that be anything but a makeshift? Open texture is the acknowledgement that all manner of things of which I'm unaware might come along to defeat my generalization; how is it possible to make confident use of it?

If that way of looking at it is correct, we are not likely to have much use for a formal logic of vagueness; for much the reasons we expected a formal treatment to be out of reach for defeasible inference, it will be unavailable here. The point of vagueness in the narrow sense is to provide an inferential warning. It's easy to assume that we want to know what the warning *says*:

[25] The argument shares its structure with a move at Millgram, 2005, pp. 276f, supporting the claim that our desires should for the most part not have precise weights or strengths. Notice that here we have the right objection to epistemicism, i.e., the view that *all* concepts are completely sharp at the edges, and it's just that we'll never know where those edges are: to take on those sorts of concepts would be *thoughtless*.

It follows that *most* of the predicates that we work with inferentially should be vague and exhibit open texture; concepts that are genuinely sharp all around their edges ought actually to be quite rare.

that we want a semantics of vagueness. But that is a mistake; what we care about is what the warning *does*.[26]

You are being warned that your inferential warrant is dicey, and the farther away you are, the more alert you need to be; you're being given *that* sort of warning because more precise warnings aren't available. (We don't really know what's going on, so *watch out*.) Formal theories of vagueness are attempts to tell you, when the point is that we don't know exactly what's going on, exactly what's going on, and they're accordingly perverse. Epistemicist theories of vagueness, if that is right, are especially perverse: adding to the warning the insistence that there is a precise point beyond which you'll have gotten too far away from your central stereotype, but that you'll never know what it is, is muddying a demand for progressively escalating alertness with a useless distraction.

If we're right in thinking that vagueness in the narrower sense will appear when we have open texture, then if the division of intellectual and evaluative labor is a bounded-rationality strategy, vagueness in the narrower sense can be expected to turn up as a side-effect of that strategy. If we were right about how to approach defeasibility, won't the path to a philosophical understanding of vagueness *also* be engineering analysis? What we want is a design account both of the techniques used to quasi-delineate vague predicates, and of the methods used to flag inferences that the vagueness undercuts.[27]

However, when we are managing defeasibility produced by division of labor, it's reasonable to hope for better techniques than merely marking predicates as vague. The knowledge is out there because it has been distributed, and just because the knowledge has been distributed, we collectively have done the distributing. We can plausibly look for methods of keeping tabs on, tracking down and marshalling the distributed resource. That won't save us from having to worry about unanticipated interactions between specialized domains, but it leaves us better off than we were, staring down an abyss of massive, brute ignorance.

[26] A point made by Austin, 1975, and also analogous to that of the analysis of knowledge provided by Craig, 1990. Knowledge is a sort of generic and transmissible certificate for information; that is, it certifies the information as suitable for general use, and you can turn around and pass the certificate along with the information to further clients. Once you understand its social function, thus Craig, for philosophical purposes you can skip traditional analyses of the *meaning* of "knows".

[27] Mosdell, 2015, makes a start on a survey of the methods we use for managing vague concepts in our practice.

5.8

A last lap: Let's allow for the sake of the argument that, in the wild, the difference between defeasible and deductive inference is clear. We argued that when defeasibility and vagueness are induced by the division of intellectual and evaluative labor, they have to be seen as part and parcel of a strategic choice: they are the price of a heuristic technique, that of parceling out a task that is too much for any one person to a great number of specialists. It's time to return to a question I raised in passing, early on: whether we should be treating deductive inference, as it arises in the course of the heuristic technique, as itself a tactical choice.

Recall that our ignorance is meta-ignorance: we can't know up front when the predicates in our deductive-in-form inferences are being deployed in their vague penumbras (although sometimes we can be pretty sure they aren't). And when they *are*, the intellectual demands that deductive-in-form inferences make are very much on a par with those required for defeasible inference. To *treat* a class of inferences as deductive is, in practice, to put all of that to one side; it is to decide that, for *these* inferences, we don't need to worry, we can wear blinders, and not concern ourselves with things that might go wrong.[28]

Returning to another way of saying what it is to treat an inference as deductive rather than defeasible, the difference is a matter of where the impacts of potential and actual defeaters are located. If an inference is defeasible, we may say that the premises remained true, but we nonetheless have to withdraw the conclusion. Whereas if it's deductive, we say that the premises weren't fully true after all—or maybe, that they turned out not to be true, plain and simple. That is, the policy choice has to do with where we're committed to locating problems and with where we'll allocate blame.

Leave to one side whether there are inferences that are in them-

[28]There is a picture that is, at the level of granularity I'm about to invoke, shared by a group of researchers who among themselves don't always agree: that on the one hand, we're good at making up arguments for conclusions we want, but bad at seeing what's wrong with the arguments we've invented; that on the other hand, we're good at finding mistakes in *other's* arguments. Accordingly, high-quality argumentation is the result of a dialogical back-and-forth. (See Mercier and Sperber, 2019, and Dutilh Novaes, 2021.)

On one version of this view, deductive argumentation represents a dialogue in which so much preemption of potential objections has already taken place that the protesting interlocutor has been reduced to silence, and his very voice can be suppressed. So a different way of putting the question at hand might be, How do we determine that we will—that we can afford to—proceed as though all possible objections have been forestalled, and that there's nothing we've overlooked?

selves deductive in nature—that have an intrinsically deductive essence, as it were—as opposed to those that aren't; here we're asking about the *treatment* we will give a class of inferences. Once again, we're confining ourselves to cases that arise against the background of division of labor that is extreme enough to merit the name, not just of specialization, but of hyperspecialization. And the question I want to leave you with is whether and why, against that background, it makes sense as part of a bounded-rationality strategy to designate classes of inference for deductive treatment: in the first place, inferences for which we have settled, in advance, that we're going to let our guard down and not work at noticing potential defeaters. With all their ignorance and fallibility, why do—why *should*—specialists sometimes reason deductively?

Interlude: Compare and Contrast

We now have a clear enough picture of what we're doing, as far as methods and content go, to position it with respect to adjacent discussions. I'll be stepping through some compare and contrast with the more traditional approach to logic and theory of rationality, as well as work in evolutionary psychology, meme theory, and the recently popular program of conceptual engineering. For each of these, I'll provide some orientation, but (fair warning) not anything that would count as a proper introduction to any of them. (I'll try to make up for some of that with discussion of a small handful of extended examples.) So here I'm presupposing some familiarity both with the argument up to this point, and with the research programs from which I'll be distinguishing it.

That choice does have this much to be said for it: if you don't know about, for instance, both evolutionary psychology and creature construction arguments, you're probably not all that worried about what makes them different kettles of fish. And if you're not worried, you're welcome to skip ahead to ch. 6. But whether you're worried or not, you're welcome to come along for the guided tour.

5.9

A first contrast you will have seen emerging over the chapters so far is that between two ways of investigating what counts—or ought to count—as inference. I'll give a relatively abstract characterization of what I'm sure is familiar to many readers from an undergraduate logic class. In the traditional approach, inferences are simultaneously represented syntactically and semantically: that is, you describe, generally in an artificial language, what properly constructed linguistic objects are, and then you add rules that

specify, given how one of them is assembled, how another can be derived from it. In parallel, you offer equally artificial set-theoretic constructions to serve as models, and show that when those rules are followed, a desirable property of the linguistic objects—usually, truth-in-the-model—is conserved through the derivation. The home of this sort of treatment is mathematical logic, but the spirit of the enterprise has carried over to the recent enthusiasm amongst epistemologists for probability theory.

We need to highlight a few features of the approach that normally go unsaid. First, the investigation is conducted independently of work in other fields; it does not draw on metaphysics, or philosophy of mind, or psychology, or cognitive science; work in formal epistemology is conducted as though computational complexity theory had never happened.[1] Logic is being treated as First Philosophy, and in a very particular sense: you can arrive at results *first*, without waiting for answers to questions asked elsewhere. Second, and in my view not coincidentally, some questions which you might think would arise quite naturally go unasked and unanswered. For instance, a moment back I gestured at proofs that deductive inferences are truth-preserving, and here the obvious question is: Why is *that* the property we want our inferences to have? There are many more such questions in the neighborhood about the ways we represent how conclusions get drawn. Why propositions, as the primary building blocks or stages of inferences? Why do we default to thinking of propositions as in the first place composed of general descriptions, along with terms referring to (or variables ranging over) individuals to which the descriptions might apply?[2]

How do such investigations get bypassed? In the background, we find a received methodology, whose first phase is the appeal to already-present opinions or judgments about whatever matters happen to be on the table—these get called 'intuitions'—which in a second phase we're supposed to systematize into a 'reflective equilibrium'.[3]

[1] In our earlier discussion of the minimal self, you might have been expecting me to look forward to the day when cloud-based services will check our credences for Bayesian consistency; instead, you'll recall, at two such points I pivoted to Williams on the history of the logic of time, and to speculating about how we might augment our monitoring of actions configured on Anscombe's template. The idea that we'll navigate our way through life on the basis of probabilities rather than yes-or-no beliefs seems to me to be unlikely for more than one reason, but here's the obstacle that's relevant in this connection: as we reminded ourselves in sec. 4.6, Bayesian updating is a computationally intractable problem.

[2] Not that these are unprecedented queries; for instance, that last one is given very thoughtful treatment in Strawson, 1971. If you're familiar with it, you will notice the continuities of approach between that treatment and what we've been doing here.

[3] Rather than embark on criticism of how things have been done in the past, here I

Third, it's assumed that we're after rules or principles that hold across the board, rather than expedients that are good enough for particular occasions. (Or rather, if we allow for such expedients, it's assumed that, at bottom, what has to license them are those across-the-board rules or principles.) Those rules or principles are open and shut; they don't need to make room for exceptions. (A more nuanced way of putting the presumption: when the need to make room for exceptions *is* allowed, it's taken for granted that you can handle them by belatedly extending the methods developed for the open-and-shut rules.) Moreover, whatever the right rules or principles of inference are, they're *already* the correct such principles: we're not in the business of making them up. You can see the assumption that everything can have been accounted for up front—thus, that you can treat methods for handling exceptions as an add-on—expressed in what we can think of as a this-is-all-there-is assumption built into the models I mentioned a moment back.[4]

The intellectual ergonomics approach that I have been pursuing here differs on all these fronts. Because I've been trying to take defeasible inference on its own terms, and because the fit with traditional techniques is at best awkward, and in fact seems to me to be unpromising, I've been trying to make sense of reasoning and inference from a design perspective.[5] And because design choices draw on a background of priorities, affordances, and feasibility constraints, I've been considering the logical questions in tandem with other problem spaces that I expect to furnish some of those constraints:

---

am only highlighting that we're proceeding differently. (For the critical assessment of reflective equilibrium as a method, see Millgram, 2005, pp. 7–10, Millgram, 2023a, sec. 3, Millgram, 2025; I turn to how the use of intuitions as a philosophical touchstone appears in the bounded-rationality perspective below, in sec. 7.4.) This way of proceeding is sometimes augmented by an intensive focus on paradoxes or puzzle cases; for instance, when the subject matter is logic, the liar and sorites paradoxes have been topoi of this kind. That is in certain obvious respects in the spirit of the consistency-regime methods I've been describing here; I'll remark briefly on some differences in our upcoming discussion of Nozick. I've provided framing for the sorites paradox in sec. 5.7, and I mean to discuss the uses made of the liar paradox elsewhere.

[4]The implicit notion that there's nothing left over outside the model is perhaps a reason it comes so naturally to call partitions in some of these models "possible worlds".

[5]That is, more or less from what Daniel Dennett used to call the "design stance" (e.g., Dennett, 1989, pp. 16f, Dennett, 1981, p. 4), but with a caveat. Dennett treated attending to design and attending to psychological characterizations as distinct perspectives, and for the most part wrote as though you'd be adopting one or the other, but not both at the same time. However, in our treatment of the self, I'm considering design that is implemented (in part, and not always) using components that are discernable only from Dennett's 'intentional stance'—that is, psychological elements that are visible only when we parse something as an intentional system.

the subject areas we've seen include the metaphysics of the self, observations drawn from psychology and cognitive science about what humans can and can't do, our collective experience with database and filesystem management, as well as what technologies, both current and shortly available, make possible. And I mean to equip us to address some of those bypassed queries: Why *these* rules? Why is it truth we should use to regulate our inferences (or *is* it)? Why do we care about consistency? Why (and when) *should* we care?

5.10

To drive home how deep the difference in approach goes, I'll use as a first illustration Robert Nozick's agenda-setting discussion of Newcomb's Problem.[6] (Since we haven't already introduced it, it will be a bit of an excursion; if it's already familiar, please excuse the recap for latecomers.) In his thought experiment, though I'll take some liberties in recounting it, the space aliens have landed, and they're psychology graduate students. They have a large grant from the Galactic Science Foundation, and they're here to run experiments on us humans.

In those experiments, subjects are ushered into a room, where they're presented with two boxes, call them A and B. Box B has a window in it, and they can see what's inside: $10,000 in cold, hard cash. Box A has no window, and they can't see what's inside, but they're told what determines the contents of Box A. The subjects are then asked to make a choice. They can take Box A, and whatever's in it. Or they can take both boxes, and whatever's in them. (To reiterate, the choice isn't Box A *or* Box B; it's A alone *or both*.)

Now, here's what the subjects are told about what's in Box A. If the aliens, using their advanced theory of Terran psychology, have predicted that this subject would take only Box A, they put ten million dollars in it, also cold hard cash. But if they predicted that this subject was going to take both boxes, they leave it empty. (No, they're not out to punish greed; it's just what they need to do to test their theory, and what they promised on their application. As you can see, this was a *very* good grant.)

And here's what "advanced" means, for our purposes: the theory seems to be working very well. So far, the aliens have run a *million* subjects (or as many as you like, if a million doesn't seem impressive enough to you). Every time subjects took just the one box, Box A, they left the room jumping

[6] Nozick, 1997, chs. 2–3.

for joy that they were now multimillionaires. And whenever subjects took both boxes, they disconsolately left the room with only the ten thousand, perhaps feeling like they'd missed the chance of a lifetime.

In this thought experiment, we're asked to allow that there's no backwards causation—that is, no time travel—involved. When a subject enters the room, the money's already in the box, or it isn't. The aliens don't change what's in the box retroactively; it's all prediction, not sleight of hand. And now, let's imagine that it's your turn. You're ushered into the room. There's Box A, which you can't see into, and Box B, in which you can see ten thousand dollars. Ever since Nozick's article, classroom after classroom (but not only classrooms) have been asked: What do you choose?

The point of Nozick's story was to show how two decision techniques, which are widely relied on and whose outputs normally line up, can be gotten to produce clashing outputs. *Dominance* tells you that if you have several actions that, depending on the background state of affairs, produce different outcomes, and one action produces a better outcome than the others whatever the background state, then opt for that action. (Here, whatever's in the opaque box, you do better by taking both boxes.) The second decision method tells you to *Maximize Expected Value*: that is, multiply the values of the outcomes by their probabilities, and choose the option with the highest product; in the thought experiment, the aliens' excellent track record is there to fix the probabilities at one (or as close to it as you like), as a way of keeping the setup simple. (There's an extremely high probability of getting the ten million, if you take only the one box, and likewise, an extremely high probability of getting only the ten thousand, if you take both.) And the problem Nozick presented his readers was: which of these decision techniques, which he characterized as *principles of choice*, is the correct one to use?

Now, the first underlying assumption here is that there *is* a right way to choose: that there is, already, a correct answer. Second, there's a *principle*, and that's what we're after: something like inference rules we list for a logic, or that, for example, philosophers have looked for when trying to understand inductive reasoning. When we write down a decision problem, in a notation the correct principle presupposes (here, representing outcomes, values, probabilities, and so on), we'll be licensed to draw the conclusion the principle dictates—and licensed across the board, because it is, after all, a principle.

From the design perspective, all of that is a questionable way to be thinking about it. Certainly nothing that anyone designs is going to be suited for all situations and all use cases. We've already encountered the

notion of a heuristic (more or less, a decision or representational technique that trades off in-principle correctness for efficiency), and in the design of the virtual overlays that we use to manage our thinking, there are bound to be such tradeoffs. Accordingly, we shouldn't be surprised, when we take a second look at what we'd always assumed was one or another in-principle right way to draw conclusions (as in the case of those two principles that Nozick set at odds), to find that we're actually looking at heuristics. Suppose we see Nozick's question as actually being about interacting heuristics—not principles—and even allow that it might be heuristics all the way up: that we never get back to principles, just to further heuristics. Then the right next question to ask would be: What heuristic do we want to use to handle some such class of overlapping and inconsistent pronouncements?[7]

Notice that it had better be a *class* of them: heuristics are normally chosen on the basis of performance on average, or anyway before you encounter the particular cases, and so you want to think about the anticipated performance when there is a range of cases to which you intend to apply a candidate heuristic. If the space aliens in Nozick's presentation of Newcomb's Problem aren't coming around with their psychology experiments frequently enough for there to be such a class of cases (and of course, they *never* come around), it's not just that 'Should you take just the one box, or both?' isn't a live question: it's not yet a question to which there is a right answer.

This is an occasion to flag a practical distinction which we're bound to make when moving ahead within the present way of handling problems. In an inferential hygiene regime, on the one hand there are the consistency constraints that we're imposing; on the other, there are representations on which we enforce those constraints. When the hygiene rules are themselves inconsistent, as in this illustration, we turn to asking how to manage the consistency regime, and here we quickly start to think about payoffs. But when

[7] It's by no means the only case philosophers have noticed in which heuristic methods give conflicting answers. Chrisoula Andreou (2023) has recently suggested that intransitive patterns of preference can sometimes be accounted for via the interaction of two types of cognitive probe, which produce formally different outputs: one of them compares nearby options, and produces two-place preferences (i.e., "I'd rather have this than that"), while the other sorts options into rough evaluative bands (e.g., per my own institution's mandated evaluations for its faculty, "excellent," "effective" and "not satisfactory"). Those interactions can make one into Warren Quinn's much-discussed 'self-torturer' (1993, ch. 10).

And Cass Sunstein (2023, ch. 6) discusses inconsistent outputs that have to do with whether the choice is made by considering and evaluating one item on its own, or instead evaluating several items comparatively.

we're figuring out how to satisfy a given set of consistency requirements over a field of representations, in central classes of cases direct appeals to payoffs are ruled out of order. (For instance, you're called on to deal with incompatible beliefs by changing what you believe, not by considering whether to restrict the requirement, and not by adopting new beliefs on the basis of what you find it *convenient* to be thinking.[8])

What's more, a good deal of my own discussion to this point has concentrated on inferences. (On my psychologistic way of seeing inference, that's about drawing conclusions.) And on the traditional approach we expect there to be rules governing how those conclusions get drawn. However, I introduced the use of consistency or inferential hygiene regimes as an expedient for forcing us to figure things out. I hope it was clear enough that while some of that can be reconstructed in retrospect as step-by-step reasoning, sometimes, when you figure things out, you have to be wildly improvisational. Not everything we get done is done to specifications generated by a design exercise. Not everything we figure out is figured out by applying rules or principles, even when we understand those rules or principles to be, at bottom, working parts of one or another heuristic technique.

5.11

In the design stance, choices get made in view of constraints, so let me turn back to a topic we've already taken up, in chs. 2 and 3, to make vivid how the constraints provided by limitations on what humans are able to easily and routinely think through can figure into design-stance explanation and deliberation. (That is, we're exhibiting another of those differences in approach, by showing what it looks like to deploy one sort of bounded rationality constraint.) You'll recall that "first person authority" is our label for your *just* knowing what you think, without doing the detective work it takes to find out what's going on in other people's minds. A good many accounts of the phenomenon are in play, we noticed, and you'll also recall that one of them is detectivism: the view that maybe you don't do detective work, but you *detect*, and you don't experience it as something you work at because it comes built in. Your brain is supposed to have equipment on board that does it for you, reading and reporting on your neural circuitry, like the brain scanners in science-fiction television.

Now, Sydney Shoemaker gives a series of arguments against detectivism that take roughly this form: an individual who lacks the detectivist's

[8] See, in the first Interlude, n. 9.

as-it-were camera in the head, but who believes that p, could rationally infer that he ought to announce that he believes that p, and moreover, that he ought to act in every way (or *almost* every way) as someone who is aware of his belief that p.[9] I won't recapitulate the elaborate and ingenious arguments that Shoemaker constructs on behalf of his putatively self-blind subject, and from which that subject is supposed to draw those conclusions. What matters for now is that the arguments are complicated: in my experience, too complicated for most undergraduates to follow. Their overall drift is that, per our practices of belief ascription, we will correctly ascribe the second-order belief to an individual who acts as Shoemaker describes: he *does* believe that he believes that p. But then, by Occam's Razor, we shouldn't postulate cameras in the head to explain a capacity we can account for via rationality that is there anyway. So there are no cameras in the head, and detectivism is false.

If you're thinking about first-person authority from the design stance, this is an unsatisfactory argument, and to show how and why it is, I'll construct a parody of it. (Sorry, Sydney.)

First, some background. Long ago, before video display terminals, programmers typed on teletypes; for those who have never seen one, these were electric typewriters modified to serve as input/output devices for a computer. Since there were no screens (remember typing, on paper?), there were no screen editors, that is, text editors that show you, as modern word processors do, the contents of a file on-screen. So programmers used line editors, which involved typing commands to the editing program on the teletype, a line of them at a time; the line editor could type out, on request, a particular line of a file. Thus one of the talents required for being a successful programmer, way back when, was the ability mentally to reconstruct the contents of the file on which one was working, on the basis of one's memory of what text one had input, and the changes one had subsequently made to it. It goes without saying that it was an unusual talent; programming was a great deal more demanding in those days.

With that in mind, here is a Shoemakerian argument against the existence of screen editors (and so, against the existence of contemporary word processing applications, such as Word). One can in principle infer the contents of one's file from what one has typed, even if one cannot see the file on-screen. (The inferences are demanding, but then, so are the inferences required when Shoemaker's putatively self-blind individual mimics normal persons.) We correctly attribute knowledge of the contents of his file to the

[9] Shoemaker, 1996, pp. 34–42, 82f, 238–240.

gifted sixties-era programmer. But, by Occam's Razor, there is no reason to postulate devices like screen editors to account for a capacity underwritten by rationality that is there anyway. (The knowledge of what is in your file supervenes—Shoemaker ought to say—on your rationality.) Therefore, there are no screen editors, and if you are only going to use a computer for word processing, you do not need to bother purchasing a monitor.

The parody shows, first, that Shoemaker's argument ignores the cognitive limitations of human beings, and how easy it is for overly demanding tasks to degrade their performance. It incidentally reminds us that Occam's Razor has to be applied with caution, because artifacts or organisms designed by natural selection are not necessarily constructed in accord with the design principle that Shoemaker takes the Razor to express. (We are all born with an appendix, whatever Occam's Razor might suggest.) More importantly, it alerts us to the need, when designing procedures for human beings to follow, to take their competences into account. (In this case, when it's something we're requiring of pretty much everyone, complying shouldn't depend on extraordinary talents and intellectual gifts.) In any case, the argument we were responding to is unsuccessful: some self-knowledge may, as far as the argument goes, be a product of the elaborate inferences which Shoemaker suggests; but most of it almost certainly is not, and may still, for all we know, be the product of a detectivist's camera in the head.

5.12

One last illustration for now: I've repeatedly suggested that if, say, you believe p, and also believe $\sim p$, we expect you to change your mind when you're called on it, because you're thereby inconsistent. I've contrasted that with mere disagreement: when you believe that p, and someone else believes that $\sim p$, no one is thereby inconsistent, and no one is thereby required to revise their beliefs. (Most times, I've added a "so far" qualification to that last claim, which I hope you've been registering.) But now, there has been a good deal of recent discussion of the so-called problem of peer disagreement: more or less, if someone is epistemically as well-placed as you are, and begs to differ, shouldn't you change your mind about whatever it is? By and large the discussion proceeds by marshalling intuitions about a handful of toy problem cases: members of a party at a restaurant who have totted up the bill differently, for example.[10]

From the bounded-rationality perspective, this back-and-forth has

[10]See Frances and Matheson, 2019, for an overview.

been conducted in a kind of theoretical vacuum. (For instance, even the basic conceptual toolkit of computational complexity seems to be missing.) It is not that we can't ever rely on others; we have to, and the epistemology of testimony has quite rightly also been much discussed in recent decades. But looking back over our shoulders at the creature construction argument we've started, we can see it as a transcendental argument for ignoring what *almost* everyone thinks: if you treat other people's mental states as deserving of consideration generally—even just other people who are epistemically well-placed—you'll be faced with a task you can't complete. Partly that's because you don't have endless time, and partly it's because when there are too many deeply different considerations to be folded together, you won't normally be able to see how to resolve a question on which they're brought to bear. In Kantian phraseology, it is a necessary precondition of the possibility of having a mind that you ignore almost everyone's opinions. We *have* to treat ourselves and others very differently; where we try to live up to the demand that we mitigate our internal inconsistencies, we cannot—with rare exceptions—aim for consistency with others.

Keeping hold of the bounded rationality frame, let's push a little further: *Which* exceptions? Epistemic injustice is also recently much-discussed, the drift being that there are whole classes of people whose opinions go unheard—and get disregarded when they *are* heard. (If you wanted to recast the complaint in a nonacademic register, it would sound something like this: They don't *listen* to us, and it's not *fair*!) But now, if in order to have a mind, you have to ignore almost everybody, you will also have to be arbitrary about it: an investigation of who is to be listened to on the merits is no more possible than actually listening to them. Epistemic injustice is also a necessary precondition of the possibility of having a mind.[11]

5.13

The creature construction arguments we've been using likely strike some readers as dangerously perched between two other research programs sharing a large and important common denominator, and quite possibly open to attack from both sides. On the one hand, there is the theory of natural selection and biological evolution; on the other, there is, although here I'm riding roughshod over terminological fastidiousness on the part of certain participants, meme theory. The shared idea is that structure emerges in, respectively, the biological and cultural worlds through the operation of

[11] For an opposing view, see Hills, 2010, p. 157.

selection pressures on replicators: on organisms making more of themselves, and on snippets of culture getting themselves copied from mind to mind to mind.

Evolutionary psychologists have provided their own sorts of reasons for holding that human beings work with niche-bound heuristics, and it's worth taking a moment to put on display the apparent overlap in topic that could lead someone to construe what we've been doing here as that sort of exercise. Accordingly, I'll recapitulate a couple of related trains of thought from that world.[12] To perform tasks successfully, you have to deploy a success criterion. But candidates for something that might count as a success criterion across the board are unobservable in real time. (These researchers are thinking of reproductive fitness, which for an evolutionary psychologist is the be-all and end-all; you can only tell how many great-grandchildren you ended up having much, much later.) So to be able to perform tasks successfully, you need success criteria that are domain-specific. Now, domain-specific success criteria will leverage idiosyncratic features of the domain: they'll be as qualitatively different from one another as you like. This means that successful task performance will be focused on idiosyncratic features of a local domain.

Next, problem-solving strategies succeed by anticipating and exploiting features of the problem domain. And the more narrowly defined the domain, the more features problems in it will share; because the common denominator of *all* problems is very, very little, a 'General Problem Solver' will have nothing to work on. So a successful problem-solving strategy will target the narrowest feasible domain. Now, for practical purposes, success in solving a problem has to be timely, but doesn't always have to be exactly right. A strategy which only works sometimes, but when it works, gives good-enough but not always exactly right answers—that's a good way of describing a heuristic. And this means that successful problem solving strategies will be (almost always) locally-suitable heuristics.

Here the important point is that these arguments are meant to show that natural selection must have produced domain-specific, special-purpose cognitive procedures or modules; thus, because we too are products of evolution, that is what we will find when we look more closely at ourselves. Whether or not you think the arguments are successful, what they are about is how we got to be the way we are—as a matter of historical fact and its causal explanation. While you could no doubt repurpose these considerations into the pieces of an appropriate creature construction argument, and

[12]Tooby and Cosmides, 1992, pp. 111, 104.

while it's tempting to leap to the conclusion that the arguments must accordingly be, really, the *same* argument, how they work and what their respective points are do not line up that way.

To someone coming from evolutionary biology, creature construction arguments are likely to come off as overly adaptationist evolutionary speculation. And that's grounds for suspicion and maybe already an objection. Natural selection doesn't look ahead to see how some problem would get neatly solved and then aim itself at the solution; pointing out that some trait is advantageous to an organism, and concluding, right then, that that must be why it's there, is to make a very basic mistake about how evolutionary explanations work. But here is not the place to talk through how evolutionary psychologists field this sort of complaint, when it is directed at them.[13] The objection we've just spoken to—Why aren't creature construction arguments just more Just So stories?—is simply a different issue.

And from the other flank, recall that the structure we're focused on is the self: a central element of our intellectual ergonomics, which flags occasions for drawing conclusions and more generally for figuring things out. I've been making the case that this device is enormously important, because our practices, technologies, and much else are largely the accreted results of one person after another being forced to draw a conclusion or figure something out. But meme theorists take themselves to have a much more parsimonious account of where our culture—and within it, our repertoire of solutions to problems—mostly comes from: differentially prolific transmission of memes (those cultural snippets) does most of the work, no intelligence required. Some memes (e.g., using "enormity" as a syonym for "largeness") get copied more; others (using it as a synonym for "atrocity") get copied less; the language changes. Other clusters of memes converge on, in one vivid example, the design of an avocado peeler. Problems get solved, but without anyone having to think them through.[14]

Under the first heading, the theory of evolution and creature construction arguments are answering different questions. The former addresses an etiological question: how did the state of affairs we find come about? For

[13]However, there is a delicate point to register. That accusation used to be couched as being a matter of illegitimately allowing teleological explanations into one's evolutionary biology. Now, "teleology" is derived from *telos*, that is, *goal*, the idea being that one is assuming that natural processes are goal-directed. Built into that way of expressing the concern is a mistaken assumption, that all practical considerations, i.e., drivers of decision, take the form of goals. The worry that practical considerations are slipping into scientific accounts where they don't belong should not be tied to that particular naivete.

[14]See Boyd and Richerson, 1985, Richerson and Boyd, 2004, Sperber, 1985, Henrich, 2016. The meme concept is introduced in Dawkins, 1989.

biological organisms, when we ask the question, "How did they get this way?" the issue of how much room we need to make in our explanations for notions like function and purpose has been fraught for a long time. But creature construction arguments speak to a different sort of question entirely: What do we want it for? What does it do for us? And here, thinking about function and purpose are quite at home.

Under that second heading, no doubt meme selection explains a good deal of our cultural heritage. But I don't recommend discounting the exercise of intelligence, as scaffolded by the sort of intellectual ergonomics that we are investigating. Reasoning, in my way of seeing things, is the high-speed alternative to the dangerously slow processes of meme selection; it is an intellectual *accelerator*. We devote resources to it because it's a faster way of solving problems than the selection-based alternatives.[15]

As a way of making that point, allow me to sketch a thumbnail history of life on Earth. (To make it a thumbnail, of course I'll ruthlessly simplify; please don't complain that I've left things out.) Early on, when it was only unicellular organisms propagating by division, the problems and challenges those organisms faced were solved through a combination of accidental mutation and natural selection: the organisms lacking the fix that met the challenge were eliminated, leaving only those that happened to have it. However, this way of solving problems is slow and wasteful, and to see why, consider a bacterium today, faced with two different antibiotic regimens. Maybe there's a lucky mutation that provides resistance to one antibiotic, and maybe there's another lucky mutation that provides resistance to the other. But if the mutations don't occur in the same place, the challenge hasn't been met; the descendants of each will have to wait until the other mutation reoccurs in its own lineage, and that may take many, many generations.

Sexual reproduction accelerates that process, by allowing organisms to mix and match genetic material. Sexual reproduction is widespread among larger organisms because it is a way of solving problems much more *quickly*; it's not a coincidence that asexual reproduction is mostly confined to mi-

[15] A preemptive correction: The recent enthusiasm for so-called System One/System Two categorization of cognitive processes characterizes reasoning as 'slow,' contrasting it with 'fast'—that is, rapid-fire—heuristics. (Kahneman, 2011, is the most familiar popularization.) The dichotomy is confused in many ways—in its own terms, leaping to classify our complicated and kludgy cognitive activity under two easy-to-remember headings is System One thinking—but in particular it gets its own most basic point backwards. To repeat myself, we engage in reasoning in the first place because it's on average faster than solving problems by evolving hard-coded reflex-based responses.

croorganisms. In the always ongoing arms race between organisms, very short generation times speed up the process of meeting challenges; but in a very short lifespan, it is only possible to construct a very small organism. The leisure to build larger organisms provided by longer generation times is enabled by sexual reproduction's speedier solutions. However, even sexual reproduction is, by our lights, a painfully slow expedient for adjusting to challenging circumstances. And so, at some point in our history, you see creatures that do better, by managing operant conditioning; conditioned reflexes allow organisms to adapt themselves to their environment much more quickly than by letting natural selection adjust the hardware, over generations upon generations.

Skipping ahead many steps, at some further point, and now we're already attending to human history, you do encounter problem solving by meme evolution. But in a manner that's analogous to the snail's-pace progress of biological evolution, in both its asexual and sexual versions, it takes a long time for memes to settle out into relatively stable clusters that handle some problem or meet some challenge. And at this point in the story, it looks like the history has been a long, long competition, in which the edge goes to techniques that solve problems more quickly than before. With that backstory in view, techniques that turn on imposing consistency requirements are just the next step in the long race for speed. The primary advantage they offer is faster solutions than meme selection allows.

5.14

Conceptual engineering is advanced as, roughly, the enterprise of reshaping concepts currently in use, as well as introducing novel concepts. Although its roots go back to the logical positivist era, it's a suddenly popular program, and the sound of the label its votaries have adopted is close enough to that of intellectual ergonomics to make them worth distinguishing.[16]

[16]See, for instance, Cappelen, 2018, Chalmers, 2025. We already find the notion in Carnap (Quine and Carnap, 1990, p. 412, and then flagged with that very phrase in the Introduction).

In Carnap, conventions are used to introduce a logic, and if the position I've been developing strikes you as Carnapian, you may have been worried about whether the outcome of the Quine-Carnap debate closes it down. As you likely remember, what was typically taken to be the decisive move was the observation that conventions have their consequences only by way of logic: the account, Quine was saying, was viciously circular. (A one-paragraph summary of the argument from Quine, 1936, can be found at the start of §4 of Quine, 1960.) Here the task is not definition, but constructing and operating virtual objects and the procedures they support. I'll return to and fill in this point below, in

I said 'roughly' a moment ago, because there's a bit of terminological hedging, accompanied by repeated attempts to qualify the claim that what's to be operated on are *concepts*. For example, at one point we're offered, as a more careful substitute formulation, the activity of "assessing and improving our representational devices."[17] What matters for our present purposes is that the focus is on the tokens we move around in doing our thinking and talking, rather than on the mechanics of the virtual system through which we are maneuvering them. The tokens may be terms in a language, or they may be concepts, as Fregeans conceive of them (abstract objects, which are expressed by predicates under which individuals do or don't fall); they may be generics, or as they're sometimes also called, Aristotelian categoricals; there are indefinitely many alternative ways of constructing representations.[18] Herman Cappelan, who is these days perhaps the most visible participant in the discussion of conceptual engineering, coyly announces that "the project isn't about concepts and there isn't really engineering. . . [but f]or all those who think there are concepts and that they can be engineered," 'conceptual engineering' is "descriptively adequate".[19]

I've suggested that metaphysics is, and unbeknownst to itself really always has been, intellectual ergonomics. By this point in our exposition it will be clear that what is in question is the virtual system—selves, persons and so on—that processes representations, and not the representations themselves (whether you think of them as concepts or some other sort of representational token). To be sure, when we design, introduce and then adjust what I've been suggesting are the proper objects of metaphysics—where selves are, again, the case on which we've been concentrating here—we'll also need to supply terminology for them. But the focus is on the virtual objects and how they work, not on our names for them.

Here I'm putting to one side here the somewhat Orwellian cast that the conceptual engineering program can take on. You should be aware that some of the more prominent participants propose redefining culturally sensitive terms in the service of various explicitly political aims. "Newspeak"

---

sec. 8.12.

[17] Cappelen, 2018, p. 3; at another point, we're told that "conceptual engineering should be seen as having as its goal to change extensions and intensions of expressions" (p. 61).

[18] For the Aristotelian categoricals, see Thompson, 2008, Part I.

[19] Cappelen, 2018, p. 4, and here are a few indicators, to help you home in on this way of thinking: Cappelan tells us that "at the center" is a "a theory of that which makes it the case that expressions have the meanings they have" (p. 7). In a series of examples, offered as reference fixers for the concept of conceptual engineering, he surveys proposed redefinitions: of concepts like 'race,' 'woman,' 'knowledge,' and so on, as well as proposed prohibitions (e.g., on using terms like "Muslim").

was a central component of the technology of political control in Orwell's dystopian novel *1984*: a regimented vocabulary in which citizens would be unable to say, and so unable to think, ideas or views prohibited by the regime.[20] What is more, we find the recurring ambition to alter the socially constructed parts of our world by reengineering the concepts used to talk and think about them. The Party that governed Orwell's fictional totalitarian state had as a motto: he who controls the present controls the past, and he who controls the past controls the future. Here the stance—and at the level of attitude, it is strikingly similar—is evidently that he who controls the concepts, provided that what they are concepts of is socially constructed, controls the things themselves.

To these linguistic and conceptual legislators, no doubt the program does not seem of a piece with Orwell's dystopia, I imagine because they believe, and believe in, the ways of thinking they're trying to promulgate. But to the people on the other end of such exercises of control, who don't believe, they will be experienced as, precisely, attempted exercises of control. It's worth remembering that Orwell's savage parody was a response to what he regarded as the abuses of language by, especially, well-meaning political activists of his own day.

Again, I don't want to make a great deal out of this; not nearly all of the declared conceptual engineers have this sort of agenda. But this contrast is worth keeping in mind: intellectual ergonomics is in the service of improved performance in problem solving generally; I am not promoting any party line that you might recognize on the current political spectrum. Intellectual ergonomics is certainly not an attempt to enforce political conformity on anyone, and certainly not by artificially inducing representational blindspots.

Concepts, along with representations in other formats, are what the program of conceptual engineering seeks to modify.[21] The dissimilarity with

[20] Orwell, 1950. Do they really think that what you don't have words for, you can't think, and what the words you use have built into them, you *have* to think? Here is Cappelan quoting a contribution to a journal devoted to state-of-play surveys: "Arguably, our conceptual repertoire determines not only what beliefs we can have but also what hypotheses we can entertain, what desires we can form, what plans we can make on the basis of such mental states, and accordingly constrains what we can hope to accomplish in the world." (Cappelen, 2018, p. 43; from Burgess and Plunkett, 2013, pp. 1096f).

[21] But registering one more bit of coyness, Cappelen, 2018, ch. 12, makes a big deal of the obverse of semantic ascent—we might as well call it semantic descent—in order to put himself in a position to claim that his "version of conceptual engineering... is directly about the world" (and why? evidently to bypass what he regards as the problematic notion of a concept: "there are no concepts to fiddle with, and there isn't any fiddling").

intellectual ergonomics that I'm highlighting here is that we are taking a design perspective on the procedures we execute in processing those representations. In particular, the more powerful and interesting techniques deploy intellectual accessories, especially the virtual overlays we induce to facilitate those procedures: first and foremost among these is the self. And our primary interest here is in these. If you like, the contrast resembles that between trying to understand and perhaps improve on the machinery on the factory floor, as opposed to taking the factory for granted, and confining one's attention to the manufactured products, as they move along the assembly line.

The monograph that I've mentioned as the current anchor of the debate has as its subtitle, "An Essay on Conceptual Engineering". It has as its title "Fixing Language." That says it all: bottom line, conceptual engineering is about language, and intellectual ergonomics isn't.

5.15

Selves are in the first place scopes for consistency regimes.[22] When the forms of consistency required of them are cut-and-dried (i.e., deductive), the operation of the device is conceptually straightforward. But over the past couple of chapters, we've homed in on the fact that we're mostly in the business of administering soft (i.e., defeasible, other-things-equal) consistency constraints. These are evidently a good deal more challenging, both as far as understanding them, and as far as implementing them go.

There has been a certain amount of foreshadowing of what I am starting to present as an urgent problem: that in a social world characterized by constantly escalating division of labor, our ability to meet these demands for consistency is being eroded. If I am right, it either already is, or soon enough will be no longer good enough. But before we face that problem full on, we need a better picture of the strategies we've inherited for living up to our various inferential hygiene mandates. To be sure, an exhaustive survey isn't in the cards, and we'll aim for a small handful of case studies.

One of these is already in place. Vagueness, we've seen, is to be positioned as an improvised defeasibility management technique, and notice that vagueness counts as falling under the subject matter of metaphyiscs

For the purposes of the distinction I'm making, his semantic descent doesn't matter.

[22]'In the first place': of course, a great deal more has been built on top of that, but what and how is best taken up in a followon work, which will treat specifically practical rationality, and the moral and social arrangements in which selves figure.

as we're construing it—that is, as intellectual ergonomics. Let me take a moment to explain.

In this book, I've been concentrating on the self, treating it as a type of virtual object overlaid on human beings, and facilitating intellectual and indeed creative productivity. But there are other devices that fall into this augmented-reality category, and are virtually located elsewhere. Think of Aristotle's *ousiai*, though for the moment, just in the stripped down version introduced in his *Categories*.[23] When we look at, say, a horse, we understand it to have a remarkable feature: if it gains a few pounds, it's still exactly one horse, not, say, 1.16 of a horse; if it loses a few pounds, it's remains exactly one horse; if it's just a foal, it's nevertheless exactly one horse. Substances, to allow the misleading translation for the moment, come with edges: there's a pretty clear boundary delineating what's counts as the horse from what's outside it, and those boundaries are temporal as well. (There's more or less a moment when it comes into existence, and also more or less a moment when it's gone.) We count human beings—persons—as substances, and so when you ask how many people are present, you know the answer is going to come back as a whole number; and when you buy plane tickets, you expect them to be priced per person, and not, say, by height in centimeters. All of that is an enormous intellectual and practical convenience.

Of course, not every chunk of the natural world will take this kind of overlay cleanly enough to permit the intellectual economies it allows. What about those chunks that the substance overlay doesn't fit well enough? Vagueness is an aspect or feature of an alternative overlay, for some of the cases in which Aristotelian substances aren't in order, as well as for predicates or properties that resist clean answers to the question of whether, in particular cases, they apply or not.

But there is more to how we've gotten along so far, in environments that call for defeasible consistency regimes. In the coming two chapters, I'll turn to the way that selection effects have supported the needed competences, and then to how persons—selves that persist over time—appear in the next lap of our creature construction sequence, and appear as a method of coping with defeasibility.

[23] Aristotle, 1984, vol. I, pp. 3–24. I'll return to his innovation briefly in 8.3.

Chapter 6

Selecting for Defeasibility

The argument in linguistics went like this: Young children learn the grammars of their native languages very quickly, on the basis of what looks like extremely scanty data. So there must be a special-purpose language-learning module, a part of the brain that endows us with the magical ability; it knows what grammars can look like, and looks out for the signals that cue it as to which grammar in particular it is being exposed to. The remaining problem, over and above figuring out precisely how the magical module worked, was explaining how it could have evolved; it is not easy to see why natural selection would have given rise to such a device. The rebuttal was that we were looking in the wrong place. What is being selected is not in the first place organisms, for their magical language-learning ability, but the bits and pieces of languages, for their learnability. Languages are going to be made up of elements that young children, however they already are, find it easy to learn, and that's why language learning happens so quickly. We don't need evolutionary accounts of magical modules if there's no magic to explain.[1]

Here I'm going to take this back-and-forth as a model for a reconsideration of defeasibility. After reintroducing the concept, I'll explain how it is that our ability to manage the defeasibility of our inferences can look like a superpower, a capability that outruns the cognitive resources available to support it. The temptation to treat it as a mysterious logical endowment, one we can try to describe but which we must just take to come with our intellectual toolkit, is almost as old as philosophy itself. But I will try out the idea that we have been misconstruing what needs to be explained because we have been looking in the wrong place. It is not that we have an impossible talent for anticipating all the ways our reasoning could be dis-

[1] Deacon, 1997, ch. 4.

rupted; rather, the working parts of our technological and social world have been selected for our ability to anticipate what might go wrong with them.

6.1

Children in my generation spent their Saturday mornings watching *Road Runner*, a cartoon in which a ravenous coyote's schemes again and again fail to take into account some aspect of their own workings, or of the surrounding landscape; they invariably implode into slapstick disaster as, say, his catapult launches him at a cliff, while his prey whizzes serenely by. And if you grew up on pulp sci-fi and double-feature B-movies, you are equally familiar with the theme of the unanticipated and catastrophic side effect: the nuclear rocket that somehow destroys the Earth's atmosphere, the robotic spaceship that induces surreal nightmares in its crew, the industrial pollution that makes marine life attack the shipping lanes, and on and on and on. In this genre, it's hard not to become impatient with the clunky writing, the thinly developed characters, and the morals being so heavyhandedly drawn, and thus to overlook the underlying formal point, which is absolutely on target—and the human limitations which account for it, the depiction of which is nearly on target.

When inference is not deductive, it is *defeasible*: even though the argument is fine as it stands, you could nonetheless come by further information or assessments—that is, *defeaters*—that would not impugn its premises, but still require you to withdraw the conclusion. Often enough but not always, defeaters are side effects you need to anticipate, and taking a real-life example with somewhat the flavor of all that old-time sci-fi, a while back regulators and policymakers were asked to endorse roughly the following bit of reasoning. Crops that are genetically engineered not to be susceptible to a particular herbicide, if widely used in tandem with the herbicide, can simplify industrial agriculture and achieve a substantial reduction in costs. (The farmer plants the genetically modified corn, applies the herbicide, and doesn't have to worry about weeds.) Of course we want to simplify processes and reduce costs. So we should approve both the modified crops and the application of the herbicide.

The inference makes sense, and underwrites its conclusion, but can be derailed by any of indefinitely many considerations, and I'll start off the list without telling you which of them became live issues, and which are figments of my overactive imagination. If the regime were to become sufficiently widespread, the milkweed that monarch butterflies eat as larvae

would be eradicated, posing an extinction threat to an iconic species. As usage of the herbicide increases, carcinogenicity will become ever more evidently a problem. Widespread use selects for herbicide resistance, entailing the need to constantly escalate dosage, and making the life of the product intrinsically short term. Approving the procedure would strengthen incentives to develop and sell crops that do not reproduce themselves in the wild; i.e., farmers unable to reserve seed corn for the following year will have to purchase it anew each year. But that sort of dependence on a sole supplier means a concomitant increase in the fragility of the agricultural sector. The gods resent humans meddling with their handiwork, and would wreak devastation, in the form of plagues of gigantic crosseyed weasels. And it's obvious that any science-fiction author worth his salt could extend this list as far out as you like; this is by no means the full roundup (sorry!) of the potential defeaters for the inference we were attributing to the gatekeeper agencies. Once any of these possibilities turns live, it makes second thoughts appropriate, and as it happened, most of them unfortunately came to the attention of governments and the public only after the fact.

But because almost anything we do is underwritten by (usually tacit) inference, and because deductive inference, in which the truth of the premises guarantees the truth of the conclusion, is rare, defeasibility management is not confined to the more lurid sort of science fiction, or to matters of technological innovation. This morning I told myself that a double shot of espresso would jump start my day and get me writing; that I mean to begin writing this very paper and make good headway on it; that I can't make a proper espresso at home; and so that I should step out to a coffeeshop, first thing. There were indefinitely many potential defeaters for the train of reasoning. The day is too beautiful; I should spend it outdoors with friends. On a weekend morning, the crowds make the coffeeshops hard to work in, and the lines too long. A caffeine addict like myself is likely to find that he has downed too many shots, and end up jittery and unable to sleep. And, even in the real world, there are live potential defeaters that sound like pulp sci-fi. A few years back, coffeeshops were a great place to contract possibly fatal respiratory infections. As I write, if you happen to live in Kiev, drone and missile attacks are a live possibility. Once again, it's clear that you can extend the list of potential defeaters as far out as your inclination and imagination suggest.

Briefly, defeasible reasoning is ever-present, and so the need to manage defeasibility is pervasive. As further experimentation will confirm, for any such inference, the well of potential defeaters has no bottom—no matter how many of them you have noticed, there are always more to be found.

6.2

A defeasible-and-defeated inference is one whose conclusion you should not draw. A setup for conclusions where you have no—or *almost* no—ability to discriminate between conclusions you should and shouldn't draw is worthless. So defeasible arguments are worth constructing and deploying only to the extent that you are alert to the occasions on which they are defeated. But how is that possible?

As the first of our two illustrations indicates, often we're insufficiently alert, and both of them suggest reasons for expecting that to be inevitable. Potential defeaters don't run out, and as far as, if you like to think of it that way, subject matters go, they can come from just anywhere—from out of left field. Some of the problem evidently is a matter of not taking into account facts and assessments that no one has yet been in a position to, respectively, discover or make; another part arises out of the intellectual and evaluative division of labor (*someone* has the relevant facts, but you don't); yet another part, from anyone's inability to register the salience of all the various things he knows or the assessments he has made to what he is thinking about at the moment. And so you would expect normal human life to follow the narrative trajectory of the pulp novels and B movies with which we began—or to be an endless episode of *Road Runner*, in which almost every expedient is accompanied by comically unforeseen consequences.

In those sci-fi flicks, impending catastrophes are forestalled at the very last second by improbable heroics, and in *Road Runner*, the inept protagonist's mishaps are never actually fatal. Wile E. Coyote emerges from beneath the gigantic boulder that has crushed him looking no worse than rumpled, but if you and I were as inept as he is we would not be here. If it helps, think of it as an anthropic argument: were we not good *enough* at anticipating what might go wrong with the courses of action on which we embark (and we'll consider in due course how high, or low, a bar that is), we wouldn't be having this discussion. Here we are, so despite the prima facie reasons for this being a task we shouldn't be able to perform, we must after all be pretty good at identifying potential defeaters for our inferences, and balking at their conclusions when the defeaters-in-waiting are defeaters-in-fact. Lo and behold, I did indeed step out to that coffeeshop, having figured that the line would not be *that* long, and having dismissed the chances of consequently dying of covid as being too low to worry about—but did not end up experiencing one of the surprises delivered by Mr Coyote's many

Acme products.[2]

Aristotle made the *phronimos*—the practically wise person, who in considering a practical syllogism is properly sensitive to any and all of the considerations that might be relevant to his deliberations—into the criterion for correct action.[3] And ever since then, philosophers have been all too willing to take for granted an apparently inexplicable ability to control the defeating conditions of their reasoning-in-progress. But inexplicable superpowers are never legitimate explanations, and especially not here.[4] Defeasibility is a way our logic accommodates our bounded rationality and our finitude in other regards: our cognitive limitations mean that we overlook things we already know, and there are a great many things we simply haven't yet seen, and don't already know.[5] That is, the problem we're looking at arises because we *don't* have superpowers, so the solution we propose had better not amount to insisting that we *do*. We ought to be completely incompetent at defeasibility management, and yet we seem to be get by well enough: thus our present puzzle.

6.3

To make further headway, we need a term for this feature: that of enabling sensitivity on the part of a reasoner to the defeating conditions of his inferences—where we're after, not a property of the *reasoner*, but of what he's reasoning *about*. I can't see a way to get all of the mouthful into a just one word, so let's say that an institution or practice or technology is *defeasibility enabling* if the people who interact with it and are familiar with it become successful (enough) in vetting inferences having to do with it for their defeating conditions. In other words, and more directly: it's defeasi-

[2]For anthropic arguments in cosmology, see Barrow and Tipler, 1986, and for discussion, Roush, 2003.

[3]For discussion, see McDowell, 1998, ch. 7, Millgram, 2005, ch. 4.

[4]Not legitimate elsewhere, either: the mistake would be roughly that of so-called strong anthropic arguments (see n. 2, above), and let me explain the distinction I'm gesturing at with an example. When you look around, you see a world that is neither too hot for life, nor too cold, one that provides a breathable atmosphere, and so on and so forth: overall, your world is very friendly to life. Why is that? The strong anthropic explanation is that the cosmos likes you—it's warmly and fuzzily benign. Entertaining such notions is what has given the Anthropic Principle its cranky sound, but the weak anthropic explanation, which is entirely respectable and a basis for low-key analogs of transcendental arguments, is that if your planet were not livable, you wouldn't be here asking the question. *Of course* the world you see will seem friendly.

[5]For the argument underlying this point, see ch. 4, above.

bility enabling if its users are able to anticipate what might go wrong with the arguments they put together about it, and under what conditions they might have to take back their conclusions.

Notice, first, that this property is, formally, a secondary quality. We understand color, the traditional exemplar of that class, by way of the responses of observers deploying the apparatus of human vision; we understand what it is to be funny—another more recently discussed exemplar—by thinking about senses of humor and what makes people laugh. And here's a more recent characterization of a secondary quality: "an 'edge' from the point of view of the visual system does not correspond straightforwardly to any known physical property of images, and theorists do not assume that the environment consists of 'edges' that the visual system is attempting to detect. Rather, the idea of an 'edge' is a psychological notion, and it is the *product* of psychological processes. It is not possible to understand what an edge is without understanding the properties of the human visual system—edges do not fall under any physical description of the world."[6]

In our own discussion, we're now picking out a property of objects with which we have to interact by way of what we can do when we reason about them. Secondary qualities are dispositions (specifically, dispositions to produce responses in those who deal with the object that has the quality), and dispositions are generally underwritten by their implementations. Until very recently, no one knew at all how color was implemented, and likely no one really knows why people laugh at what they do; we use these dispositional concepts nonetheless, and that is the situation we are now in with regard to the concept we're introducing.[7]

Notice also which items we're focusing on: you can think of institutions, practices, and technologies as the medium-scale working components of a society, and they're what's going to figure in the upcoming argument. If instead you asked whether your fellow citizens were themselves defeasibility enabling, you would perhaps be starting to investigate an overlooked virtue: one's being such as to allow other people to anticipate under what conditions they'll have to take back conclusions they draw about one. The trait is probably important enough, but the topic is one for a different occasion.

[6] Oaksford and Chater, 1998, p. 151.

[7] For this reason, secondary qualities are sometimes also called "response-dependent properties". Hardin, 1988, and Gold, 1993, are dated but still useful starting points for philosophers wanting to find their way around what underwrites color perception; Hurley *et al.*, 2011, is a recent attempt to say what makes things funny. Wiggins, 1991, generalizes the concept of a secondary quality past the sort of sensory properties for which British Empiricists had formerly developed the notion.

Now, the deployment of practices, institutions, and technologies normally involves inference, that is, a participant or user drawing conclusions about what will happen when he, say, gets in line, or logs on to a server, or registers for a conference. It follows that if he is unable to anticipate sufficiently many of the conditions under which those inferences won't go through, he'll experience failures that range from things merely not working to utter disasters. For instance, in the former category, I'm sometimes a bit slow on the uptake, and it hadn't occurred to me that, since the last time I was overseas, the utilities, web hosting services and so on that I log onto remotely would have begun blocking foreign access, no doubt on the assumption that anyone trying to reach them from outside the US would be either a criminal or a hostile state actor. I had formerly concluded that I could pay my bills online while travelling, and having failed to identify the defeating condition ahead of time, was faced with a series of procedures that I couldn't make work. In the category of disasters, the security precautions I just mentioned are a desperate attempt to cope with an IT environment in which unanticipated defeaters for one's inferences can all-too-easily threaten the viability of an institution. A representative pending example: a couple of decades ago, my university, like many others, made grading a task that instructors perform online. It should have been obvious at the time—but wasn't—that eventually changes to students' grades, and indeed academic certifications of all kinds, would be sold on the dark web: as we now realize, once a process is made accessible online, it can't be secured. Certification is one of the basic functions of an academic institution; what is a diploma or a transcript worth, once no one can tell the ones that have been earned from hackers' edits to the registrar's database?

Looking at that link from the other end, if an institution, practice, or technology is not defeasibility-enabling, it will often enough prove dysfunctional—where, again, the failures range from things merely not working, to catastrophes. But if there's a way people do things that fails frequently and badly enough, the people who do things that way are going to be sidelined; either they themselves will look for another path forward, or, one way or another, they'll lose access to various resources. And when the users of or participants in a technology or institution or practice come to be sidelined and have less in the way of resources, that technology, etc., will be deployed to a lesser extent in the sequel. The upshot is that when practices, institutions, and technologies are not defeasibility-enabling, over time they will tend to fall into disuse and be displaced by alternative ways of doing things.

If that is correct, then if you live in a society in which this win-

nowing process has had time to do its work, when you look around, you will see practices, institutions, and technologies that are preponderantly defeasibility-enabling. Since your environment is mostly your social environment, such practices and so on will be a very large part of what you do your reasoning about. So by and large you will be pretty good at anticipating defeaters for your inferences (anyway, inferences with this range of subject matter); you will appear, and perhaps feel, canny and positively gifted in this regard, and be suitably confident in your knack for getting things to work. But, and now reversing that cliche about giving children in competitions magic amulets and such, the magic is not in *you*: it was in the selection effects we have just been sketching, all along.

6.4

How reliable—and how necessary—is the selection effect which I've just suggested accounts for much of our success in defeasibility management? Let's pause to consider a handful of objections. The first, to the effect that there isn't all that much success to account for, can be made out on the basis of examples I've already floated: no one seemed to have anticipated the problems that arose around developing herbicide-resistant plants, for instance. A second pulls in the opposite direction: I have found repeatedly that conversation partners and audiences (and reviewers) find it hard to believe that we overlook and generally fail to notice defeaters for our inferences, or that, when we do, it will too often turn out to matter a great deal. A third is that when people do overlook defeaters, very often the explanation does not have to do with the difficulty of the task, but rather with incentives: as Upton Sinclair once put it, it is difficult to get a man to understand something, when his salary depends upon his not understanding it.[8] We've all encountered what looks like wilful ignorance on the part of professional advocates, or managers who don't want to be the one to tell their boss why his idea won't work.[9] The university administrators who moved grading online were rewarded—it counted as an 'accomplishment,' to be listed on annual activity reports and resumes—and not about to listen to naysayers. And last for just now, won't the scope of the explanation be too narrow? What about the natural sciences, where we're evidently able not only to consider defeaters, but to decide pretty definitely when we've considered enough of

[8] Sinclair, 1935, p. 100.

[9] For a case study of some of the professional advocates, see Oreskes and Conway, 2010.

them, and that it's time to move on?[10] But nature in the raw can't have been shaped by the selection processes we've just proposed.

Under that first heading, there are two points to bear in mind. For one thing, while we make mistakes and while there are a great many things we don't notice, perhaps we're willing to complain about the inferential choppiness we encounter because most of the genuinely awful catastrophes are indeed preempted by the selection processes we're contemplating. Their effects have ratchetted our standards up, by making us lose sight of the dismal baseline: we tend to suppose we're not doing all that well mostly due to not making the right comparisons, and we're not making the those comparisons precisely because we're doing quite well. An unfiltered social environment would perhaps be imaginable by way of a museum of abandoned inventions.[11] In any case, recall that the argument required only that we be good *enough* at identifying defeaters; that standard—roughly, not going through life in the manner of Wile E. Coyote—is far short of perfection.

But for another thing, keep in mind the precondition that turned up in the center of our argument: that the process has had time to work. When a strikingly new technology, with accompanying agricultural procedures and economic arrangements, is brought on line, why should we presume that people will be good at anticipating where their reasoning about all of that might run off the road? This is an issue to which we'll return.

Under the second heading, confidence in one's ability to anticipate what one might have to take into account, and how one needs to hedge, isn't a reason to dismiss the problem space we're discussing; on the contrary. In the face of both its formal characterization and the illustrations we've seen, the tendency to think that defeasibility management is easy enough not to require explanation itself requires explanation! I expect that much of the confidence is to accounted for by a mix of two effects. One is so-called fundamental attribution error.[12] We do experience high levels of competence in surveying our inferences-in-progress for potential defeaters; if the hypothesis we're developing is on track, much of that is due to the structured environment in which the inferences take place. But we are prone

[10] Galison, 1987, discusses these sorts of decisions from an historical and sociological point of view; Latour, 1988, describes a feature of the scientific enterprise with a family resemblance to the account I'm giving here.

[11] An idea I owe to Eli Pitkovsky.

[12] Ross, 1977; but the phenomenon was noticed long before. John Dewey tells us that "we are all natural Jack Horners. If the plum comes when we put in and pull our thumb we attribute the satisfactory result to personal virtue" (1988, p. 174). An ancilllary explanation might draw on the depressive realism literature: a psychologically healthy individual will tend to overestimate his level of control over outcomes.

to construing the competence as an intrinsic feature of our own personality (along the lines of: well, I keep my eyes open and my wits about me, and what's so hard about *that*?), one that can be taken for granted. The second effect is perhaps a relative of the phenomenon Bastiat marked with the phrase "what is seen and what is not seen" (1995). Defeasibility management is in the first place a matter of noticing what might go wrong with your inferences. What you don't notice, you're not aware of, anyway unless it's forcefully brought to your attention in retrospect. So to the extent that you're not very good at discerning defeaters, you will tend to underestimate the scope and depth of the problem.

6.5

Before asking what next questions the account puts on our agenda, we need to register its limitations. For one thing, it only purports to explain success in defeasibility management when what we are thinking about are institutions, practices, and technologies—those medium-scale components of a society. But we have to be able to reason competently about much else: I remarked that raw nature is beyond the scope of the account, and inasmuch as science, a good deal of it anyway, is the investigation of raw nature, we have yet to make sense of inference in what has become, in the modern era, a triumphant string of intellectual accomplishments.[13] Even now, it is possible to leave civilization for the wilderness, and in the not-all-that-distant past, human settlements were tiny islands out of which people had to venture into an untamed environment.[14] The restricted scope of the account is as it should be, if the lenses through which we are seeing defeasibility management make it out to be an engineering problem: in the world of engineering, problems are solved piecemeal. Just as there is no across-the-board unified theory of transportation—transportation by car is very different from transportation by bicycle, by plane, by train, by reusable launcher and so on—so we can expect different domains and types of inference to be kept on track

[13] You do sometimes encounter the notion that the success of science is itself due to a related selection effect: "research is," as Medawar, 1967, p. 87, put it, "the art of the soluble." That is, if an area of research doesn't generate problems that we're able to solve, it never begins to count—or stops counting as—science. (Though it's not clear that Medawar, despite originating the phrase, had the conception that gets casually attributed to him; he seems to have been a dyed-in-the-wool Popperian.)

[14] Henrich, 2016, ch. 3, samples the dismal track record of explorers stranded in unfamiliar environments of this kind, some of which is apparently due to shortcomings in defeasibility management.

by very different expedients.

Moreover, even within the social world, we may find institutions or practices or technologies that resist the selection pressures at the heart of the hypothesis I've asked you to entertain: where, for one reason or another, our inferences in the neighborhood of, say, a practice can't be reliably vetted for defeating conditions, and even produce the penalties we've been anticipating—yet the practice remains in play. Premodern medicine persisted through, as far as we know, all of human history; the point for just now is not that the treatments were ineffective, but that neither physicians nor patients were in a position to anticipate a treatment's side-effects.[15] Perhaps sheer desperation kept people going to those doctors and barbers. Nowadays, as some of the examples I've used indicate, networked IT systems seem to cripple our defeasibility management skills; even genuinely disastrous security compromises do not seem to motivate rigorous airgapping: after those disasters, firms seem to go on much as before. Perhaps here perverse incentives are at work, as when shifting costs to users, who are given no choice as to whether to interact with the system, relieves the institution and postpones the day of reckoning past when the decisionmakers—CEOs, university presidents, or even five-star generals—will have moved on.[16]

As those gestures indicate, we shouldn't suppose that what we need to explain with the piecemeal account we are starting off is uniform and across-the-board success. That sort of success is presupposed when defeasibility management is treated as a skill—or rather, as a personal superpower. But there are all too likely to be regions in which success is not the norm, and so it is not an objection if the reach of defeasibility management as we are describing it falls short of what the nonexistent Aristotelian *phronimos* is imagined to be capable.

6.6

The account I've advanced might seem to constitute an argument for an idea floated in passing by Frank Herbert—the science-fiction author of *Dune* fame. Recall that the argument requires that the selection pressures on institutions and other social components have time to work; it may take quite a while for a social building block that does not enable defeasibility management to fall by the wayside. Today that premise of the argument is no longer realistic; we live in a society in which innovations, technological and other-

[15] For a respectable overview, see Wootton, 2007.

[16] The complaint about shifting costs to users is made by Bruce Schneier:

wise, are being rolled out ever more rapidly. If we do not come with built in superpowers, how can we expect to maintain our inferential competences when faced with novelty? Herbert's Bureau of Sabotage, a government service whose agents gum up the works—they're the ones sent out, perhaps, to fix it so that your car is not going to start tomorrow morning—is justified as a way of slowing down progress, and so giving people time to adapt.[17]

However, a policy of slowing down innovation is very likely also unrealistic. And if that is right, our discussion puts an urgent action item on our agenda.

Until this point, we've kept to a level of abstraction at which the feature of institutions and so on that we're considering is captured dispositionally, in terms of counterfactual conditionals about effects on those who interact with the institution, etc.: if such a person were to entertain an inference regarding the institution, or practice, or technology he would be sufficiently able to vet it for defeaters. But secondary qualities are *implemented* (sometimes, as in the case of color, very variously); if a practice or institution or technology is defeasibility enabling, that is in virtue of much more tangible features.

We don't know what those features are, so let me make up an acknowledgedly completely speculative illustration. In the philosophy of explanation, mechanism has been recently claimed to be its central distinguishing feature: we are supposed to understand how things work by showing them to be composed of more or less discrete but interlocking working parts.[18] I am skeptical about the claim as regards explanation across the board, but it's possible that this structure contributes to institutions, practices, and technologies being defeasibility enabling: a user can look for defeaters by considering each component of the mechanism, to see whether it would be doing its job, by considering the interfaces between components, and so on. Structure that supports mechanistic explanation might be, sometimes, a partial implementation of the dispositional property of being defeasibility enabling.

If the argument to this point is on track, we have very good reason to do our absolute best to figure out what can make an institution, practice, and so on defeasibility-enabling. The pace of innovation may not allow time for those social selection processes to do their work, but if we understand how to build in features that make inferences about institutions and so on straightforward to vet, we can adopt policies that mandate those features

[17] Herbert, 1971, p. 14, and Herbert, 1977, pp. 44, 66f,

[18] See, e.g., Craver, 2007, or Bechtel and Richardson, 1993.

in future social—and other—devices.

It's quite likely that there is already a good deal of inheritance of features of this sort, from one generation of institutions and practices to the next: technologies, social and otherwise, are not usually creations *ex nihilo*, and generally fold in working components of earlier technologies. So even a newly-invented institution, practice, or technology can be presumed to already come with some of the implementing features we were considering. (E.g., perhaps they will be modularized in the way presupposed by mechanistic explanation.) But we won't want to rest on our laurels: over and above the pace of innovation, in our ever-more-complex society, a tangible feature that formerly made an institution defeasibility enabling may do so no longer. Evolution and natural selection, here, of institutions and so on rather than biological organisms, will have to be supplemented by artificial selection and by design.

Here we have set before us a research agenda in the philosophy of logic that we see to be hands-on and of great practical import. If we do not take seriously the problem of figuring out how to support our ability to vet our defeasible inferences, we will be lurching toward a future of science-fiction disasters, in which lives will more and more resemble episodes of *Road Runner*. With the right level of focus, and a willingness to take matters in hand, that is a future we can still hope to avoid.

Chapter 7

A Life Hack with a Future

The problem of personal identity is the question—the metaphysical question—of what makes someone the same person over time. It's a fraught question, among other reasons because the answer is supposed to make sense of prudence or self-interest—of the 'special concern' that you're presumed to have for your *own* future, as contrasted with how things will go for other people—and because the answer is tightly tied up with your death: the first moment when there is no longer anyone who is identical to you is the moment when you have died. There are competing answers, some of long standing: just to mention a few of the more familiar candidates, one's personal identity is carried by one's body, or the organism that one is; it is a matter of psychological continuity, that is, that one's future self will remember who one now is as its past, that it will act on intentions one now forms and so on; one's personal identity is put in place by a narrative of a life; last for now, you are who you were because you're possessed by one and the same immaterial spirit.[1] (And of course there are also the dismissive answers: *maybe* you can define personal identity, if you like, but it's not what actually matters.[2]) In the philosophical literature as we have it, the proposed answers compete to match more 'intuitions', that is, judgments prompted by the description of one or another set of circumstances; usually these are to the effect that, in those circumstances, the character as described later on

[1] See, for instance, Locke, 1979, Bk. II, ch. 27, Wiggins, 1980, ch. 6, Olson, 1997, Shoemaker, 1984, Schechtman, 2017. Madell, 2015, promotes as it were a Cheshire-Cat variant of that last, currently unpopular view: just as the Cat vanished, leaving only its smile, in this formulation the Cartesian soul has vanished, leaving only the unanalyzable identity of the person.

[2] Parfit, 1984; see also Eklund, 2004, for discussion of the claims that personal identity is merely conventional and that the concept exhibits indeterminacy.

is or is not the same person as the one described earlier. In this discussion, those circumstances can be exotic, with brain transplants, teleportation, memory recordings overwriting real memories, and even gadgets that morph you halfway into Greta Garbo all standard bits of the repertoire.[3]

Here I want to reframe the problem, and let me be up front about the ingredients that are going to start off the pot. I do agree that this is a metaphysical question, but where the tradition tends to construe metaphysics as in the business of unearthing very special facts, I understand it to be intellectual ergonomics: the objects that metaphysicians centrally take an interest in are generally constructions of our own, introduced as accessories to one or another logic; that is, they are there to facilitate inference and make reasoning possible and effective. So I'm going to be explaining personal identity in those terms.[4]

The appearance of quasi-magical superlative facts that is the hallmark of old-school metaphysics is, if I am seeing it correctly, often the effect of misperceiving a demand as a fact, when the demand is accompanied by frequent enough success in meeting it in one way or another. And I am going to suggest that this confusion has indeed been at the bottom of our engagement with the problem of personal identity. The philosophically important task is to account for the demand; it is a mistake—what philosophers impressed by Gilbert Ryle used to call a *category mistake*—to argue over which of the various possible implementations meant to satisfy a requirement *is* the requirement.

We'll circle back to discussing those intuitions and the appeals made to them; the approach I'm going to take instead is a creature construction argument, and in case you're not yet familiar with the notion, these exhibit the items being treated as a solution to a design problem, and so, as making a kind of functional sense that justifies having them; in this respect, they're close kin to state of nature arguments.[5] Here's a little motivation for the approach. Think of persons as analogous to corporations—as the etymology of "corporation," i.e., embodiment, would suggest. As with persons, there has to be counting as the same firm later on: for example, when the IRS demands social security tax records going back seven years, it can only insist on the documents of that very firm. And there's a technical legal answer

[3] Of course there has been methodological pushback; see, e.g., Williams, 1973c, Wilkes, 1993, ch. 1, Gendler, 2002, and Gendler, 1998. I'll take up Johnston, 1987, which develops a complaint about this sort of conceptual analysis, below.

[4] For this reconception of metaphysics, see Millgram, 2009a, ch. 12; I'll return to it in our Conclusion.

[5] You'll find the technique exhibited in Grice, 1975, Bratman, 2007, *passim* and esp. 49f.

(which differs from country to country) as to how you identify and reidentify a corporation. (In the US, I'm told, part of it has to do with who's on the board of directors and what the mailing address is.) *That's* pretty obviously not something philosophers need to care about; what we want to understand is why we have them: what corporations *do* for us.

The treatment I am about to offer of personal identity *over* time will use as its launch pad an account of, as one might put it, personal identity *at* a time. So let's start with that.

7.1

Imagine that you're a mad scientist who is trying to design a society of androids, and that you're using the account given by Bernard Williams in his final book, rather as though it were a manual. As Williams thought of it, we humans jointly maintain a large shared pool of information, which is usable if the information is true. That will be the case provided only true information is added to the pool. Consequently, there are two relevant virtues we have to insist on, namely accuracy and sincerity, and if we can just take care of both of those, we'll be fine.[6] Your androids, like busy worker bees, are going to gather information for their common pool, and because you've motivated them to be careful and forthright, you expect them to be as dramatically successful in their small-scale ecology as humans have been in theirs.

As it turns out, you find yourself unpleasantly surprised. Caution, sincerity, and truthfulness generally do not prevent the androids' informational resource from being polluted, and even poisoned, by small and larger errors.[7] You wonder whether you could set your creatures to identifying these and rooting them out; at least, when one piece of information is incompatible with another, that would indicate a quality problem to be diagnosed and resolved. Perhaps they could crosscheck the data for incompatibilities.

Once again you are unpleasantly surprised. Just checking for incompatibilities—even for truth-functional consistency—is not computationally tractable. (It's not that you don't know how to instruct them to do it, but that it takes too long; they can't get more than barely started.) And deciding what the right thing to do is, once an incompatibility has been identified,

[6] Williams, 2002; the approach is surprisingly Cartesian—surprisingly, because Williams had thought carefully and critically about Descartes (1978).

[7] See Millgram, 2009a, sec. 5.2, for the point spelled out more slowly, with anecdotal support.

in principle ought to bring to bear any relevant information in the pool; but surveying all the different sets of items pulled from the informational pool, even if they could assess each set, would take longer than your androids have—or rather, than the universe has. Worse, they can't successfully assess each of those sets; like human beings, and for the same computationally deep reasons, they can only wrap their minds around a few items at a time.[8]

Finally you hit on a solution: checking for incompatibilities in the information that the androids have acquired is the right entry point, but to make the task computationally tractable, you have to partition the pool into fairly small fields of representations. To simplify matters, you assign each such small field of representations to an android, you define what will count as an inconsistency in that field, and you task each of them with identifying and then resolving inconsistencies in their respective fields.[9] And once you have done this, you realize that your solution tracks our human mode of proceeding fairly closely: after all, if I believe that p, and I believe that $\sim p$, I am inconsistent, I am contradicting myself, and I ought to change my mind; but if I believe that p, and *you* believe that $\sim p$, we disagree, but there

[8]That is, first, satisfiability is NP-complete (for a textbook overview, see Garey and Johnson, 1979, sec. 2.6); on that last, see as well Miller, 1956.

For that point in the middle, I'm following the insightful argument in Cherniak, 1986, esp. pp. 49–52, 57, 61–70, 127–129 (and which we rehearsed in ch. 4.2): to survey all the pairs of representations in the pool, so as to check them for relevance to each other; next, all triples; next all the quadruples. . . would, all in all, for n representations, take $O(2^n)$ operations—that is, it is also not computationally tractable.

But now, do you really need to consider every element of that power set? Why won't checking a quadruple for consistency also be checking the triples it contains? Almost always, inferential connections are defeasible—a concept we'll reintroduce in sec. 7.3. For the moment, that's to say that even when the inference goes through, a further piece of information (or evaluation, or goal, or whatever) could turn up that would make you take the conclusion back. So a triple might be inconsistent, per an appropriately defeasible notion of consistency you're deploying, even when a quadruple containing it *is* consistent, if the additional element is a defeater for one of the inferences and defuses that inconsistency. For defeasible connections between items in the pool, each element of the power set of the set of items in the pool needs to be checked separately.

[9]Some of these are likely to be the flavor of consistency we concentrate on in deductive logic textbooks, but not all such representational hygiene requirements will have that sort of home. As we have emphasized, not only is the requirement that inconsistencies be extirpated defeasible, the consistency requirements are themselves normally defeasible.

Sometimes we identify emotions as inconsistent, as when you love someone, but also hate him. But we also are often unable to relinquish an emotion, as a way of smoothing out such inner conflicts. In cases like these, we sometimes impose a sort of backup requirement: that the inconsistency be *explained*, in this kind of case, by describing how each response is accounted for, in terms of, say, some aspect of one's own personality, and some aspect of the emotions' object. (For the suggestion, I'm grateful to Buket Korkut-Raptis.)

is, so far, no inconsistency, and no one is—so far—required to change their mind.

At this point, you notice that you have equipped your androids with the rudiments of inference: they have a *logic*. Just like people, if an android has p in its field of representations, if it also has $p \supset q$ in that field, but the field doesn't contain q, or for that matter contains $\sim q$, the android is inconsistent, and is required to update those contents: by adding q, and overwriting $\sim q$ if necessary, or by dropping at least one of the representations p or $p \supset q$. (Here I'm simplifying a little, by counting the gap as an inconsistency; but notice that it tracks our own practice this way: if someone balks at drawing a conclusion he's committed to, after it's been pointed out, not denying the claim but remaining silent—or for that matter, if he seems determined to remain wilfully ignorant—we'll often accuse him of inconsistency.[10]) You have put your androids in a position to *figure things out*, and it dawns on you that *this* is the process that does the bulk of the work of maintaining the quality of the information on which the androids act; Williams was naive to think that the sort of good intentions he had appealed to could manage it.[11]

Let's introduce some terminology. A *self* will be such a scope of application for such consistency requirements at a time, that is, in something like a specious present: it's the synchronic notion. Beyond those limits or such a scope, a representation is assigned to someone *else*, or sometimes—when it is, say, a book or a memo—treated as nobody's in particular; you can think of insistence on the so-called intentional fallacy, much discussed in literary studies, as reflecting a policy of disembodying certain texts in this manner. (Just to be clear, not every scope for a consistency regime is a self; for instance, scientific disciplines impose the rather differently managed expectation that researchers will respond to in-field objections to their results. This is a member of a *family* of techniques.[12]) Of course in real life things

[10] I'm grateful to Jordi Fernandez for reminding me both of our attitudes to wilful ignorance and self-deception.

[11] In case you're wondering how these consistency zones are to be demarcated, here are sample considerations that might reasonably figure into the gerrymandering exercise.

On the one hand, consistency management generally only works well—it only assists you in figuring things out—when inconsistencies are relatively sparse. So the demarcation of selves should take account of, so to speak rifts, in the consistency topography: different cultures, for instance.

On the other, setting up many consistency zones with a high level of content overlap, and letting them run in parallel, can serve as error correction (Millgram, 2025).

[12] In the sciences, we normally impose reproducibility as a precondition for application of the consistency regime, which is at any rate unusual when it comes to selves. Or

are more complicated; like state of nature arguments, creature construction arguments involve ruthless simplification. In particular, almost everyone manages context-specific consistency regimes, which they switch between as appropriate. As Proust remarks, "our person, in the appearance given to it by our body... is made up of several persons superimposed."[13] For the purposes of our creature construction argument, I'll leave the extra bells and whistles to one side.

Person will be the diachronic notion, and here's a bit of linguistic softening up for the choice: we're used to talking about past and future selves, but not about your past and future persons; we're used to worrying about personal identity over time, but not about self-identity over time. So our creature construction problem is: how do we get from the selves to the persons?

7.2

There's a straightforward reason for selves, as we are conceiving of them, to be temporally extended, and let's put it on the table before proceeding to more illuminating issues.

Although I've been inviting you to picture a self at a moment, facing the consistency regimes being imposed on it, such a self can't comply with them if it lasts *just* for a moment. The demand being imposed on a field of representations within a self—that is, the field covered by the scope of application of that demand—is for inconsistencies to be identified, and the representations within the field adjusted so as to conform to its inferential hygiene regime. That's a *process*, which is to say, it takes time, and if we simplify, and consider just the inferences from already accepted premises to not-yet-accepted conclusions, we're reminded that arguments are idealizations of such processes. There's the field of representations prior to the adjustment, there is the field of representations afterwards, and the demand treats them as *the same field*, which is to say that the operations that synchronically extended selves are there to facilitate presuppose temporally

again, we complain about inconsistent legal and regulatory guidance, and, long ago, Han Feizi objected to inconsistencies in rulers' assessments of their subjects (Ivanhoe and Van Norden, 2005, p. 343). I'm grateful to Mathias Frisch for pressing me on this point.

[13]Proust, 1988, p. 40. Occasionally these are quite local: when I amusingly argue for an outrageous position that everyone knows I don't hold, I'm held to consistency within the conversation, but not with my other opinions. (Thanks to Thi Nguyen for reminding me of this type of case.) Moreover, we have (and need) many ways of softening the demands of a consistency regime, some of which are discussed in Millgram, 2009a.

extended selves.[14]

Still, so far you might think we're committed only to very short-term selves. Our toy argument schema, from p and $p \supset q$ to q, can be executed in a snap. Perhaps we don't need more than twenty-second-long persons, more or less in the manner promoted by Galen Strawson?[15] Why not allow the scopes of consistency requirements to shift, once an argument is completed? How is it that we take ourselves to have one and the same person, birth to death, rather than a welter of overlapping person-fragments?

A brief and insufficiently noticed discussion by Bernard Williams addressed itself to Parfit-style psychological continuity accounts of personal identity, and it can serve us as a model.[16] Psychological connectedness and continuity (the properties and relations that underlie personal identity on Derek Parfit's "Complex View") come in degrees. Williams adopted the term *scalar* to describe such accounts, and suggested that if a view of persons on which personal identity is scalar is accepted, "this fact should receive recognition in moral thought." And he considered typical cases of moral obligation, such as promises, canvassing ways one might try to reflect the scalar nature of the underlying properties and relations. Could the action one takes be adjusted along a scalar dimension? (E.g., might I have to pay only part of the money I promised, proportional to the degree to which the person in front of me is psychologically continuous with the person to whom I made the promise?) Could the stringency of the obligation covary with the degree of psychological connectedness? (E.g., might I have to pay all of the money, but less urgently?) Could I be unsure whether the obligation applies, where my uncertainty is proportional to the degree of psychological discontinuity? (E.g., might I understand that if I pay the money, I have to pay all of it, but be unsure whether I do—or precisely to whom?)

The silliness of the suggestions is unavoidable, and it emerges from Williams's discussion not only that we don't see satisfactory ways to reflect the scalar nature of psychological continuity within anything like our current morality, but that scalar versions of personal identity aren't socially *work-*

[14] See MacIntyre, 1990, ch. 2, for very insightful related discussion, engaging the 'genealogical' critique of the presumption that is built into any argument: that there is stability of commitments for its duration.

[15] Strawson, 2009—but perhaps not. As we'll consider in a while (i.e., hold that thought, and for much longer than twenty seconds), a great many arguments take a good deal longer; we can ask whether that explains—given that its author purports to be describing his own inner life—why the book is so remarkably underargued.

[16] Williams, 1981, pp. 5–8. A quick reminder: Parfit allows personal identity to be definable, but thinks that what's actually important is his 'Relation R,' that is, psychological continuity.

able. There have to be straightforward answers to such questions as how much money I owe someone, and so personal identity has to be a yes-or-no matter, rather than a matter of degree.

Occam's Razor—do not multiply entities beyond necessity—is typically heard as an injunction regarding theory choice, that is, as though it meant: if it's simple, it's (more likely) true. I don't know about that, but one thing is for sure, and maybe this is Occam's Disposable Safety Razor: if it's simple, it's *simple*. Imagine redrawing the scope of a consistency requirement after each adjustment to a field of representations. Imagine what it would take—on the part of the iteratively reconfigured inferential agent trying to meet those demands, and on the parts of third-party monitors—to keep track of the constantly shifting selves. Given a particular inferential hygiene regime, a field of representations no doubt contains many inconsistencies; if several of them are in the course of being addressed, but the repairs have different times-to-completion, *when* should we fold down the short-term self? Redrawing the boundaries of the self may introduce new inconsistencies, or legislate old ones away; the collection of problems to solve could change startlingly from moment to moment, erasing whatever progress has already been made on resolving them.

The straightforward design choice—not that we've proven, despite the insuperable appearance of the task, that it *couldn't* be done otherwise—is to stabilize the scopes of those consistency requirements: to demarcate fields of representations that get repeatedly updated, but which persist until some clean—even abrupt—stopping point.[17] And there's a natural farthest-out such point. Ought implies can, in that there's normally not a lot to be gained by imposing a demand that someone can't live up to. We observe that when people drop dead—and since we don't want to put a question-begging construction on the thought, I mean that in the most literal collapse-on-the-floor, death-rattle sense—they stop being able to execute inconsistency-eliminating adjustments to the fields of representations for which they have been assigned responsibility. In other words, given what we're thinking of selves as introduced to do, the default way of managing their temporal

[17]Relatedly, you might have been pondering whether selves, as I am describing them, are going to turn out to shift from moment to moment, with every change in context. Now, sometimes—as in frivolous conversations—consistency regimes are inexpensive throw-aways. But when selves matter, the expense of maintaining, coordinating and administering them tilts the balance of the pragmatic considerations in favor of keeping around relatively few, mostly very stable configurations, typically tied to institutional roles or the like. The drivers of these choices are closely related to those in the main text: we need straightforward and workable design choices.

dimension is to count it as all-or-nothing persistence, sustained up until their deaths.[18]

Those moment-to-moment demands for consistency induce a somewhat different diachronic requirement: not that one's opinions and so on stay the same over time (on the contrary, you're being asked to adjust them to remove one inconsistency after another), but that you demonstrate ongoing good-faith efforts to stamp out one consistency regime violation after another.[19] If you assert that p one day, and you assert that $\sim p$ the next,

[18]But of course sometimes the relevant capabilities give out before death, and then, not coincidentally, we sometimes hear relatives say things like: "she's not there anymore." While after you die, your executor may be responsible for managing various consistency requirements imposed on your estate.

And why shouldn't we count that latter as your continued existence? The requirements your executor is trying to satisfy are typically quite restricted, having to do primarily with legal obligations and what you wrote into your will; they don't include, e.g., adjusting your opinions. And the executor understands his job to be: winding it all down.

Apropos, Johnston, 1987, p. 82, whom we'll engage in due course, warns us not to mistake a persona for a person; you can, he reminds us, outlive a persona; a persona can outlive you, but your mind cannot: "a person cannot be outlived by... his own mind." Regarding the status of that announcement, Johnston tells us "that, if anything deserves the name of a conceptual truth about the relation between persons and minds, it is [that] claim." The appeal to conceptual truth is coming from someone who purports not to be "engaging in the self-defeating search for truths about ourselves that are both purely conceptual and yet specific enough to be interesting" (75), but I'm nonetheless willing to accommodate this one, by explicating the notion of a mind so that it comes out true, as in ch. 3, above.

One has the sense, perusing Johnston's work (but I'm not claiming to read his mind) that what especially matters to him is phenomenal experience—a present sensorium, and memories of past sensory experience—and that to someone like this fear of death is really fear of losing consciousness once and for all. Now, different things matter to different people; for my own part, a dementia that left sensory experience and even experiential memory intact would count as even more devastating, and Bräuer, 2023, makes a case that having their own taste matters to many people as a central feature of their own personality. More generally, as products of a long evolutionary history, we're kludgy: we have indefinitely many sides to us, and when we produce an account of persons or selves, we're opting to highlight some rather than others of those aspects. That choice should not be made on the basis of the vividness, to someone or other, of some of those aspects; instead, we will want to look to the reasons for selecting the objects of our philosophical theory.

[19]To add a bit of emphasis to that point, Klein and Nichols, 2012, p. 695, recap the state of play like so: "How can I be the same person who existed forty years ago, given all the apparent differences...? Hume's reaction to the problem was to maintain that... [t]he fact that a person's psychology is constantly changing entails that that there can be no strict identity..."; on the other hand, "Reid and Butler... maintained that... there must be something that really is unchanging." That is, the shared assumption is that the theory of personal identity is there to capture what doesn't change. On the present view, that's

you're liable to get asked why you changed your mind. When you are, you're being given an opportunity to show that you've been living up to your obligations as a self: to say that you adjusted what you think, or even switched to a quite new take on things, as a way of eliminating an inferential hygiene violation. To be sure, there are many other possible reasons for a change of mind, or of heart; maintaining an inferential consistency regime is only one of your epistemic responsibilities. (You're also supposed to be responsive to what you have seen or heard—just for instance.) And over long periods of time, there will be a great many inconsistencies resolved, sometimes in cases where you could have settled on any of various fixes. So those queries tend to focus on short-term changes of view: If you assert that p, and are reminded that as a five-year-old you used to think the contrary, nobody will ask you to explain.[20]

7.3

That was the more straightforward, but also less interesting lap; I'd have been remiss to leave it out, but let's proceed to a more illuminating reason for extending selves into persons. And first, let's add a further concept to the menu: defeasibility in inference or reasoning.

We (I mean, we analytic philosophers) are normed into deductive logic, in which (i) the form of the inference explains why it's okay to draw the conclusion; and (ii) what that form provides is an ironclad guarantee that, if the inference's premises are true, the conclusion is true also. That is, nothing else you can subsequently learn (as long as it doesn't impugn those premises) will require you to take the conclusion back, and here's a fancier way to say it: when inference is deductive, adding more premises only ever increases the number of conclusions in the pile, and so the size of the pile of conclusions is a monotonically increasing function of the size of the pile of premises.

---

got things upside down: personal identity is there precisely to make room *for* psychological change, which can be quite dramatic when it serves the purpose of figuring things out.

[20] Does this mean that personhood turns out to be a matter of degree, after all? (I'm grateful to Jonas Harney for pressing this question.) Perhaps it would, if we were allowing consistency requirements to fade over time. But with a couple of categories of exception I'll discuss below, we impose consistency demands synchronically; so if you are now inconsistent, we may demand revision; if you are no longer inconsistent, we won't. Your different views at different times are not an inconsistency, but rather, as I've been framing them, an *occasion*, and if too much time has passed for an explanation of the change to be interesting or even possible, then it's no longer that occasion.

In real life, most reasoning isn't like this. The normal, *defeasible* case is that you can have an inference that's just fine, as far as it goes; the premises are fine, and you should draw the conclusion... but you could come to learn something else that would make you take the conclusion back. You can add premises, and have the pile of conclusions shrink: so inferences like these get called, in some fields, "nonmonotonic"; the extra premises are "defeaters," whence the term "defeasibility".

For example, Iris Murdoch was a deeply insightful, exquisitely refined philosopher. It's a fairly robust induction that when someone like that writes a philosophical work, you'll find it to contain muscularly sustained trains of thought, and to be something definitely worth reading carefully. So if you come across a book by Murdoch that you haven't yet read, you should take it to contain muscularly sustained trains of thought, and to be something definitely worth reading carefully. Fine, but remember that Murdoch was an author of philosophical novels, and maybe the induction is defeated by that further heads up. And also, *Metaphysics as a Guide to Morals* (and to a lesser extent *The Fire and the Sun*) were from the period when Murdoch was starting her descent into dementia. (Famously, she died of Alzheimer's.[21]) When you learn that, you aren't so sure of the conclusion, and maybe you just drop it. Notice, while we're at it, that the illustration fudges the distinction between theoretical and practical reasoning, as a way of making the point that the concept of defeasibility generalizes, past the cases where the status of steps in one's reasoning that you're tracking is truth. In practical reasoning (reasoning about what to do), it's procrustean to insist on that, but you can still find yourself needing to take back conclusions—about what's worth reading, per my fudge—in just the same way.

We don't know how to spell out the weaker not-quite-guarantees that inferences like these come with. If you start to list potential defeaters for some inference you're considering (if you start to fill in the *ceteris paribus* clause, by saying what the 'other things' that have to be 'equal' are), you'll find that the list never ends: no matter how many such circumstances you've noticed, you can always think of another one, and what's more, you'll find that there's no emerging order in those lists—the defeaters don't converge on a handful of subject matters, so to speak, that would allow you to handle them in bulk. Rather, they end up being as different from each other as you like.

Think for a moment about what this means for administering selves, as we introduced them. Selves are scopes for inferential consistency regimes.

[21] For her husband's account of her illness, see Bayley, 1999.

What does this sort of administration involve?[22]

When the inference rules in question are deductive, that's fairly straightforward to manage in real time: in the caricature I used earlier, if I say that p, and I also say that $p \supset q$, and I refuse to conclude that q (or to drop a premise), you get to sanction me, by calling me "inconsistent" and "irrational"—and maybe, per the early stretches of Robert Brandom's *Making It Explicit* (1994), the local convention might allow for beating me with sticks. But what if the inference in question is advanced as defeasible? If I go ahead and draw the conclusion, maybe that's wrong, because the inference is defeated. And if I don't go ahead, maybe *that's* wrong, because the potential defeaters aren't live, after all. And it looks like we can't do a conclusive check, for two reasons.

Again, because there are indefinitely many potential defeaters, you can't page down the checklist, to ascertain whether one of them is not just a potential, but a live issue: the checklist doesn't come to an end. And second, because the potential defeaters really can be as different in content as you like, generally there isn't a mechanical way to determine, given a potential defeater, whether it is actually a defeater. Normally the way this bit of it is managed, it seems to me, is by trading explanations back and forth: you explain why the defeater is pressing (in our illustration, when you're in the early stages of Alzheimer's, you gradually stop being able to manage lengthy trains of thought).[23] The bottom line: where monitoring a self for

[22]One thing we can expect it to involve, as I'll suggest in a moment, is policing. Now, policing, whether it's done to you by others, or done for you by yourself, takes time: to notice the errors and deviations, to get called out on them, to insist on correction and so on. So for the requirements constitutively imposed on selves to come out real, rather than empty pronouncements, selves have to be extended into persons.

That quite abbreviated train of thought may be identified as Wittgensteinian in spirit, so let me be up front about the way it's departing from the usual way of proceeding. Wittgensteinians insist that policing has to be a collective enterprise, one which requires that what is policed be done in public. (There is no such thing as a 'private language'.) While that seems to be for the most part about right, there are important exceptions: the deeply innovative philosophers, among whom I count Wittgenstein himself, invent conceptual vocabularies and modes of inference which, at the outset, only they understand. Normally, it takes a few generations for the new ways of thinking to be mastered by a larger community; so policing must sometimes be managed by the innovator himself. That is to say, Wittgenstein can be adduced as a refutation by counterexample of his Private Language Argument.

[23]The strategy of trading explanations and objections is explored by Mercier and Sperber, 2019, and Dutilh Novaes, 2021. The latter sees the practice as the predecessor of our deductive practices.

When you do this, you're prepared for explanation-based objections to a proposed defeater. For instance, Murdoch compresses her deepest insights into snippets of prose,

compliance with deductive inference requirements is cut-and-dried, checking compliance with defeasible consistency requirements is a much more dicey matter.

We're thinking of inferences as propelled by consistency requirements: if you accept the premises, but not the conclusion (anyway when it's pressed on you), you're inconsistent. Here we're contemplating a softer, trickier such constraint: other-things-equal consistency requirements, where if you accept the premises, but not the conclusion, and other things are equal (that is, the requirement isn't defeated), you're inconsistent. (And there's a mistake that's the flip side of that violation, namely, drawing the conclusion, even though other things *weren't* equal, that is, even though the consistency requirement *was* defeated.)

Slowing down to spell out the train of thought a little more, here's the problem posed by defeasibility, when you're thinking of selves as scopes for inferential consistency regimes. Administering a consistency regime requires the ability to flag and correct inconsistencies. So selves presuppose and require the ability to check inferences for correctness. Now, when the inferences are deductive, that's generally straightforward: deductive inference is formal, and so the form of the inference is enough to support real-time monitoring. However, for human beings, most inference has to be defeasible. And when an inference is defeasible, its conclusion only follows 'other things equal': that is, when none of its potential defeaters are *actual* defeaters. So to check whether an inference (or the failure to proceed with one) violates the underlying consistency constraint requires checking whether other things *are* equal: whether any of its potential defeaters are actual defeaters. But we observed that a defeasible inference has indefinitely many potential defeaters. And when it's indefinitely many items, a checking procedure doesn't get a grip: you can't check whether the underlying consistency constraint is being violated. Although when you identify a live defeater, you can say it's not being violated, you're never in a position to say that there are no potential defeaters that are actual defeaters. Which gives us an apparently worrisome conclusion, namely, that you can't administer selves that figure things out via (mostly) defeasible inferences. If selves are administrative devices, that's as much as to say, sorry, you can't have selves.

But if you can't flag and correct inconsistencies when a putative inference is about to be executed, and you also don't want to just abandon this very clever hack, the fix looks to be: making room for the correction

---

and those insights were stored up over a lifetime of philosophical thought; so even if the organization of the book will be spotty, those rewards could still be there.

to happen *later*. Often enough, it becomes clear in due course whether the defeasibility conditions of an inference were correctly assessed: the possibility you hadn't noticed, or had dismissed as unimportant, in retrospect is what should have made you call it; conversely, the reason you did call it turns out to be something you could have ignored. And when you realize that—or have it pointed out to you—that can be an occasion not just for recalibrating the manner in which you conduct defeasibility management, but for rethinking the issues connected to that other-things-equal consistency constraint that you belatedly realized you had violated. If we're going to call temporally extended selves "persons," we can put the idea this way: *one* reason persons are there is to be the space in which the practices of correction needed to support defeasible consistency regimes have time to do their work.[24]

7.4

Let's pause to consider a couple rounds of pending objections, starting with matters of methodology. The ongoing discussion of personal identity has been directed toward saying *what it is*, where that means specifying in just which circumstances someone at a later time is the same person as someone at an earlier time. And that specification has been taken to be responsible to our judgments as to when that's the case. Whereas I've put all that aside, in favor of explaining the point—or rather, as we'll see, *a* point—of having diachronically extended selves. Why haven't I just changed the subject, and anyhow, what justifies my willingness simply to disregard our intuitions?

As alert readers are likely to have noticed, the frame we've adopted to

[24]How much space does that take? That is, how long do we want persons to last? Since we're considering the design of persons as an engineering problem, the requirement is not that we *always* have enough time to sort things out and learn our lessons, but that we manage to do what we need to do acceptably often.

Updating our first pass over the question, the point of correction is to allow a self to fix its mistakes, because in the course of doing so, it will try to figure things out—to investigate how best to repair an inconsistency, for instance—and those results are payoffs we're after. Fixing mistakes in a field of representations means adjusting *that* field of representations. So temporally deferred correction presupposes a self that remains a going concern until the correction can arrive. Corrections to defeasible inferences typically look like this: only later—maybe quite a bit later—does it become apparent that overlooking some factor was an error. So corrections often wait on your being brought up short by a defeater proving actual—and who knows when *that* happens? That's to say that corrections to defeasible inferences can take indefinitely long to turn up. Since we don't know how to build persons that last forever, we could opt instead for persons being indefinitely temporally extended: they should stick around for a while (but not any precisely specifiable while).

explore our topic treats humans as, first and foremost, boundedly rational: maybe there's an in-principle-correct way to draw conclusions (and maybe not), but creatures like us have to make do with limited cognitive and computational resources.[25] So we'd expect those intuitions to be the outputs of heuristics: suitably inexpensive processing that produces correct-enough answers in the typical cases, at the cost of giving poor answers in the sorts of cases that matter less, especially because they're unlikely to come up. But now, as we remarked early on, the arguments about personal identity that drive the literature are centered on intuitions about exotic cases, that is, precisely the intuitions that are least likely to be worth much. For instance, one class of such cases is sometimes called "branching": someone waking up one morning as two different people on the two sides of the bed. If people don't actually branch, why *wouldn't* a decent heuristic sacrifice correctness, when assessing whether either of such a person, so to speak, is the same person as before?[26]

We briefly analogized people to corporations, so consider how heuristics get used in sorting out corporate identity over time. Whether something is the same company for legal and accounting purposes, and how it's reidentified in practice can be very different; generally, people's intuitions here have to do with branding and the logo, on a sense of what lines of business the firm is in, and the like. Suppose we try to elicit people's snap judgments, as to whether what they see later is the same corporation as earlier on. Library Bureau, an early American library supplies business, was acquired by Remington Rand, which manufactured office machinery, computers, but

[25]Quickly explaining that 'maybe not': Old School work in bounded rationality makes sense of what we do as approximating ideal rationality, given what's feasible for us. But the New Wave approach to bounded rationality acknowledges that often enough what's ideally rational is not well understood, or even not well-defined, and we still have to come up with workable solutions.

How can that latter approach make sense? Isn't that a bit like offering a shortcut to someone who doesn't know where he's going? (For this very vivid way of putting the complaint, I'm grateful to Sergio Tenenbaum.) Now, yes, it's very like that; a while ago I described a back road that would take a friend to southern Utah, although she didn't yet have maps of the different national and state parks that would let her decide just where she wanted to go. We don't have a satisfactory theoretical grip on bounded rationality that stays bounded all the way up, and that's a major part of the New Wave agenda. But it's clear that we have to get that grip; there's already a smallish literature pointing out the need to set satisficing thresholds for choice sets with no maximal elements (see, e.g., Landesman, 1995).

[26]Though as Rovane, 1990, points out, if branching were normal, the ways we think and talk would have to readjust themselves around that fact. We can add to that: what our heuristics prioritize would have to change.

also occasionally firearms, and which in turn became a division of Sperry and subsequently Unisys—all while retaining its legal obligations. Much later, Remington sold its Library Bureau trademarks to another firm, which did not come to count as continuous with the previous corporation. Here, intuitions will generally be an ineffective guide, and why expect intuitions about personal identity to serve us any better?

Could we perhaps do better by shying away from the exotica, and sticking with what everybody just knows about how to keep track of who is who in ordinary life? I'm not optimistic, and to show why, I'll draw on a widely read essay by a philosopher who asks us to rely on practices we appeal to on a day-to-day basis. (His thought is that they're connected to their subject matter by counterfactuals: the practices *would* normally, ordinarily work; so persons have to be what would ordinarily, normally be successfully traced and reidentified by our practices.) And what practices are those? Evidently, we're told, "our offhand methods of tracing by way of carefree observation of bodily continuity, sometimes augmented by observing some sort of consistency in behavior."[27]

This recommendation does stick with what everybody just knows, but now for the much-needed reality check: I'm quite sure that nobody follows me around to verify that my bodily continuity is uninterrupted, not ever, and only my very closest friends are positioned to notice—or so much as pay attention to—consistency in my behavior. (Though as I've been arguing, we do track consistency of other sorts.) How do we *actually* reidentify people?

[27] Johnston, 1987; on the way, he complains about traditional conceptual analysis, which he calls "the method of cases," so let's quickly mark how that objection differs from ours. In the case of personal identity, he has it, in our linguistic and cultural community people have been trained on various and different conceptions of personhood—usually, we're given to understand, those associated with competing religious and secular ideologies. Since training on a conception shapes one's intuitions, our linguistic and cultural community will contain a sort of spread of intuitions on the topic of personal identity, rather than exhibiting tight convergence. Averaging over them, we end up with a vague and unspecific gesture towards a poorly demarcated conceptual *zone*, rather than tight necessary and sufficient conditions. So the 'method of cases' is unsuitable for the subject matter—though it might suit other topics in philosophy, possibly the analysis of knowledge. —As we'll see shortly, an inability to supply necessary and sufficient conditions is, by the lights of the present account, no cause for complaint.

In a sequel (2007, p. 41), Johnston prods us to move on to "the method of real definition," which we are told means "using all the relevant knowledge and argumentative ingenuity we can muster in order to say *what it is to be* the given item or phenomenon." Although the piece also seems to advance the earlier proposal, this pronouncement abandons the method of seeing what our methods of reidentification track: "Do *whatever*" isn't a method, and so there *is* no 'method of real definition.' For this reason, I'll confine myself here to the contentful earlier proposal.

Sometimes by using distinctive objects presumed to be in one's possession and under one's control: you'll be asked to show your ID; nowadays two-factor authentication may deploy your mobile phone or an RSA security token; in the Biblical story of Judah and Tamar, he leaves her, as a sort of security deposit, his "signet, cord and staff".[28] Sometimes we demand information presumed to be secret, such as a password, or, in the current United States medical system (and remarkably naively), the patient's date of birth. And then there's recognizing someone's voice (deprecated, due to recent developments in voice synthesis), face recognition (both automated and not), and in special cases, DNA sequencing.[29]

[28] *Gen.* 38:18; for context, see Millgram, 2008, pp. 76–89. But you might be entertaining this objection: when I reidentify my friend by recognizing her face, or when the police officer matches my face to my driver's license, isn't that just a proxy for and means of verifying bodily continuity?

In reidentifying the person by his face, we can let the detour through uninterrupted bodily continuity drop out, and a proponent of intuition-based philosophizing will allow me some science fiction to make the point. When my friend enthuses that for the past several months she's immersed herself in an extra-mystical brand of yoga, one that grants you the deepest possible relaxation by teaching you to dematerialize your body—and that she sometimes manages to do so twice a week—I'm perhaps amazed at her new skill, but I don't take myself not to have recognized her when she stepped into the coffeeshop. My recognition evidently wasn't advertising itself as insurance against yogic dematerializations.

When animals analogously reidentify their conspecifics, or humans, by their scents or distinguishing cries, the question of uninterrupted bodily continuity analogously doesn't come up. Whyever expect us to have made this our real target?

[29] And how well do those allegedly connecting counterfactuals hold up? There are forged ID cards and passports; sometimes people can't tell by looking at the picture if it's the same person; face recognition and other biometric software fails frequently; IDs get stolen or borrowed on the way into the liquor store. Don't even get me started on what's almost certainly happened to your passwords; nowadays, identity theft is a large-scale problem.

The older casual methods of reidentification didn't work particularly well, either; there's a well-known sixteenth-century illustration (Davis, 1983). When, after a prolonged absence, someone purporting to be Martin Guerre appeared in the neighborhood in which he had grown up and lived, people who had known him well—perhaps including his wife—were unable to say for certain whether or not he was the person who had left.

In short, we use our methods because in ordinary circumstances, they work *somewhat*, and we can't be bothered to do better. More broadly, i.e., when it's not just for people, use of ineffective methods of reidentification is not an outlier. For most of human history, diseases were tracked on the basis of their symptoms; now that we have a good deal more medical knowledge, we are aware that infectious diseases will often be asymptomatic, that symptoms vary over the course of an illness and from person to person, and that different diseases share symptoms. Indeed, applying Johnston's method gave us hodgepodge concepts such as "cold". (There is no such disease as a cold; rather, there are rhinoviruses, coronaviruses...) You wouldn't have found out what diseases were by asking what premodern methods of reidentifying them were reliably tracking. We generally track a great

What philosophers are inclined to say about what we do—how we reidentify people, but we find this to be the case more generally—is in my experience peculiarly disconnected from what we in fact do. The disconnect unnervingly resembles aspects of late medieval culture, as rendered in Huizinga's classic *Autumn of the Middle Ages*.[30] In his telling, the realities of (just for instance) the power politics of the day were apprehended via the conceptual and evaluative repertoire of chivalry, that is, irrelevant and mythology-soaked cultural forms that had long since ceased to fit the practices of a differently configured world. It's as though philosophers are unable to allocate attention to their own lives and the activities in which they engage day to day. This sense of ungrounded fantasy, both in rarefied philosophical discussions of personal identity and in the more mundane appeals to methods of reidentification, makes various large questions pressing, but sticking with the one we're considering right now: What could judgments—'intuitions'—that are rooted in this sort of discoupled understanding of our human world be worth?

We're eschewing not only intuitions, but the previous method's objective of providing necessary and sufficient conditions for a person at an earlier time being the very same person as one at a later time. Have we just given up on the philosophical task, that of understanding personal identity?

We needn't think of that objective and that latter task as one and the same, and let me offer an analogy. The academic mission involves both education and research, and so it is very important to have an understanding of what each of these is, and quite as important for practical purposes to be able to tell good research from bad, and successful from unsuccessful pedagogy—in something like the way it is important for practical purposes to be able to reidentify people. Nonetheless, experience tells us that nothing good can come of academic institutions attempting to define either in the sort of way that lays out necessary and sufficient conditions.

There's an underlying reason, which will also turn up when we think about how we pick out persons over time. Research can go just *anywhere*, and it needs to reshape itself to its subject matter. Education needs to reshape itself both to what is being taught (which could be just about any-

---

many items using methods that *don't* reveal what they are; in retrospect, again and again, such techniques have proven not to work. That is, they didn't support the counterfactuals Johnston's method requires—they *wouldn't* work—and briefly, the presumptions that method makes cannot be taken for granted, but have to be supported by argument on a case-by-case basis.

[30] Huizinga, 1996; formerly translated and more widely known as *The Waning of the Middle Ages*.

thing), and to what the instructees bring to the process of instruction. David Wiggins once insightfully observed that human beings need to be endlessly resourceful in their social relationships and arrangements; we've been insisting that we're in the business of figuring things out, and so, taking his point, we should allow humans to reshape themselves, in ways that could sidestep any tight list of necessary and sufficient conditions for being extended over time, and *a fortiori*, anything along the lines of the traditional proposals.

Wiggins advanced that argument as supporting a demand that we see people as organisms rather than social constructs, mechanisms or artifacts, and let's allow that train of thought to play out, as a way of giving our argument some traction.[31] A bit of backstory: Williams had complained about accounts of personal identity that had it carried by one's body, that we lack a philosophical account of what a body *is*, and so, of what makes something the same body over time. So Wiggins took up the explanatory burden: by body, what is meant is an organism of a particular species, and the species form, to borrow an Aristotelian phrase, will determine what it takes to be the same organism over time.[32] That is, species come with, roughly, func-

[31]Wiggins, 1980, pp. 179–182, and what was the train of thought? First, if *person* is an artifactual kind, he reasons, then only the features of persons that register as elements of its design come into play when it functions. Now, persons are deliberative, choosing beings. So only features of persons that we're aware of, as components of the design of our deliberative process of choice, can figure into it. (Today that's presumed to be preferences and desires, thought of as input to selection of means, and that was Wiggins's own focus in the course of his discussion, but the formal point is more general, and so that's how I've couched it.) However, it's absolutely crucial that we be able to think outside the box: to deliberate in a way that's not confined to any antecedently specified range of inputs, or to any antecedently specified formal procedure. So it's absolutely crucial that we *not* be—and that we not remake ourselves as—an artifactual kind. The resource that allows us to respond in ways not captured by a design specification of us-as-artifacts is our biological or creaturely nature. So we have to construe ourselves as biological creatures—thus, with species-based identity criteria.

Moreover, Wiggins holds that one of the deep features of persons is that they can engage with and get along with one another: man is a 'political animal'. Engaging others, in a way that often enough lets you get along, requires this: that when you look *further*, past current understandings and disagreements, you sometimes—often *enough*—find common ground. So persons must be presumed able often enough to find common ground, past and beyond what they've already had in their sights. The responses underwriting that ability are rooted in the biological nature of the organism, and in the case of humans, in ways their nature plays out in what Wittgenstein called their forms of life. So, once again, persons must be understood as biological creatures: as members of a species (the human species) whose nature underwrites all that potential reach of agreement.

I'm on board with much of that, but take exception to the assumption that artifacts and social constructs can't provide an equally deep fund of resources.

[32]For the complaint, see e.g. Williams, 1973a, sec. 2. There was a competing response

tional blueprints, which determine what it is to stay the same individual: a butterfly starts out as a caterpillar, wraps itself in a chrysallis, emerges as a winged creature, and is that same individual throughout, because that's what butterflies do; but if *you* cocooned yourself and turned into a large flying insect, you'd have died and been replaced by something else, because that's *not* what human beings do.

All well and good, but in this day and age, the species form had better be real biology, not Aristotelian speculation, and so we can fill out the suggestion with a position developed by Thomas Pradeu, who has recently argued that we should look to the immune system. The immune system monitors the organism and tracks abrupt changes at the molecular level; immune response is criterial for what constitutes the organism. And because the immune system is always learning what the organism it defends is, it registers continuity through change, and so how the organism's identity develops over its history.[33] Let's take the approach on board for a moment, not that I necessarily endorse it, but for the sake of the argument.

We already tamper with immune response, in order to transplant organs and prompt resistance to infectious disease. It's science fiction, but not outlandish science fiction, to suppose that in due course we'll see immune response being manipulated to preempt the rejection of organs from a particular donor. We're told that husband and wife are of one flesh (*Gen.* 2:24); if the husband donates a kidney to his wife, with this sort of immune system tweaking, would that become literally true? Would they end up the same person for legal purposes? If the Wiggins-Pradeu criterion for personal identity that I just sketched *were* our own, and this prospect materialized, it would be necessary to adjust or replace the criterion, to preempt unreason-

---

to it, taking 'body' to mean, in the first place, a material body: a physical object capable of serving as an address at which to locate thoughts and other elements of a psychology (Strawson, 1971, chs. 1–3). That type of account seemed to press one in the direction of criteria for being the same material body over time. However, in a famous but still unpublished lecture series, Saul Kripke compellingly argued that ordinary material bodies were as basic as it got; they didn't come with—or need—criteria for reidentification.

[33]Pradeu, 2012, a view with some surprising consequences: for example, your microbiome (e.g., intestinal bacteria which are tolerated by your immune system) counts as part of you; cells that are genetically your mother's, but which are retained in your body, and conversely, cells with your genome retained in your mother's body count as, respectively, yours and hers.

But can this is right? I've had it put to me that the immune system is too complicated and messy for there to be a simple way to say what it is, by saying how it works. And if you think about it, there's a plausible reason for that: immune systems are in an evolutionary arms race with attackers; arms races produce kludges; the upshot of an ongoing series of kludges is normally something that's too hacky to describe cleanly.

able outcomes, and briefly, what matters for philosophers is understanding the point of personal identity, and not the fool's errand of trying to spell out necessary and sufficient conditions for it.

7.5

We've been trying to understand the point of personhood as building out synchronic consistency regimes—selves—in ways that enable them to figure out ever more (and ever more difficult) things. But isn't this an overintellectualized approach to understanding what we are? And if we're going to be more practical, shouldn't we rather be focusing on the consistency conditions that we impose on people *over* times—say, to retrieve an instance we've already mentioned, your responsibility for paying back the amount of money you borrowed to the person from whom you borrowed it?

Now, we shouldn't assume that consistency is a rarefied, and so to speak merely theoretical matter. We assign to selves not only the task of identifying and resolving contradictions between their views about how matters of fact stand, but of rectifying and eliminating specifically practical incoherences. Sometimes these are construed as means-end inconsistencies, as when you are not taking effective steps to an avowed end; sometimes these incoherences are made out as incompatible plans; sometimes, more ambitiously, Kantian conceptions of "contradiction in the will" can get invoked. There are consistency concepts developed for patterns of preference as well.[34] But we can also push back on the thought that sorting out theoretical inconsistencies is somehow unworldly. The payoffs of getting these ironed out are in the end as hands-on practical as it gets.

I've already mentioned that creature construction arguments are relatives of state of nature arguments, so recall how the classical arguments of the latter kind highlight one functional dimension of the state; e.g., Hobbes's justification for political states was that they provide security services. What we can call the libertarian fallacy is leaping to the conclusion that anything over and above the conclusion of such an argument is an illegitimate activity on the part of your government. But of course any modern state is multifaceted and multifunctional, with rather the look of a Swiss Army knife: governments provide parks and parades, roads and running water, and in the US, taxpayer-subsidized entertainment in the form of the public univer-

[34]For an overview of Kantian consistency, see O'Neill, 1989; Davidson *et al.*, 1955, famously introduces a consistency requirement over preferences; for the classic survey, see Luce and Raiffa, 1957.

sity's football team. Any real state is much more than an army and a police force. Our creature construction argument likewise tells us what persons are, on the basis of a take on what selves are; but if you'll forgive me the locution, there is much more to you than your self, and you are not merely your personhood.

The aspect of the way we run our lives that I'm focusing on is underdiscussed, and in my view, distinctive of human beings in a way that harks back to a misunderstood philosophical precedent. When Aristotle looked for the human *ergon*—for what it is we do and how we work—he ruled out activities shared with other living beings, and that restriction has come to seem quaintly archaic. (Why *shouldn't* what we centrally do be something that other creatures also do?) But much of the way we live is visibly quite different from the natural histories of the other living beings with whom we share the planet, and seeing how that falls into place around what is, as far as we know, a uniquely human technique does seem to me to be philosophically revelatory. This technique inflects pretty much everything we do.

On the one hand, the reason for imposing an inferential hygiene regime is getting people to figure things out: when you're faced with an inconsistency (in your beliefs, your intentions...) and you have to think about how to change what you claim, or perhaps are pursuing, so as to remove the inconsistency, you *reconsider* things. Doing that intelligently requires a good deal of flexibility: you could adjust *this... or that*.

On the other hand, those diachronic consistency constraints are in the first place imposed precisely where we *don't* want flexibility; they're not a way of prompting and enabling people to figure things out. Rather, they're a way of making people *just* do it. (Why the contrast? In part, because you can't change the past, so consistency adjustments mean bringing the future into line with past decisions and determinations.) In the *Genealogy of Morals*, Nietzsche dramatizes how brutally contracts were enforced long ago; sticking with our illustration, where a loan now requires a repayment later, "if he should fail to repay... the creditor could inflict every kind of indignity and torture upon the body of the debtor."[35] Torture, Nietzsche tells us, generated a memory, "and it was indeed with the aid of this kind of memory that one at last came 'to reason'". Even if that's true, this is not where the more interesting, creative, original, and productive sorts of

[35]In Nietzsche, 2000, at Essay II, §§5, 4. Even if the indignities are now inflicted on one's credit rating, loan sharks are not that far back in the past; the approach is not so distant as to be unfamiliar.

reasoning arise.[36]

7.6

Personal identity over time is widely supposed to serve as a platform for a number of interrelated attitudes. You are supposed to have a "special concern" for your own future self, a concern that is different from the stake you have in how other people are doing. In part that is a matter of being motivated by your own self-interest. The primary mark of that special concern is the way that a benefit to you at one time can compensate you for a cost at another. (Secondary marks include emotions such as fear and regret.)

When a self sets about resolving inconsistencies within its field of representations, it ought not go about it by, say, flipping a coin to decide which of the incompatible representations to delete. After all, the point of tasking selves with the effort of resolving incoherences was to have them figure out one thing after another. That *can* be just drawing a conclusion from premises you believe, but in the approach we're taking, inconsistencies aren't just *bad things to be gotten rid of*: rather, they're cognitive fuel. They trigger investigations, which sometimes are mere surveys of one's assumptions, looking for what went wrong, but more demandingly can involve fieldwork, experiments, and in the most dramatic cases, deep conceptual change. As we academics are aware, getting to the bottom of such a puzzle can take both a good deal of time, and also a great deal in the way of resources. Surely the real justification for tenure is that it allows and encourages a researcher to take as long as he needs—decades, if necessary—to figure something out; there's a lot of variation by field, but research can involve activities that cost more than any one grant will cover.

[36]That said, we can allow that there are also attempts to spell out diachronic consistency requirements put in place in the service of figuring things out. Bayesian updating is recently much discussed, though not implemented in real life; perhaps the most plausible of the philosophers' proposals is Michael Bratman's investigation of plans (2001). A plan or policy, once adopted, frames subsequent deliberation; as you fill it in, the subsequent choices—or subsidiary plans—are required to be means-end consistent with your earlier and now ongoing intention. While Bratman allocates various tasks to plans, one of them is indeed to help you figure things out. If I survey all of my options at any given time, my choice problem is not manageable: for instance, I could this very instant drop what I am doing, get a cab to the airport and the first plane to Brazil, and spend the rest of my life in Rio. (How on earth am I to decide whether it's a better option that what I currently expect to be doing? And there are indefinitely many such swerves to examine, once I start in on it.) Because that course of action is excluded by my plans for the afternoon, for the week, the year, and indeed farther out, I don't consider it... and so deciding what to do becomes a tractable problem for a boundedly rational agent like me.

Not every self is an academic researcher, and that's no doubt a good thing, but the point generalizes: a self that is committed to figuring out how best to resolve the contradictions for which it is held responsible needs both the time and the resources to do it. And so if a mad scientist were designing a fleet of such selves, he would be well-advised to equip them with a secondary mission, that of making sure that they have the time to make headway—that each person takes steps serving its own persistence—and the resources to support investigative activities.

Research is what you're doing when you don't know what you're doing: because investigations can develop in unpredictable ways, the specific resources they require can't be anticipated.[37] Accordingly, the resources that an effective self attempts to accrue had best be generic, that is, able to support a wide range of possible activities. Examples of such generic resources might include financial depth, social positioning, a liberal arts education and a panoply of widely deployable skills.

When an individual maneuvers to build up a supply of such generic resources, he is behaving in a manner analogous to a political state, as it advances its national strategic interests. A state protecting its interests is engaged largely in *positioning*. Compare a simpler but quite familiar model of strategic cognition: in the GOFAI approach to chess—that is, "Good Old-Fashioned AI," as contrasted with contemporary connectionist machine-learning approaches—the program faces a move tree on which it can't do an exhaustive traverse: just like human players, the complexity of the chess problem doesn't allow it to be treated like a game of tic-tac-toe (even though *formally* they're on a par).[38] Instead, a heuristic is used to assign rule-of-thumb values to states of the board, which will normally

[37]The remark is usually attributed to Wernher von Braun. And as Wilkes, 1993, p. 54, remarks, "Longer-term interests are rarely... specific and particular, and the further away in time they are, the less specific they tend to be."

In an earlier note, we highlighted practical inconsistency, and although there are different spins to put on it, we can treat the following as an observation that applies to them across the board: over and above the costs of research, ameliorating such inconsistencies is often a matter of deploying resources. More crudely put, sometimes you can spend your way out of them, and here's an amusingly vivid illustration, courtesy of Madeleine Parkinson. She recounts how, in Istanbul, she took a cab to the wrong airport, and ended up paying the driver enough for an airport-to-airport ride that allowed her to make her flight after all—a feat which seemed to involve dispensing with the rules of the road.

Resolving practical inconsistencies by throwing money at them is a typical enough fix, and generalizing, it makes sense for a self that is committed to ameliorating and rectifying practical inconsistencies to have, as a secondary mission, assuring the availability of resources that will make it frequently enough feasible to do so.

[38]von Neumann and Morgenstern, 1944, pp. 112–125.

roughly assess positioning. Allocating its limited resources, the program examines a small subtree, generally just a few moves deep. Each node of that subtree is a state of the board, and a move is chosen on the basis of heuristic values given to the most distant surveyed nodes (to the fringe of that subtree).

It will be natural to layer onto the design of the self as we have been conceiving it the motivation to pursue its self-interest, where that concept is to be construed on the model of the interests of political states. And here's the last-for-now surprise our creature construction is suggesting. Self-interest won't be a matter of overall preference- or desire-satisfaction, and *a fortiori* not of overall hedonic reward. Philosophers since Aristotle have tried to understand the interest in one's future self as anchored by an assessment of the life as a whole.[39] However, our mad scientist would plausibly opt to augment the cognitive equipment of his temporally-extended androids rather differently. Just as a political state does not generally make strategic choices on the basis of a conception of its well-being over the course of its entire history, but rather focuses on how well resourced it will be to deal with future challenges and contingencies, at some horizon of strategic prediction, we human beings bypass global assessments of a life in favor of heuristic assessments of our futures, typically at some relatively near-in predictively accessible horizon.[40]

[39]And one's objective is assumed to be getting that assessment to come out suitably positive; a recent and characteristic expression of the sentiment: "what matters to a person is the amount of well-being in their life as a whole" (Crisp, 2006, pp. 161f). On your deathbed, you hope to be able to give the *right* answer to: Was my life, taken as a whole, well-lived? Was I *happy*?

Adam Smith complained that what Hume had gotten right about the way people appreciate 'utility' was actually perverse: we end up valuing the means that could be applied to bring about ends more than we actually care about the ends themselves (1976, pp. 297f). (And he compliments himself on the observation: this perversity "has not, so far as I know, been yet taken notice of by any body".) If the means are generic, and acquiring them is positioning, we have accounted for anyway some portion of the phenomenon.

[40]The point is related to the observation in Lenman, 2000, that the global, end-of-time assessments standardly required by consequentialist theories are not available to decisionmakers—where here the point is scaled down to human lives. (To be sure, human beings have a rough sense of how much time at most they have left on the planet, whereas states don't have the cutoff point set ahead of time. But generally for the purposes of planning and assessment, people can't see as far out as their expected lifespan.) For other worries about taking as a target an overall assessment of one's life, as assembled from more local assessments, see Velleman, 2000, ch. 2, and Millgram, 2010.

But surely there is this disanalogy: the chess program's heuristic evaluations are calibrated in view of an in-principle objective, namely, checkmating the opponent. For heuristic assessments of one's future positioning to be possible, doesn't there have to be some-

This imperative to resource acquisition and positioning involves cross-temporal tradeoffs, in particular, of costs and benefits across time: to build up a skill set over the long term, for instance, a self has to be able to assume the costs of education over the short term; to salt away an emergency fund, a self has to be able to defer spending here and now. Seeing a sacrifice made now as justified by the rewards that accrue later is, in the terminology of the personal identity debate, "compensation". Once our mad scientist has selves engaging in research—in investigations that might extend indefinitely far into their future—and strategically positioning themselves so as to be around for that indefinite future, and so as to be able to support their research activities, cross-temporal compensation—the ability to treat causally-linked costs and benefits at different times as weighing in together in one's deliberations—is a capacity that he needs to fold in.[41]

We needn't slide into thinking that compensation makes sense to the temporally extended agent because he sees his life as a whole: because, to him, all of it is—as Thomas Nagel sometimes put it—equally real. That is, we're not assuming the tradeoffs are being made so as to maximize the overall or aggregate well being of the life. People obviously do pay attention to their interests, but those interests amount to positioning, which is being judged in more or less the way that chess programs evaluate states of the board a few moves out; there need be no overall assessment of a life as a whole in the offing.

If this is the right way to look at it, we need to put on our agenda, for a further occasion, a reassessment of aggregation approaches to social welfare.

---

thing that counts as winning the game of life? Recall our earlier discussion of the agenda of New Wave bounded-rationality theory: we need ways of selecting heuristics that don't presuppose already having at hand a nonheuristic finish line.

[41] Following up on a previous note, the point also applies to selves that position themselves so as to be able to spend their way out of practical inconsistencies, on the occasions in which that is a sensible option.

While this is a fairly weak version of compensation, as seen in persons, compensation is also fairly weak: cross-temporal compensation exhibits temporal asymmetries, discounting, and even forms of discounting that themselves engender new practical inconsistencies: Ainslie, 1992, Hanson and Ainslie, 2009; Nagel, 1978, p. 71n1, attributes to Bernard Williams the observation that "Thank God that's over" displays (entirely reasonable) temporal bias.

It is also worth registering that cross-temporal compensation—in particular, treating later benefits as justifying present costs—makes sense only in sufficiently stable environments, in which up front costs *can* tie down future benefits. In our Weberian economic system, parents think that their children are more likely to succeed if that attitude is successfully inculcated, and so they read them bedtime stories about grasshoppers and ants, but in environments where investments are too likely to be wiped away by events, the stance may be far less appropriate.

States, we observed, pursue positioning; what voters opt for—a plausible collective analog for summing an individual's preferences—is simply another matter. Under what circumstances, and for what reasons, would we try to quantify, add up, or otherwise balance individuals' positioning in the course of making decisions for a group or population as a whole? This does not strike me as a simple question.

7.7

And here's where we've arrived. When you try to implement selves, you either find yourself already committed—or sometimes it's just a reasonable next step—to building out some of the main features of personal identity over time.

What exactly the nuts and bolts of the implementation are—how the demands that have to be met to have a working self, or its natural extensions, *are* met—matters a good deal less than the requirements themselves. In our own case, we find a complicated (even kludgy) system that helps itself to familiar raw materials: bodies that are fairly stable and normally discrete packages, memories that allow their owners to keep track of a history of revisions to a field of representations, and to undertake even lengthy investigative activities. These are implementation constraints, not *what it is* to be the same person over time.

Surveying those requirements, taking stock of how they are motivated, and what they allow us to do and to be, we see selves to be what nowadays gets called a life hack. Because the device works so well, and works so much better when it is augmented in the ways we have been sketching—ways that configure selves into persons with extended life histories—a self is not just a life hack, but a life hack with a future.

Chapter 8

Conclusion: Being Expelled from the Garden

I have adopted the somewhat unusual procedure of alternating pairs of chapters composed in a familiar essay format with substantial interludes, in which we've stepped back to see the moves we've made in a broader context. The interludes have considered what methodological posture—and what conception of metaphysics—makes sense of those moves, how the approach is distinguished from other ways of handling the various subject matters, and what further philosophical problems are being put on our to-do list. As we bring the book to its close, I want to take yet a further step back, to where we can ask what conception of philosophy itself is implicit in the argumentation we've traversed. Doing that will allow us to set an agenda for a next lap of investigation, and one that I will argue has in fact become urgent.

In the course of engaging a number of ongoing discussions within philosophy, on questions such as first-person authority, extended minds, defeasibility and personal identity, I've been laying out an administrative metaphysics of the self. And now that the approach has come into focus, we need to do a second pass over two related and—from some quarters at least—likely deeply felt objections. Where do I get off making the self out to be a device supporting a consistency management technique? Am *I*—you may have been asking yourself—supposed to be some sort of gimmick or contraption? Doesn't everyone care a great deal about the self that they are? Isn't consistency something that is primarily a concern for academics, and not even all of them? And so isn't the fit poor, that is, between the more or less universal stake people have in their selfhood, and what looks like a special interest in the way that, on this account, selves matter?

Second, and more abruptly, how can explications of this sort be *philosophy*? In our previous pass over the question, I addressed the sense that to see metaphysics as intellectual ergonomics makes its subject matter unacceptably mundane. But even once that is no longer an objection, there is a great deal of discussion of the everyday that is by no means philosophy. To meet these qualms head on means not treating that last question as merely rhetorical, but as deserving a proper answer.

In putting my cards on the table, and articulating the conception of philosophy in play here, I will also be speaking to what is quite clearly a severe quality control problem in the field. Philosophy has a frankly awful reputation. On the one hand, outsiders looking in see a playground for massively intellectualized daydreaming, where professors build castles in the sky out of possible worlds, sense data, grounding relations, and many other fantastical concoctions. On the other, philosophy is understood to be—as a rideshare driver once baldly put it to me—a subject that has to be mostly just common sense. Wasn't philosophy launched by someone insisting that the question he was trying to answer was how he—how *anyone*—should live?

The reputation is by no means undeserved, because while there is something that philosophy is for—and that something is practically important, and central to the ways that we human beings address the challenges we face—it's also all too easy for philosophy to collapse into a sort of imitation of itself. The failure is almost inevitable once the awareness of what philosophy is about is lost, and in spelling out how the real thing gets displaced by the ersatz, we will not only be vaccinating ourselves against what is in a very real sense a moral virus, but also be explaining why it makes sense to settle on the bureaucracy of inference as a frame for our self-understanding.

Let me use as my point of entry a recent and insufficiently appreciated discussion of the social dynamics that drive philosophical traditions, part of which is a story about how they emerge.

8.1

With the right institutional framework—roughly, economic support for a class of people who are in the business of defending opinions about something or other, and who have sufficient freedom of maneuver in adjusting and developing positions and argumentative support for them—philosophy can arise out of, it might seem, almost any subject matter: the crude protophysics of the ancient Greeks, the theological disputes of medieval Islam

and Christianity, the political and social concerns of the warring states of China, scriptural concerns in India and so on. As it becomes apparent to participants and onlookers that arguments are not going to settle whether (sticking with just one of these) everything is made of water, air, fire, earth or numbers, one of the players advances a skeptical view; another player starts to investigate the mechanics of argument, that is, turns to the study of logic; another player responds to the skepticism by looking for a skepticism-resistant anchor—normally, something with the look and feel of the Cartesian *cogito*—and suddenly the debate is recognizably philosophy.[1]

A moment ago, I registered agreement with Randall Collins—from whom I'm borrowing the account—that a philosophical tradition could apparently crystallize out of almost any subject matter. But while his sketch of the launch phase of a philosophical practice strikes me as plausible enough, at second glance, that claim, put as I just had it, doesn't look quite right. What launches philosophizing is the evident futility of the previous back-and-forth; neither empirical investigation nor a priori argument are going to settle the question of which of the four 'elements' makes up the universe. In retrospect, these protophilosophical debates were the stuff of late-night dorm room bull sessions, and not anything like science, or even science in the making, and not, on second thought, merely because they weren't going to be successfully resolved. Even the very best philosophizing does not normally establish anything that looks like a result, a theorem that can be taught as uncontroversial fact to novices in the field, and real philosophy does not have whatever *this* problem is. Rather, ancient Greek protophysics was characterized by systematic failures of reference: talk of "the wet" and "the dry," for instance, just wasn't connected up to anything. What comes 'before philosophy,' to borrow the title of a once-standard survey, looks to us as though it is merely going through the motions of thought.[2]

Now, in his extensive treatment of the ensuing traditions, Collins doesn't discuss the way in which the alchemy that transmutes arguing about nothing into philosophy is so often reversed. But it's remarkable how easily philosophy lapses into debates that are functionally on a par with the futile controversies of ancient Greek protophysicists and their ilk. Too frequently, we don't notice that it has happened, in part because they're conducted within institutional frameworks erected or appropriated by philosophy, and in part because the subject matter is inherited from a properly philosophical conversation. The participants are arguing about, as it may be, the correct

[1] Collins, 1998.
[2] Frankfort *et al.*, 1960.

definition of knowledge or the semantics of the 'actually' operator; they seem to be thrashing out the details of the correct consequentialist (or virtue-theoretic, or Kantian) ethics; they seem to be colonizing new topics, such as the grounding relation or the problem of peer disagreement or higher-order metaphysics. But they're now engaged in, and let's introduce the term we need, *hypophilosophy*.

To give an account of the distinction between philosophy and hypophilosophy will be to try to say what philosophy is. Consumer warning up front: it wouldn't do to be too confident about it. After all, philosophers have been arguing with each other over that question as well, ever since the very beginning of the enterprise. And full disclosure, I myself sometimes see different and very possibly competing conceptions as compelling, and in any case, as definitely in play.[3] Here I mean to be making the best case I can for one of them, where the proof of the pudding will be in the eating; I'll advance the proposal with, of course, a creature construction argument.

8.2

Suppose you're a god, building creatures to inhabit your newly created world: not, however, the omniscient God of the theologians, but rather the kind of god who wouldn't necessarily realize that you needed an Eve until Adam plaintively reminded you. Because you have quite limited powers of foresight, you use the Garden of Eden as your development and testing site, and there you are, with a pair of humans that you've got walking, talking, and performing other basic but crucial tasks successfully. You mean to load one or two more software packages onto the wetware before you introduce them into the wild, and you're going to do it biochemically, via a procedure with the aesthetics of a Cronenberg movie, namely, by having them eat a fruit. What kind of fruit should that be?

If you were the theological God, the one who knows ahead of time what kind of natural and social environments your humans are going to face over the long term, you would no doubt compile a list of positive and negative evaluations to guide them and their descendants safely and productively through their lives. And so you'd arrange to have your humans eat from the tree of knowledge of good and evil. Alternatively, you could work up a list of foolproof rules, borne by the tree of commandments. (You'd need a good deal more than ten of them, however.) But as it happens, you're not

[3] Tersely reviewed at Millgram, 2023b, pp. 260f.

that kind of god, and who *knows* what your humans will have to cope with down the road?

Here's one device you *have* settled on. Your humans are going to be reasonably good, sometimes even very good, at picking up on and enforcing representational hygiene regimes. That is, your humans have already been built to maintain and extend internally stored systems of representations that drive their behavior, and these hygiene regimes impose global—rather than representation-by-representation—constraints on those systems. The humans feel themselves impelled to adjust the representations, to make them conform to the constraints when they notice them being violated. Some of these constraints the humans will later come to think of as logical (as when having two representations, one of the form p, and the other of the form $\sim p$, in the store of representations counts as a violation), but not nearly all of them. For instance, when someone demands that you verify that your intention is such that everyone in your shoes could act on the same intention, without thereby making you ineffective, that's traditionally classified as a moral requirement.[4] And as we know, for a while some of these hygiene constraints have been budgetary.[5] So you have your humans eat of the fruit of the tree of consistency, and this is very clever of you: noticing and trying to remove inconsistencies can be expected to engender all sorts of strategies that will help them cope with novel, fluid, hard-to-anticipate environments. For instance, when humans are trying to decide which of p- and $\sim p$-shaped representations to delete, they're prompted to *investigate* the question of whether p; when they're trying to balance the budget, they find themselves needing to *reconsider priorities*.

But now your earlier quandary has just reproduced itself: you can't predict what environments and concomitant challenges your humans are going to encounter, or what resources those environments will supply, and so you can't be sure which consistency constraints it will be helpful for them to enforce. Budgets can be very useful, but not until there is currency, and writing, and numerals, and arithmetic, and for advanced forms of budgeting, spreadsheets; moreover, as the Austrian economists emphasized, if you're going to do accounting, you need markets to generate prices. As you wander around the Garden of Eden in the heat of the day, you realize that you

[4]For a standard presentation of the constraint, see Nell, 1975.

[5]In the real history, as opposed to our just-so story, we don't want to underestimate what it takes to get such a constraint accepted. Deringer, 2018, documents the slow process by which numerical monitoring of government and corporate financial activities became normal in eighteenth-century England, and so shows that even a commitment to budgetary consistency is not to be taken for granted.

have no idea whether human history is going to produce those technologies. Some of the humans will eventually insist that the constraints that they call 'logical' are written into the metaphysical nature of things (whatever that means, you will one day mutter to yourself); but you, as a not-so-knowledgeable god or demigod, can't discern which of the representational hygiene constraints that reflect deep structural features of the world it will be intellectually and practically useful to enforce. There's only one solution, which is somehow to delegate to the consistency-hungry creatures you're about to release into the ecosystem the task of selecting the consistency requirements themselves. And so you do, by sending them along to eat of the fruit of yet another tree, which we'll get around to naming in just a moment.

8.3

Now, what will it look like, when they get around to doing what you've set them up to do? The humans are receptive to the invitation to adopt new and even rigorous inferential hygiene regimes, and so a character who goes around proleptically insisting on them, in the course of demanding that people explain and justify their practices, is likely to get uptake. He might even give rise to a highly formal spectator sport, as one speculative account of the shorter Platonic dialogues has it.[6]

Flagging an issue to which we'll return, here's what that 'proleptically' was about. The term marks the peculiar register in which parents sometimes tell their children that good boys and girls like them brush their teeth—that is, the attempt to realize the desired outcome by anticipating it in a description. The character we started to imagine a moment ago, who in some ways looks a great deal like Socrates as Plato represents him, acts as though his novel and thus unfamiliar demands for consistency were what consistency *already* required: as though this were what consistency just *was*.[7] That's misleading; someone can always make up another way to be consistent. Trivially, it's important nowadays for one's spelling to be consistent, but back in Shakespeare's time, it didn't used to be. These days, decision theorists will describe your preferences as inconsistent if they can't be represented by a von Neumann-Morgenstern utility function; this

[6] Ryle, 1966.

[7] As Vlastos, 1994, observes, Socrates never argues for the requirements of the *elenchus*, or for that matter raises the question of its definition; rather, he suffices with explaining what they are.

is historically recent usage. Jewish dietary law sorts foods into different classifications (dairy, meat, and neither), and requires that meals containing an item in one of the former two classifications not contain items from the other. As Ian Hacking and others have documented for us, the notion of probability that we take for granted today—and so, the requirement that probability assignments be consistent—is of recent provenance.[8] So why the perennial temptation, one to which philosophers regularly succumb, to think that something or other is what consistency *just is*?

Let's return to the question of how your humans are supposed to do an intelligent job of selecting consistency regimes. If you've done your part especially cleverly, you might have coded into the final fruit plate the impulse not only to impose consistency regimes and rethink representations and the practices that depend on them when those regimes are violated, and not only the impulse to invent and adopt new ones, but also the impulse to systematize *those* consistency regimes as systems of representations, and to impose appropriate consistency regimes on *them*. That is, you might set them up to reify their representational hygiene regimes into theories, which they can then investigate, and which may prompt them to reconsider their priorities—and amongst which they can then intelligently choose.

Some of these theories will be familiar to philosophers today as logic, but those will normally be accompanied by ancillary theories as well. For instance, inferential consistency regimes that mark successful representations as true or as knowledge will invite theorizing about these statuses; inferential hygiene requirements that constrain credences will invite theorizing about probability. And it is not only this sort of tag for a step in an inference that will attract theoretical attention. For example, looking down a bit later on your creatures, you notice that one of them has invented just such a requirement: a Principle of Noncontradiction, applied to predicates, on which one and the same thing cannot be, in the very same respects, both F and $\sim F$.[9] This version of inferential consistency, so different from the propositional consistency constraints you had been imagining, needs its users to demarcate its scope of application, namely, the things that can't have those incompatible predicates. (For instance, that Principle of Noncontradiction will only be usable if what it is applied to has edges, i.e., places where it stops; otherwise, it is all-too-easy for me to be, say, both pale and not pale—because I'm tanned, *over here*, where where we would

[8] Hacking, 1975, Daston, 1988.

[9] Aristotle, 1995, at IV 3–6; for background discussion, see Politis, 2004, ch. 5 and Gottleib, 1995.

normally say *I* am, but pale *over there*, where we would normally say *you* are.) That is, consistency regimes of this sort generally involve developing additional intellectual accessories—in this case, scopes of application of a principle—which themselves become subject to further consistency regimes. And, sure enough, the inventor of this particular Principle, you admiringly notice, does develop a theory of the scopes of application of his Principle; he calls the most important of these *ousiai*, which we render as 'substances,' and his successors spend north of two thousand years arguing over whether a theory of those intellectual accessories can itself be consistent.

By now, you've no doubt figured out what we're imagining you serving to your creations before you expel them from the Garden of Eden. They will be made to eat of the fruit of the tree of philosophy.

8.4

Philosophy, then, on the picture that is emerging from our creature construction argument, is centrally the enterprise of inventing, rethinking, revising and, when appropriate, replacing consistency regimes. So now let's anticipate what will happen when that focus is lost.

We can suppose that by this point there is already a patchwork quilt of theories in place, both about inferential consistency itself and the intellectual accessories involved in such a hygiene regime. There will have been ongoing efforts to adjust them to conform to second-order consistency regimes; these present as attempts to figure out what the correct theories of statuses and intellectual accessories like truth, knowledge, substances and so on are. But such constructs are contentful only insofar as they are part of a very practical enterprise: an ongoing inference management program (or more broadly, a program of representational hygiene management).

Units of currency, like dollars or euros, are analogous constructs, used in financial accounting procedures; grades—I mean, the sort of grades a teacher gives out in a classroom—are similar constructs used in ongoing processes of assessing academic performance. Imagine that somehow theoreticians managed to forget these very concrete facts: that they came to spend their time arguing about what a dollar *really was* (*metaphysically* really was: if you remember your history of economics, think bimetalism), or about what philosophical theory of grades was in reflective equilibrium with widespread intuition. The moment this happens, a perfectly reasonable theoretical activity, one that has to do with imposing consistency on, respectively, resource allocation decisions and pupil assessment, has turned

into the theory of nothing.[10]

And in just that way, when the practice of evaluating and redesigning consistency regimes lapses—because philosophers come to take their logic and other such matters for granted, or for whatever other reason—the exercise of projecting those consistency regimes into theoretical form and imposing further consistency requirements on them slips into hypophilosophy. Detached from the will to rethink one's intellectual and practical ground rules that gives those theoretical constructs their life, what looks like metaphysics, like philosophy of language, like metaethics and like epistemology. . . is suddenly the theory of nothing.

There's a distinctive look and feel the slide into hypophilosophy can have, one having to do with the way that philosophy differs from many other intellectual disciplines. A former colleague once remarked to me on the nearly opposite ways in which the term "philosophy" is used inside and outside the field. When a layman announces that his philosophy is such and such, he means that he is informing you of convictions he is not about to rethink; whereas it is no coincidence that philosophers do not use the phrase "my philosophy is," for what a philosopher properly does is rethink his convictions. Philosophy etiolated, although its practitioners will not use the giveaway words, could for the most part be prefixed by that *my philosophy is*.[11] Whether or not the contrast is as cut and dried as this, we've put ourselves in a position to entertain an explanation for it.[12]

Philosophy, we are allowing, is in the first place the enterprise of developing and assessing consistency regimes—which we noticed can take the form of arguing about substantive theses, when the consistency requirements are put in material mode. Because such consistency regimes govern your reasoning, there is nothing more intellectually basic than they are. This means that when you assess them, design alternatives to them, and invent novel inferential hygiene regimes, there's no firm ground to stand on: nothing you can simply take for granted. This means in turn that, anyway when engaging the central subject matter of philosophy, you are always in

[10]Rather similarly, Rawls, 1955, points out that there are only strikes, bases, home runs and so on within a game of baseball. If you start working on a theory of bases, fouls, etc., only without the baseball, you will be developing a theory of nothing.

[11]The adjective is appropriated from Austin, 1975, pp. 22, 92n, 104, 122.

[12]Is it that cut and dried? After all, sometimes the layperson *does* change 'his philosophy,' and there are a great many philosophers who defend views that they will not consider abandoning under any circumstances. But the status of the layperson's 'philosophy,' as a fund of convictions on which he draws in addressing other issues, remains even if the fund experiences some churn, and defenders of their theoretical fortresses, even if they do not change their minds, are in any case supposed to defend them.

the mode of Kuhnian revolutionary science. What characterizes so-called normal science, in Thomas Kuhn's description of it, is reliance on a toolkit of stable inferential techniques; these make up much of what he famously called 'paradigms'. Amongst the academic disciplines, philosophy is unusual in that there is no normal-science phase, nothing that can *just* be the workmanlike activity of theory construction and application. Because what is being revised is that basic inferential toolkit, philosophy *normally* requires the intellectual posture appropriate to scientific revolution.[13]

Can that be right? Surely getting to the bottom even of a proposal for the revision of a consistency regime will require seeing what it amounts to when it is fully realized; if the proposal is sufficiently ambitious, no one person will be capable of getting to that point; so won't it be necessary to develop the proposal cooperatively? And won't that mean having philosophers who work within it, and so who treat it as Kuhnian normal science?

Something in the way of cooperative development is no doubt necessary, but when the participants cease to be self-aware about the enterprise—when they forget that the point is assessing and debugging the consistency regime itself—the track record shows, or so it seems to me, that their achievments turn out not to be fit for purpose. And the usual sorts of training for normal science—learning to accept the dictates and methods of your philosophical school as given—are not conducive to self-awareness. Adopting the posture of normal science ("we're on the right track, we have results, but we 'must do better'") sets you on the short and slippery path into hypophilosophy.[14]

Our history allows us a good deal of confidence in that determination. Normal science generates results. Certainly it would have been premature for Socrates to insist that philosophy was not going to succeed in producing theories that counted as established results. But we have an up-of-two-thousand-year track record that he did not, and it now looks to be the conclusion of a well-supported induction that philosophy shares with hypophilosophy not just an apparent subject matter, but this inability. The point of doing philosophy cannot reasonably be to arrive at those theoretical conclusions.[15]

[13] Kuhn, 1970.

[14] The quoted phrase is the title of Williamson, 2006.

[15] But now we can notice that the observations we used to launch our discussion need to be still further qualified. As a matter of history, philosophy does seem to arise, again and again, out of, as people used to describe *Seinfeld*, a show about nothing. Having identified the primary role of philosophizing as the development, assessment, and management of consistency regimes, we can explain the character of those occasions by invoking a rule of thumb, that science tends to appear late on the scene, after philosophizing itself has

Taking a moment or two to characterize the collapse into hypophilosophy, it seems to be in the first place a matter of your intellectual activities becoming disengaged from their objects, in a way that drains what you say of the content it formerly had. If you move the gearshift into neutral, disengaging the engine of your car from its driveshaft, and then start revving the engine, turning the steering wheel and slamming on the brakes, you are only *pretending* to drive, and hypophilosophy is very much like that: it is Pretend Philosophizing.[16] Alasdair MacIntyre famously compared the state of our moral discourse to the pseudoscientific language of a futuristic dystopia, in which science has undergone generations of suppression-by-bookburning. "What we possess," he suggested, "are the fragments of a conceptual scheme, parts which now lack those contexts from which their significance derived."[17] Where philosophy collapses into hypophilosophy, we see this sort of disconnect from context and the consequent hollowing out of an activity—which seems to go on, just as before, while no longer being *about* anything.

8.5

Why does this happen so easily, and how is it that we don't notice? If I am right, philosophy is inferential engineering, but other somewhat similar engineering enterprises—think of what is involved in designing the rules of the road—don't seem to engender the sort of confusion we have been considering: no one imagines that 'right on red' is a nonnatural normative truth, that is, a special sort of invisible fact. Nevertheless, the historical record suggests that, when it comes to philosophy, this shift has occurred not just once, but repeatedly. The most visible indicator is disciplinary revolt: to mention only two recent examples, our own founding fathers, the logical positivists, articulated their dissatisfaction with their Hegelian predecessors by describing their output as literally nonsense; the Cartesians dismissed their scholastic predecessors as engaged in merely verbal quibbles. (Again and again, sooner or later, we *do* notice.) We have all been taught the

---

become entrenched.

However, the various special sciences generate and rely on their own consistency regimes, which then come to need all of the services that philosophy is in the business of supplying. Thus we find Whewell, 1847, an early example of philosophy of science as we know it, in which the conceptual and methodological problems of the sciences are an occasion for philosophizing that—with a little squinting—looks like someone trying to understand and critically reconstruct their evolving consistency regimes.

[16]The gearshift metaphor can be found in Thompson, 2006.

[17]MacIntyre, 1981, p. 2.

objections to the logical positivist criterion of meaning, and that contributes to our thinking of their move as in the first place technical. But it must have been, as it was for the Cartesians, something on the order of an observation. The German idealists, like the scholastics, had at one point been thoughtful, penetrating philosophers; so the recurring complaints, usually couched in metaphors of emptiness or hollowness, or as charges of triviality, witness that philosophy has repeatedly been corrupted into hypophilosophy.[18]

We shouldn't suppose that the predecessors whose endeavors were so brutally dismissed thought they were engaging in Pretend Philosophy. In his *Attack upon 'Christendom'*, Kierkegaard bewailed the state of religion in Denmark: a country full of churches, which were filled with churchgoers, who had been baptized into the Christian religion, in a country whose official, established religion was Christianity—and despite all that, a country that contained vanishingly few Christians.[19] Kierkegaard did not mean, as someone might today, that the government was providing unneeded services to a now-secular population; in the Denmark of his time, almost everyone took themselves to be practicing believers. Rather, he was reminding us that you can think yourself to be a Christian, and go through the motions of Christianity, even though your faith has been entirely lost, leaving behind ossified institutional and behavioral remains that we could, varying the vocabulary we have started to use, call *hyporeligion*. Philosophy is evidently like *that*: there is no presumption of first-person authority—you will recall, one of the topics of previous chapters—with respect to whether you are philosophizing. But how is that possible? How can you lapse into hypophilosophy unawares?

A few turns back, we noticed that novel consistency regimes tend to be advanced in a proleptic register: as though the new proposal were what consistency already just *was*. You, as the god or demigod of our creature construction story, might have anticipated that. Suppose you had thought

[18] MacIntyre, 1990, pp. 159f, provides a diagnosis which we can treat for the moment as the null hypothesis. Spelling it out, a professionalized, institutionalized version of philosophy will focus on sharpening up entailments: if you hold *this*, *that* follows. But, as one sometimes says, one philosopher's *modus ponens* is another philosopher's *modus tollens*: you can choose whether to accept *that*... *or* reject *this*. So a professionalized version of philosophy ends up settling on conclusions by weighing tradeoffs, and tradeoffs are made on the basis of substantive evaluations. However, professionalization precludes stating substantive evaluations as unargued starting points. And so positions in this sort of professionalized, institutionalized philosophy are adopted on the basis of evaluative commitments from outside the field. Now, progress in a field happens when theories and positions respond to substantive evaluative assessments that are internal to the field. So, MacIntyre concludes, professionalized, institutionalized philosophy won't exhibit progress.

[19] Kierkegaard, 1968.

it would be best to send your humans out into the world equipped with a starter set of consistency conceptions. Then you would hand out the ones you guessed they would need first, to get by on for long enough to start developing their own. (And if you weren't going to bequeath them any yourself, then necessity would turn out to be the mother of invention, and the ones they were going to need first would be the ones they would manage first—or not at all.) Suppose that among those would be consistency checks suited for mundane facts about what philosophers have gotten in the habit of calling medium-sized dry goods, that is, suited precisely to descriptions of items that are already *just there*.

Then when the time came to consider and assess novel consistency regimes, along with their intellectual accessories, and so to reify them into a form that could be checked for consistency, your creatures' default would be to leverage that earlier investment in factual consistency, by casting the representation of the hygiene regime as a description of mundane fact. But the ensuing practical benefits of doing it that way come with costs, in the first place, a hard-to-override tendency to think of consistency, truth, knowledge, probability and so on as things that are just the way they are.

There was in addition history that promoted this way of construing the philosophical subject matter; here I'll just mention some of the relatively recent past.

During the Enlightenment and its immediate aftermath, what crystallized into the profession of philosophy as we currently know it was the battleground of an ideological war. Spinozists attempting to undercut the legitimacy of the established institutional structures were resisted by an emerging professoriat whose sponsors had a use for a moderate Enlightenment ideology, one that was up-to-date enough to be creditable, yet which would still underwrite the extant churches, monarchies, aristocracies and so on.[20] Because the point of the ideology-for-hire was to insist that the already-entrenched institutions were nonoptional, metaphysics, moral theory and all the rest had to be presented as a theory of *how things are*: as a straightforward matter of fact. After all, ways of doing things, and devices for doing them, have alternatives. So an academic philosopher who positioned himself as being in the intellectual machine-tool industry (that is, as making the intellectual tools which are used to make intellectual tools: in the present view, the real business of philosophy) would have been thereby unemployable.

Finally for now, proper philosophy is *difficult*. The intellectual task

[20]Israel, 2001, and Israel, 2006.

on its own is hard enough for a few generations of philosophers of science to have been at a loss when the time came to explain how it could be managed rationally.[21] But it is also emotionally difficult, and our creature construction narrative gives us a way of explaining why. If in the course of staffing your development and testing facility, you had made the mistake of hiring an over-eager serpent as a lab assistant, and that lab assistant had made the mistake of convincing your humans-in-preparation to eat of the tree of knowledge of good and evil, your humans would have ended up going out into the world running incompatible software packages: one of them providing a fixed and affectively entrenched system of evaluations, and the other, a suite of applications meant to allow the development and revision of, among other things, systems of evaluations. We inherit, not from a talking snake but from our evolutionary and older cultural history, preloaded responses that are tantamount to such an entrenched system of evaluations.[22] So philosophy done right is bound to be experienced as an inner struggle.

Perhaps there is more to the story. But whatever the full explanation, philosophy is for the most part taught today as an enterprise that produces theories about matters of fact. Students are exposed to a run of failed theories, and are made adept at the ins and outs of a number of theories that seem to be still in the running. (There are occasional exceptions, in Wittgenstein-influenced pedagogy, for instance.) And so it is not as though we all started out knowing what philosophy was about, and only then lost track of what we were doing; rather, we were almost all of us raised into hypophilosophy.

The citizens of Kierkegaard's Denmark knew that they were Christians, because it said so on the birth certificates, and our academic professionals know that they are philosophers in almost the same way: they have PhDs certifying that they are competent researchers, they are employed as professors of philosophy, they go to philosophy conferences, they publish ever more journal articles in philosophy journals—and throughout, as though there was a storehouse of accumulated philosophical results that could be disseminated, archived, and recited to students.[23] Of course in all

[21] Kuhnian paradigm shifts were famously said to be irrational, and Fisch and Benbaji, 2011, is a recent example of philosophers struggling with the problem.

[22] For a partial description of the workings of one of those preloaded responses, see Kelly, 2011.

[23] If philosophy is, as I take it to be, an engineering science, shouldn't it be possible to certify practitioners, etc.? That's trickier than you might think. While one can study the inferential engineering techniques currently in use, as well as past devices that have been

of this there need be scarcely any philosophy at all. But the presumption that it is philosophy does not have to be moral perversity on the part of its practitioners; for a long time now, hypophilosophy has been our original sin.

Specialists in more established subfields of philosophy—metaphysics, epistemology and the like—are sometimes inclined to ask of less established subfields—feminism, history of ancient Chinese philosophy, bioethics and so on—why what they are doing is *philosophy*. These questions are heard as dismissive in large part because the burden of explanation is presumed to be asymmetrical, falling exclusively on the latecomers. But if I am right about the prevalence of hypophilosophy, everyone who means to do philosophy owes an explanation of why what he is doing *is* philosophy. That means first of all giving a well-motivated characterization of what philosophy is. Familiarity does not count, and no one is exempt; we have seen why work in metaphysics and related areas is especially likely to prove to be no more than hypophilosophy.

8.6

Philosophy as I've just described it is ruthlessly practical; it is driven by the needs of logic, understood as an engineering science. But the next ingredient we need in the mix is the acknowledgement that philosophy can also be playful; it can be advanced and regulated not just by the imperatives of the clients of inferential hygiene regimes but by autonomous philosophical aesthetic sensibilities; it can take on problems and approaches to them that have no obvious practical payoffs. No doubt that's to be expected. Recall that creature construction arguments resemble state of nature arguments in many respects; when you are convinced by a Hobbesian state of nature argument that security functions are legitimately provided by the state, it would be a mistake to conclude that when you examine actually existing states, they will provide *only* security.

To explain how we can allow that there is so much more to the practice of philosophy, and the motivations for engaging in it, while still retaining the distinction between philosophy and hypophilosophy, rather than invoke a further creature construction, I'm going to use an analogy to introduce a new concept—one of which we will make heavy use.

We earlier on suggested that units of currency are artifacts of a repre-

superseded by newer ones, if philosophy really is always properly conducted in the mode of Kuhnian revolutionary science, its practice cannot be routinized. And when a practice cannot be routinized, such certifications have very modest value.

sentational hygiene regime, and that they belong in the same broad family of metaphysical constructs as Aristotelian substances. So let's return to money, which is often enough explained by way of its own state of nature argument. Here the state of nature is a barter economy, in which goods are directly exchanged one for another; a litany of inconveniences is recited, and then we are told of the stroke of brilliance in which a particular commodity is designated as the medium of exchange. And so we are supposed to conclude that the point of currency is to facilitate the exchange of goods: the currency in your possession represents the extent to which you have a claim on the socially available goods.[24]

Money absolutely serves that need, but in more advanced economies that is by no means all it does. Other functions, such as the efficient allocation of capital, are being served.[25] To be sure, I am the last person to suggest that all the activities of our hypertrophied financial sector can be accounted for in terms of some practical function or other.

Nonetheless, the pirouettes and tsunamis of high finance are intimately bound up with the functionality exhibited by that familiar state of nature argument; I will call the relation *tethering*. At any point in time, not all of the money can be exchanged for goods. However, if one day it happens that my dollars no longer purchase a cup of coffee (or if not coffee, then various other goods that are not themselves financial instruments), the dollars will no longer be money, and even if our financial institutions continue to trade complex instruments on intellectually fascinating markets, they will no longer be engaged in finance, but rather (we can say) in *hypofinance*.

Think of tethering as an explanatory option that navigates between the Scylla of reductions and the Charybdis of reifying what you're trying to account for into a mysterious freestanding fact. On the one hand, reductions purport to analyze away a puzzling or problematic concept or phenomenon in less problematic terms; a theory of money that specified, exhaustively, the social interactions in which it figures might count as such a reduction. But somehow these projects never seem to work out successfully. On the

[24]Graeber, 2014, objects that this cannot be the history. That is probably right, but no objection to a state of nature or creature construction argument; these do not purport to be historical. (Hobbes's point does not depend on there ever having been the founding conference at which people assented to the social contract as he describes it.) But Graeber's objection has another side to it, which is that premonetary economies have other ways of addressing the litany of inconveniences (e.g., instead of finding someone who wants to barter with you *now*, you ask a favor of someone who gives you what you need, thereby creating the obligation for you to repay him with a similar favor *later*). So I don't necessarily want to endorse the argument I'm using in the analogy at hand.

[25]For various roles played by currency, see Townsend, 1990.

other, keep in mind the philosophical awkwardness of the sort of stance in which Moore ended up, when he found himself insisting that 'good' was a nonnatural property.[26]

Like finance, philosophy is tethered to its original and fully practical uses. Not nearly all of the activity of a developed philosophical tradition is to be explained directly as serving the engineering of novel logics and analogous consistency regimes. But once the connection to its tether is lost, just as what looks like money degenerates into merely ornamental slips of paper, and just as what looks like banking turns into no more than ritual, so philosophy just as rapidly becomes hypophilosophy.[27]

8.7

We're now in a position to tie up two largish loose ends, both having to do with philosophical method. First, I've been fielding creature construction arguments throughout—in this last instance, to make a case for a conception of philosophy, and to reinforce my earlier claim that we should understand the self through the lens of consistency maintenance. But since we are seeing patterns of argument as underwritten by consistency requirements, and those requirements as properly legitimized within the design stance, we ought to ask, taking that stance, why creature constructions arguments are themselves legitimate.

Of course, from the point of view of an as-if designer, creature construction arguments make sense. But that's for an engineer, seeing things from above. What about the creature's own point of view? Why should such arguments have bite *for us*?[28] The problem for us is not to justify the

[26]Such positions somehow seem to attract variants on the label 'realism'. Right now I'm not going to embark on the sizable task of arguing you into agreeing that the both of these are extremes you should do your best to avoid. However, here is some softening up, on the reductionist side.

In his *Genealogy of Morals* (in sec. 13 of the second essay), Nietzsche famously argues that institutions, practices, devices and so on which have long histories won't take functional accounts (2000, pp. 515f). The idea is that over such a history, the item in question will have been repeatedly repurposed, and each such repurposing, in adding some features, deleting others, forcing further features into new roles, etc., makes it kludgier and kludgier. If Nietzsche is right about how the process plays out, the late-stage product will, in the end, resist not only functional accounts, but reductionist analyses.

[27]How is the discussion of hypophilosophy itself philosophy? It does seem to be fairly closely tied to the central philosophical enterprise, but it too can engender activities that count as philosophical only by way of being tethered to matters of inferential hygiene.

[28]Millgram, 1991, sec. 5, lays out the objection, with the case made by Gilbert Harman for a coherentist account of belief revision as its foil.

ways of man to God, or even to justify the ways of God to man, but to justify the ways of man to man.

Let's return to the creature construction argument that we still have in progress. You've equipped your humans with the promethean gift of philosophy: they're able to rethink, invent, develop and adopt inferential consistency regimes, along with the intellectual accessories—i.e., the metaphysics—supporting that functionality. Your humans are going to be out in the world, doing all of this on their own. That means that they will have to be able to assess prospective consistency regimes, as well as relatively minor adjustments to the ones they currently deploy.

Selves are scopes for consistency regimes. So to adopt a new consistency regime, or to modify how you operate one you have, is to alter the overall design of your very self. So you, the demigod designing your humans, will need them to produce assessments of how their selves function, when alterations are made to their consistency management. They will need to adopt the design stance, and one extremely helpful technique for doing so is argument by creature construction. So, and this will be one more conclusion to extract from the creature construction argument we're now wrapping up, you'll want your creations to be able to construct and use creature construction arguments. (But that last step is a placeholder for a much more tangible design argument, one that would consider comparative costs of alternative strategies.) Thus when your creations eat of the fruit of the tree of philosophy, you can quite reasonably decide, they will at the same time be imbibing facility with and apprehension of the force of creature construction arguments—as applied by them, to themselves.

Arguments trace out patterns in the requirements imposed by consistency regimes. That will be true of creature construction arguments also. (That is to say that our own argument has its own account of what arguments are; does our argument match that account?) It will, however, be left as an exercise to your creatures to articulate the conception of consistency that underwrites that pattern of inference. I don't know that we're in a position to do that yet; design choices are underwritten by practical reasoning, and our theoretical understanding of practical rationality is not in the shape we would like it to be.

8.8

Turning to that second loose end, I sometimes have it suggested to me that the views being developed here are 'pragmatist'. Here I do have to push

back. Pragmatism was a philosophical tradition framed by a distinctive complex of philosophical commitments, a rich heritage of positions and putative results, and an understanding of what counts as an argument that is alien to analytic philosophers. It is a mark of how distant the tradition is that the vague impression that someone cares about getting things to work is enough for analytic philosophers to think someone is a pragmatist. But there are no more pragmatists; although there are many insights yet to be harvested from the history, it's simply not intellectually possible anymore, in something like the way it's no longer possible to be a British empiricist. And accordingly, this book is not advancing pragmatist views.[29]

That said, there is a way in which the ideas we have in play can be made to speak to an issue that the later pragmatists experienced as pressing, and in which we also have an interest.[30] You'll have noticed that, at various points throughout the book, I've needed to resist an impression it's easy to get: that the task of identifying and resolving contradictions in the service of figuring things out is an overly intellectualized conception of what, for instance, personal identity is really about. (And I've been reminding readers of the various ways in which extirpating contradictions can be as hands-on practical as anything gets.) Now, John Dewey fielded an analogous response to the pragmatists' version of the worry, which was to make thinking out to be a matter of getting disrupted activities reconfigured and back on track.[31] You think in response to a *problem*, and your thinking is successful when it resolves that problem.

The problem was, however, that the notion of a problem was undertheorized, and came to function as a *de facto* primitive in pragmatist philosophizing. Dewey had memorably presented the idea this way:

> a traveler faring forth... marches on giving no direct attention to his path, nor thinking of his destination. Abruptly he is pulled up, arrested. Something is going wrong in his activity... there is shock, confusion, perturbation, uncertainty. For the moment he doesn't know what hit him, as we say, nor where he is going. (1988, p. 137)

That is, he was taking a problem to consist in a habit's being disrupted, sometimes by novel circumstances, but sometimes by another habit.

While that sort of disruption is certainly a valuable indicator, it is not an adequate analysis. When my habit of sipping my way through a cup of

[29] Millgram, forthcomingb, spells out this cluster of claims.

[30] I'm grateful to Michael Bratman for pressing me to take up this question.

[31] E.g., Dewey, 1981, Dewey, 1988, pp. 54, 90, 121f, 125, 134, 172, 207.

coffee is interrupted by a request for a lift to the train station, of course I reply, and not disingenuously, "No problem." Are automatic timeouts and bandwidth caps to derail a social media addiction creating a problem, or providing a benefit? And sometimes it's that some habitual way of proceeding continues uninterrupted, undisrupted, and unrethought which is precisely the problem; as a teacher, that's my view of some of the intellectual habits of my students, and as a researcher, that's what I think about literatures or subfields that look like they need some shaking up. The problem here is not that we haven't *defined* "problem," but rather that in one after another case we don't have a clear take on why there's a problem: on what the problem *is*.

Pressing on that shortfall a bit more, 'habit' was one of Dewey's technical terms, and we have a choice about which way to take it. Was he invoking the ordinary and descriptive notion of a habit? As I've just suggested, a habit disrupted isn't necessarily a problem. Or do we construe habit more demandingly, so as to bear a prescriptive burden, where, more or less, a habit is how you do things, or rather, how *to* do things? In that case, having a habit break down would perhaps mean that you've got a problem; how you're doing things isn't *working*. In that case, either running problem-free is built into the concept of habit, or alternatively, 'working' has become the unexplained primitive.

The problem of what to make of problems remained outstanding at the collapse of the pragmatist tradition. In our more recent analytic tradition, a fallback assumption has kicked in, that a problem invokes a goal or objective to which there isn't a clear pathway.[32] In my view, however, that conception is also unsatisfactory, because of the special status granted to instrumental rationality. Means-end reasoning can't be assumed to be what problem solving consists in; you can face problems that have to do with what your ends are, as well as problems that don't have to do with means or ends at all. Altogether, pragmatism folded down while it was still unable to say what it is to be a problem, what it is to solve one, and why problems need to be solved; we have not done particularly better in the meantime.

However, keep in mind that the practical imperative is not that of defining or analyzing problems in general, but rather of being in a position to recognize when you have one, and why it matters to you to solve it. And here we have, in our creature construction arguments, provided ourselves with most of what we need, for an important range of problems.

By designating fields of representations as scopes for the enforcement

[32] As in Fisch and Benbaji, 2011, pp. 202f, 208f.

of consistency regimes, and by defining the accompanying conceptions of consistency, we have firmed up and given content to a central class of problems for particular types of agents, and given such agents the wherewithal to identify their problems. What's more, we've shown how to justify these sorts of consistency regimes. To understand why an inconsistency of a given kind should count as a problem for a given type of agent, we ascend to the perspective in which we can see how setting such and such consistency requirements solves a problem for our demigod (or mad scientist or robotics engineer or similar embodiment of the design perspective). As we've argued, the occupier of that perspective also has to be construed as boundedly rational, and as implementing similar techniques for figuring things out; those techniques (no doubt some of them will be inferential consistency regimes with specified scopes) will make sense as solutions to problems from the perspective of a meta-designer...[33]

A moment ago, I resisted treating all problems as set by frustrated goals, and now we can fill in the objection. Problems posed by such consistency requirements are only artificially construed as driven by a goal or objective: the *problem* is the violation of a consistency constraint. Why 'only artificially'? Turning to a different branch of the practical reasoning debate, it's sometimes objected to the means-end-only conception of practical inference that, before you can look for means, often enough you have to specify what your objective is concretely enough to support that search, and that that has to be reasoning, too.[34] Likewise, sometimes you have to specify your puzzlement or quandry more concretely, in order to get it to amount to a problem, as opposed, say, to a riddle.[35] Laying down inferential consistency regimes is one very useful technique for analogously specifying a perplexity or predicament into a problem. If you like, it gives you a goal (or various goals), but only derivatively; the specification, of a perplexity into a consistency violation, tells you that the problem is, precisely, *that*. You address it, in part, by setting goals.[36]

[33]To reiterate, we need not have captured all the problems agents will have to solve, and pointing out that we have a technique for firming up and making tractably concrete certain types of problem is not the same thing as giving necessary and sufficient conditions for counting as a problem, across the board. But here we have been sidestepping (and trying to displace) that sort of definitional ritual, in favor of explanation by creature construction.

[34]For an overview, see Millgram, 2008.

[35]For the contrast, made out as between a riddle and question, see Diamond, 1991.

[36]The creature-construction perspective we're considering is close enough to that of a game designer for us to want to do just a little compare and contrast, where the discussion serving as my foil here is Nguyen, 2020.

In short, we have problems posed in terms of intellectual hygiene regimes and the tasks involved in maintaining them because firming up problems in this way solves other problems. And because we *do* firm them up in these ways in many classes of case, we *are* able to say—in *those* instances—what the problem is, and why it is one.

8.9

With tethering in mind, let's remind ourselves of where we've arrived. Despite the field's reputation as the most unworldly of the philosophical subdisciplines, we're construing metaphysics as an *applied* science. We've proposed analyzing the central objects of traditional metaphysics—selves and persons, so far, but with substances, universals, and so on also on the eventual agenda—as virtual constructs whose central function is to facilitate reasoning.

Here is the anchoring functionality we've identified for selves. If I believe that p, and I also believe that $\sim p$, I'm contradicting myself; I'm inconsistent; if I don't change my mind when it's pointed out to me, I'm irrational. Whereas if I believe that p, and *you* believe that $\sim p$, we merely disagree. No one is—so far—required to change his mind. Selves serve, first and foremost, as the scopes for consistency requirements. These consistency requirements drive inference, and more generally prompt investigations; they're a way of making us figure things out.

We identified what throughout human history has been an opaque aspect of the procedures and methods that this virtual overlay is there to

---

Games do set problems, but almost always in terms of a goal or finish line, that is, the objective of the game. Playable games can be won; normally, the objective can be achieved. However, the problems set by our creature-construction engineers need not be solvable. There need not be a clear finish line. And in the more general case, a legitimate way of addressing a problem is to make it go away (or wait for it to go away), rather than solving it in its own terms; but you don't win a game by having the problem set by the game go away.

What makes sense of these related and even similar sorts of challenges posed to agents being nonetheless different? The problems set by games are there to, in one way or another, enable play; after all, much of the time we play games so as to have fun. The problems set by the imagined agents in our creature construction arguments will generally have to do with exploring, coping with, and coming to control or anyway manage an environment that is given but not entirely in view: there, and not well understood.

As Nguyen points out, the demands made by games and the capabilities of the players are normally fitted to one another. The world, and we in it, are not like that: we are boundedly rational, and the world is too large and complicated for us to master. Selfhood is one way of coping with the disparity.

support. The schema I just rehearsed, introducing the notion of a consistency requirement, was a misleading special case: most of the consistency requirements we administer are defeasible. While we've had informal practices for negotiating their application—especially, trading explanations and objections back and forth until a local agreement is reached as to whether 'other things' are 'equal' or not—the logic of defeasibility has never been well understood. We tried to make sense of it from a design standpoint, but that hasn't equipped us with anything like an algorithm, or even a clear criterion, for determining whether it's correct to move ahead with some bit of nonmonotonic reasoning.

Because selves have been an extensively deployed and long term administrative technique, a great deal has been built on top of the virtual object we've been discussing. This is not the place for a close up examination—again, I am reserving that for a further installment—but we can start to catalog some of those reuses. We treat selves, and their temporal extensions, persons, as the basic units of our moral order. We treat them as the primary market actors, which is to say that our economy is assembled out of selves. In democracies, we treat them as the political actors who vote and who hold elected offices; we treat them as the primary bearers of political rights. There is, in addition, longstanding and deep emotional and cultural investment: just for instance, we are encouraged to understand our lives as narratives, where we are the protagonists of those narratives—and that role has built into it the core functionality of the self. We experience self-improvement as a personal imperative, and in close and especially intimate relationships, we reveal—often enough, with trepidation—our true selves. In the mythology of the Great American Road Trip, you are to embark not only to find America, but yourself.

And already we can anticipate a looming emergency: all of this superstructure is tethered to the core functionality of the self, that of being a scope for consistency management requirements. When the tether to currency's core functionality, that is, serving as a medium of exchange, is severed, we observed that apparent economic activity is hollowed out; when that happens in real life, the most frequent trigger being hyperinflation, the normal consequence is economic disintegration. Likewise, discontinue selves, and the activities and procedures tethered to them will dissolve. There would be no point in delegating to actors in a market decisions about buying and selling, if you couldn't assume that they were consistent—well, consistent *enough*—in their choices. There would be no point in asking voters for their input if you couldn't expect them to have more or less coherent preferences, and more ambitiously, to, anyway sometimes, figure things out. There is

only something that you really believe, or something you really want, if you are consistent enough to allow those ascriptions to take hold; but when you went on that road trip, to look for your real self, wasn't a central part of that finding out what you believe and what you want? There would be no point in demanding moral agency from someone if you could not also reasonably demand enough in the way of consistency from them (and so, when we can't, we don't: then there's the memory care facility). If selves were to lapse, more or less across the board, into—as we are now prepared to say it—*hyposelves*, all of that moral and political and economic activity would persist, if at all, as no more than mummery. But, as with the economic analog, in that case we should not expect it to persist.

Despite the lingering Cartesian tendency, that of treating the self as the most stable feature of our epistemic and agential landscape, we have started to see how it is that selves *are* well on the way to becoming no more than hyposelves. The self is an administrative device, one that plays a role in consistency management. If consistency management ceases to be feasible, whether entirely or for the most part, the device loses its functional characterization: in something like the way that a baseball field, once no one plays baseball anymore, will no longer be a baseball field, a scope for consistency management, once consistency management is futile, will no longer be a self.

We were bringing to bear insights from emerging work on bounded rationality, in order to explain how the design of the virtual overlay has served problem solving in computationally limited agents. But we also noticed another collective bounded rationality technique: the division of intellectual and evaluative labor. And we observed that as this sort of division of labor escalates, the older, informal methods for handling defeasibility stop working. The casual back-and-forth in which someone points out a consideration your argument has overlooked is no longer effective when you don't so much as know which expert to run your argument by, when he won't understand your train of thought, and when you won't understand what he might have to tell you. Defeasible consistency constraints are by far the largest part of the requirements selves are there to comply with; once defeasibility management is no longer a competence we can expect, selves are no more than hyposelves, and at that point, the further roles we have selves enact are no more than going through the motions.

And I'll add one more dependency to the list of endangered practices at which I've gestured. On the conception of philosophy I've advanced here, philosophy is the business of understanding, debugging, augmenting and reinventing consistency and inferential hygience regimes. These are, first

and foremost, consistency regimes meant *for* selves. If we lose the selves, we will lose the philosophy as well. We have already called to mind the sort of hypophilosophy to which we have, until now, been most prone; this is another.

8.10

How are we to proceed? I'll wrap up by indicating some of what seem to me to be the pending tasks.

I'm going to put aside for now the option of declaring that the ship has sailed: that we're about to be done with selves, and we might as well get used to the idea. We've been analogizing selves to currency, so notice that when currencies collapse, the response is not to announce that we'll just do without: governments attempt to reconstitute or reintroduce a currency, and if they do not, individuals find substitutes, usually the relatively stable currencies of other countries. We have invested so much in selves, and built so much on top of them, that they have become indispensible, in something like the way that money is. Nonetheless, I allow that it would be a worthwhile philosophical project to make out what life—and social arrangements—without selves could be.

Leaving that aside, first and foremost, we need an initial survey filling in the catalog at which I gestured: a study, in my view best conducted using creature construction arguments, of the central additional devices we have tethered to selves. This would take moral and political philosophy for its primary territory; as I've indicated, our practices of moral and political agency treat selfhood as a feature of the basic units with which they work. But because agency eventuates in decisions, and the reasoning that goes into decisions is practical reasoning, the consistency requirements that drive the underlying inferences are themselves practical. And because these are not well understood, the survey I have in mind would do well to be accompanied by an investigation of practical rationality, focused on what it takes to administer practical consistency requirements.

Second, we need to consider how we can improve our performance in managing defeasible inference.[37] As we've seen, the logical function of the ceteris paribus clauses that mark defeasibility was to preemptively make room for what finite and boundedly rational creatures like ourselves have overlooked. And because someone can have overlooked just *anything*, man-

[37]Elsewhere, I've considered such expedients as cross-educating philosophers on other specializations (2015, sec. 11.5). Here I'll proceed to another possible option.

aging the openendness of those other-things-equal clauses can present the appearance of the sort of task that requires a magic wand.

However, our diagnosis of the problem we are facing makes it out to be significantly less desperate than that. The problem was that, because we now live in a conceptually much more complicated and *much* more cluttered world than formerly, division of labor has become division of intellectual and evaluative labor. We overlook so much more because, as I've put it, when you look at division of intellectual labor, the other side of the coin is obviously massive ignorance, of most of what it's not your job to know. When you look at division of evaluative labor, the other side of the coin of is massive evaluative incompetence, with respect to almost all those things it's not your job to assess. Supporting consistency management and the reasoning it drives by flagging and vetting out-of-field defeaters for inferences-in-progress is a more restricted, and so less demanding task than the general problem of defeasibility management. In the fully general case, a consistency requirement can be upended by a fact or a priority that no one knows about, or has ever known. But in ignorance generated by the division of intellectual and evaluative labor, what you don't know about, because it's not your area of expertise, someone else, in a differently demarcated field, does—the information is available, but elsewhere.

8.11

Here I'm going to suggest a very speculative path forward. (It is the early days of the technology, and I'm aware that matters may look rather different very quickly.) In modern AI, we see the (lossy, to be sure) compression of remarkably large bodies of information, which we see subsequently evoked with high sensitivity to salience. You can imagine the AI assistants that tap you on the shoulder when you're overlooking something, in particular when it's something that only someone with expertise not your own knows about: given what you're doing, there's something in another specialization that you had better check in about. Very plausibly, even current-level large language models can accompany that with a science-journalism explanation, and indeed, one that would count as high-quality science journalism. And such an assistant can also tell the person to whom you need to talk, at the same level of science-journalism clarity, what problem you are facing.

That's not necessarily everything we need. Elsewhere, I've borrowed a label, "pataphysics," from a Dada novelist: he defined it as the science

of exceptions.[38] Machine learning, generative AI, and more generally Big Data connectionist techniques—that is, modern AI, as opposed to GOFAI, the last generation's rule-based expert systems—is about finding patterns in large bodies of data. Unless noticing exceptions is simply a matter of tracking larger patterns, quite possibly modern AI will fail at pataphysics, and pataphysics is a good tongue-in-cheek characterization of the general problem of defeasibility management. Knowing that someone else's expertise is relevant to what you're trying to do is not yet to have negotiated gaps in understanding, clashing modes of assessment, and other problems involved in attempting to solve problems that bridge areas of specialization. Nonetheless, this sort of scaffolding is definitely worth exploring.[39]

Selves built out into diachronically extended persons are, as I've phrased it, a life hack with a future. But whether this is an expedient we will be able to continue to rely on—that is, if it is to be in a different sense a life hack with a future—depends on our success in making consistency management broadly effective, despite increasing division of intellectual labor. Our exploration will have to be experimentation. How well will a statistics-driven model find connections where there are not already well-worn pathways to follow? How well can an AI agent function as a cross-disciplinary translator and mediator? Philosophers are used to working at their desks, or rather, as the cliche has it, in their armchairs; the research agenda we're considering now is something to pursue in an AI lab.

8.12

Last but not least on our agenda: In the conception of philosophy I've advanced, on which part of its job is the development and introduction of

[38] Jarry, 1996, p. 21; Millgram, forthcominga.

[39] And is the scaffolding just for humans, or does it make sense to reproduce it as a harness for AI agency? It's too early to know, but we should be aware of a case to be made that human and AI intellection are not going down the same path.

In a widely known online thinkpiece, Richard Sutton has suggested that the "bitter lesson" of seven decades of artificial intelligence research is that content- and task-specific problem-solving techniques (that is, the heuristics that power human cognition) get swept aside, one after the other, as the raw computational power of our hardware grows. (It grows exponentially: the rough rule of thumb, that the number of transistors on a chip doubles every eighteen months, has been dubbed "Moore's Law".) Advances in machine intelligence are driven by scaling, pure and simple, and Sutton concludes that we should not be wasting our time trying to get computers to solve problems the way we do.

That approach won't work for humans: nobody's brain doubles in size every year and a half. We will continue to need those content-based and task-specific techniques—of which consistency management, scoped to selves, is one.

novel consistency regimes, and on which some of those consistency regimes count as logic, we will need to face up to what has recently been renamed the "adoption problem".[40] How do we convince someone to adopt a method of inference and a mode of reasoning that he does not already deploy? You might well wonder whether it is even possible: if adopting a new logic is to be done rationally, mustn't it be done either on the basis of the logic one already has, or on the basis of the logic being newly proposed? But if you are abandoning the older logic in the course of endorsing the new and different logic, then in the former case, doesn't it mean that your argument for adopting the newer logic was no good? And won't an argument for adopting the newer logic, conducted in accord with that newer logic, beg the question? Won't it fail to the convince anyone who hasn't already accepted its conclusion?

Recall the contrast between discovery problems and invention problems.[41] The position we have been developing sees philosophy, and philosophy of logic in particular, as focused on invention problems. So notice that when we observe a succession of what we understand to be intellectual tools, we should not conclude that the ones that have been discarded did not serve their purpose perfectly adequately in their own day. Analogously, when the horse and buggy were replaced by the automobile, they were not *refuted*. And that somewhat unfastens the dilemma we just posed.[42]

---

[40]Kripke, 2024, and see also other essays in that issue of *Mind*.

[41]I have previously attempted to account for what can seem like the dismal philosophical track record—I mean, that there are no established results—on the assumption that philosophy is an attempt to solve discovery problems; see Millgram, 2013. Here I am taking a different tack, and developing the thought that the enterprise is addressed to invention problems.

[42]We can return momentarily to the Quine-Carnap debate, which we brought into view above, in sec. 5.14, n. 16. Its closure point was Quine's observation that, because logic is required to implement the conventions, logic by convention turns out to be: logic by convention, plus logic.

Conceding for the sake of the argument that circularity is an objection to a definition or to a conceptual analysis, superficially analogous circularity in implementations is unobjectionable. Tools may be used to manufacture more of the very same kind of tool, as when Intel's chips control steppers that produce more of the very same chips. What goes for tools generally goes for software in particular: I actually remember writing, in an introductory computer science class, a text editor in that very editor, and a Lisp interpreter that ran inside a Lisp interpreter. (Both were toy versions, of course.) As indicated in the main text, such implementations are normally ramped up via previous generations of a toolkit: this year's workstations are being designed on last year's.

If the virtual objects that make up our metaphysics, along with the procedures they are there to enable, are software—if they are a kind of augmented-reality overlay on ourselves and our world—then what goes for software (and for tools in general) goes for metaphysics

Here I think we have an underexplored resource to draw on. When I look out over the history of philosophy, I see philosophical traditions organized around their distinctive conceptions of philosophical argumentation. One way to describe logic might be: the science or study of what counts as a good argument. At the outset of such traditions, we find philosophers who managed to convince first themselves, and then those they managed to recruit, to adopt their novel mode of argumentation. And so we can expect the founding figures of philosophical traditions to have addressed themselves, each in their own way, to the adoption problem.

It does seem to me that, given where we have found ourselves, a series of case studies is called for. These will identify the modes of argument central to different philosophical traditions, home in on their occasions of introduction, and reconstruct how it was that philosophers were brought to accept a form of argument they had not previously acknowledged. And that is where I will leave things for now.

---

and logic: Quine's objection to Carnap simply passes them by.

Bibliography

Abbate, J., 1999. *Inventing the Internet*. MIT Press, Cambridge.

Adams, E., 1966. Probability and the logic of conditionals. In Hintikka, J. and Suppes, P., editors, *Aspects of Inductive Logic*, pages 265–316, North-Holland, Amsterdam.

Adams, F. and Aizawa, K., 2008. *The Bounds of Cognition*. Wiley-Blackwell, Oxford.

Ainslie, G., 1992. *Picoeconomics*. Cambridge University Press, Cambridge.

Andreou, C., 2023. *Choosing Well*. Oxford University Press, Oxford.

Anscombe, G. E. M., 1985. *Intention (Second Edition)*. Cornell University Press, Ithaca, NY.

Aristotle, 1984. *Complete Works*. Princeton University Press, Princeton.

Aristotle, 1995. Metaphysics. In Barnes, J., editor, *The Complete Works of Aristotle*, pages 1552–1728, Princeton University Press, Princeton. Translated by W. D. Ross.

Armstrong, D. M. and Malcolm, N., 1984. *Consciousness and Causality*. Basil Blackwell, Oxford.

Austin, J. L., 1975. *How to Do Things with Words*. Harvard University Press, Cambridge, 2nd edition. Edited by J. O. Urmson and Marina Sbisà.

Auyang, S., 1998. *Foundations of Complex-System Theories*. Cambridge University Press, Cambridge.

Bar-On, D., 2005. *Speaking My Mind: Expression and Self-Knowledge*. Oxford University Press, Oxford.

Barrow, J. D. and Tipler, F. L., 1986. *The Anthropic Cosmological Principle*. Clarendon Press, Oxford.

Bastiat, F., 1995. What is seen and what is not seen. In de Huszar, G., editor, *Selected Essays on Politcal Economy*, pages 1–50, Foundation for Economic Education, Irvington-on-Hudson.

Bayley, J., 1999. *Elegy for Iris*. St Martins Press, New York.

Bechtel, W. and Richardson, R., 1993. *Discovering Complexity*. Princeton University Press, Princeton.

Bendor, J., 2010. *Bounded Rationality and Politics*. University of California Press, Berkeley.

Bendor, J., 2020. Bounded rationality in political science and politics. In Mintz, A. and Terris, L., editors, *Oxford Handbook of Behavioral Political Science*, Oxford University Press, Oxford.

Bilgrami, A., 2006. *Self-Knowledge and Resentment*. Harvard University Press, Cambridge.

Boghossian, P., 1989. Content and self-knowledge. *Philosophical Topics*, *17*, 5–26.

Bowman, M., 2012. *Are Our Goals Really What We're After?* PhD thesis, University of Utah.

Boyd, R. and Richerson, P., 1985. *Culture and the Evolutionary Process*. Chicago University Press, Chicago.

Brandom, R., 1994. *Making It Explicit*. Harvard University Press, Cambridge, Mass.

Brandom, R., 2001. Action, norms, and practical reasoning. In Millgram, E., editor, *Varieties of Practical Reasoning*, MIT Press, Cambridge, Mass.

Bratman, M., 2001. Taking plans seriously. In Millgram, E., editor, *Varieties of Practical Reasoning*, MIT Press, Cambridge, Mass.

Bratman, M., 2007. *Structures of Agency*. Oxford University Press, Oxford.

Bratman, M., 2014. *Shared Agency*. Oxford University Press, Oxford.

Braudel, F., 1992. *Civilization and Capitalism 15th–18th Century, Vol. III: The Perspective of the World*. University of California Press, Berkeley. Translated by Siân Reynolds.

Bräuer, F., 2023. Aesthetic testimony and aesthetic authenticity. *British Journal of Aesthetics*, *63(3)*, 395–416.

Brewer, T., 2009. *The Retrieval of Ethics*. Oxford University Press, Oxford.

Burge, T., 1993. Content preservation. *Philosophical Review*, *102(4)*, 457–488.

Burgess, A. and Plunkett, D., 2013. Conceptual ethics i. *Philosophy Compass*, *8(12)*, 1091–1101.

Cappelen, H., 2018. *Fixing Language*. Oxford University Press, Oxford.

Carnap, R., 1963. *Logical Foundations of Probability*. University of Chicago Press, Chicago, 2nd edition.

Carruthers, P., 2011. *The Opacity of Mind*. Oxford University Press, Oxford.

Cartwright, N., 1983. *How the Laws of Physics Lie*. Clarendon Press, Oxford.

Cartwright, N., 1999. *The Dappled World*. Cambridge University Press, Cambridge.

Cassam, Q., 1994. *Self-knowledge*. Oxford University Press, Oxford.

Chalmers, D., 2025. What is conceptual engineering and what should it be? *Inquiry*, *68(9)*, 2902–2919.

Cherniak, C., 1986. *Minimal Rationality*. MIT Press, Cambridge, Mass.

Clark, A., 2003. *Natural-Born Cyborgs*. Oxford University Press, New York.

Clark, A., 2011. *Supersizing the Mind*. Oxford University Press, New York.

Clark, A. and Chalmers, D., 1998. The extended mind. *Analysis*, *58*, 10–23.

Collins, R., 1998. *The Sociology of Philosophies*. Harvard University Press, Cambridge.

Conlisk, J., 1996. Why bounded rationality? *Journal of Economic Literature*, *34*, 669–700.

Conway, M. and Playdell-Pearce, C., 2000. The construction of autobiographical memories in the self-memory system. *Psychological Review*, *107(2)*, 261–288.

Cooper, G., 1990. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, *42*, 393–405.

Cowan, R. S., 1983. *More Work for Mother*. Basic Books, New York.

Craig, E., 1990. *Knowledge and the State of Nature*. Clarendon Press, Oxford.

Craver, C., 2007. *Explaining the Brain*. Oxford University Press, Oxford.

Crisp, R., 2006. *Reasons and the Good*. Oxford University Press, Oxford.

Dagum, P. and Luby, M., 1993. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, *60*, 141–153.

Daston, L., 1988. *Classical Probability in the Enlightenment*. Princeton University Press, Princeton.

Davidson, D., 2001. Knowing one's own mind. In *Subjective, Intersubjective, Objective*, pages 15–38, Oxford University Press, Oxford.

Davidson, D., McKinsey, J. C. C., and Suppes, P., 1955. Outlines of a formal theory of value, I. *Philosophy of Science*, *22(2)*, 140–160.

Davis, N. Z., 1983. *The Return of Martin Guerre*. Harvard University Press, Cambridge.

Dawkins, R., 1989. *The Selfish Gene*. Oxford University Press, Oxford.

Deacon, T., 1997. *The Symbolic Species*. W. W. Norton, New York.

Dennett, D., 1981. *Brainstorms: Philosophical Essays on Mind and Psychology*. MIT Press/A Bradford Book, Cambridge, MA.

Dennett, D., 1989. *The Intentional Stance*. MIT Press, Cambridge.

Dennett, D., 1991. *Consciousness Explained*. Little, Brown and Company, Boston. Illustrated by Paul Weiner.

Dennett, D., 2004. *Freedom Evolves*. Penguin, New York.

Deringer, W., 2018. *Calculated Values*. Harvard University Press, Cambridge.

Dewey, J., 1981. Theory of valuation. In Boydston, J. A. and Levine, B., editors, *John Dewey: The Later Works, 1925–1953, vol. 13*, pages 191–251, Southern Illinois University Press, Carbondale. Introduction by Steven M. Cahn.

Dewey, J., 1988. *Human Nature and Conduct*. Southern Illinois University Press, Carbondale. Introduction by Murray G Murphey.

Diamond, C., 1991. Riddles and Anselm's riddle. In *The Realistic Spirit*, pages 267–289, MIT Press, Cambridge.

Dietz, R. and Moruzzi, S., editors, 2010. *Cuts and Clouds*. Oxford University Press, Oxford.

Dutilh Novaes, C., 2021. *The Dialogical Roots of Deduction*. Cambridge University Press, Cambridge.

Earman, J., 1992. *Bayes or Bust?* MIT Press, Cambridge, MA.

Eklund, M., 2004. Personal identity, concerns, and indeterminacy. *Monist, 87(4)*, 489–511.

Evans, G., 1982. *The Varieties of Reference*. Oxford University Press, Oxford.

Finkelstein, D., 2003. *Expression and the Inner*. Harvard University Press, Cambridge.

Fisch, M. and Benbaji, Y., 2011. *The View from Within*. University of Notre Dame Press, Notre Dame.

Foucault, M., 1975. *The Birth of the Clinic*. Random House, New York. Translated by A. M. Sheridan Smith.

Foucault, M., 1988. *Madness and Civilization: A History of Insanity in the Age of Reason*. Random House, New York. Translated by Richard Howard.

Foucault, M., 1995. *Discipline and Punish: The Birth of the Prison*. Random House, New York. Translated by Alan Sheridan.

Frances, B. and Matheson, J., 2019. Disagreement. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University.

Frankfort, H., Frankfort, H. A., Wilson, J. A., and Jacobsen, T., 1960. *Before Philosophy*. Penguin, New York.

Frege, G., 1978. *The Foundations of Arithmetic (2nd revised edition)*. Basil Blackwell, Oxford. Translated by J. L. Austin.

Galison, P., 1987. *How Experiments End*. University of Chicago Press, Chicago.

Garey, M. and Johnson, D., 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York.

Gendler, T. S., 1998. Exceptional persons: On the limits of imaginary cases. *Journal of Consciousness Studies*, *5(5–6)*, 592–610.

Gendler, T. S., 2002. Personal identity and thought-experiments. *Philosophical Quarterly*, *52(206)*, 34–54.

Gertler, B., 2003. Self-knowledge. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*, Stanford, CA.

Giere, R., 1999. *Science without Laws*. University of Chicago Press, Chicago.

Gigerenzer, G., Todd, P., and the ABC Research Group, 2000. *Simple Heuristics That Make Us Smart*. Oxford University Press, Oxford.

Gold, I., 1993. *Color and other Illusions*. PhD thesis, Princeton University.

Gopnik, A., 1993. How we know our minds: The illusion of first-person knowledge of intentionality. *Brain and Behavioral Sciences*, *16*, 1–14.

Gorman, M., editor, 2010. *Trading Zones and Interactional Expertise*. MIT Press, Cambridge.

Gottleib, P., 1995. The principle of non-contradiction and Protagoras: The strategy of Aristotle's *Metaphysics* IV 4. *Proceedings of the Boston Area Colloquium in Ancient Philosophy*, *8*, 183–198.

Graeber, D., 2014. *Debt: The First 5,000 Years*. Melville House, Brooklyn.

Grice, P., 1975. Method in philosophical psychology (from the banal to the bizarre). *Proceedings and Addresses of the American Philosophical Association*, *48*, 23–53.

Haber, M., 2025. Introduction to 'Formalization and the meaning of "theory" in the inexact biological sciences'. In Ankeny, R., Dietrich, M., and Leonelli, S., editors, *Scaffolding: Selected Contributions of James R. Griesemer to History, Philosophy, and Biology*, Springer, Cham.

Hacking, I., 1975. *The Emergence of Probability*. Cambridge University Press, Cambridge.

Hanson, C. and Ainslie, G., 2009. *Thinking about Addiction: Hyperbolic Discounting and Responsible Agency*. Rodopi, Amsterdam.

Hardin, C. L., 1988. *Color for Philosophers*. Hackett, Indianapolis. Foreword by Arthur Danto.

Hardwig, J., 1985. Epistemic dependence. *Journal of Philosophy*, *82(7)*, 335–349.

Harman, G., 1986. *Change In View: Principles of Reasoning*. MIT Press, Cambridge, MA.

Hart, H. L. A., 1968. The ascription of responsibility and rights. In Flew, A., editor, *Logic and Language (First Series)*, pages 39–53, Basil Blackwell, Oxford.

Henrich, J., 2016. *The Secret of Our Success*. Princeton University Press, Princeton.

Herbert, F., 1971. *Whipping Star*. Berkley Publishing, New York.

Herbert, F., 1977. *The Dosadi Experiment*. Berkley Publishing, New York.

Hieronymi, P., 2005. The wrong kind of reason. *The Journal of Philosophy*, *102(9)*, 437–457.

Hills, A., 2010. *The Beloved Self*. Oxford University Press, Oxford.

Hlobil, U., 2018. Choosing your nonmonotonic logic: A shopper's guide. In Arazim, P. and Lávička, T., editors, *The Logica Yearbook 2017*, pages 109–123, College Publications, London.

Holland, J., Holyoak, K., Nisbet, R., and Thagard, P., 1986. *Induction: Processes of Inference, Learning, and Discovery*. MIT Press, Cambridge, Mass.

Hölldobler, B. and Wilson, E. O., 2009. *The Superorganism*. W. W. Norton, New York.

Horty, J., 2012. *Reasons as Defaults*. Oxford University Press, Oxford.

Huebner, B., 2014. *Macrocognition*. Oxford University Press, Oxford.

Huizinga, J., 1996. *The Autumn of the Middle Ages*. University of Chicago Press, Chicago. Translated by Rodney Payton and Ulrich Mammitzsch.

Hurley, M., Dennett, D., and Adams, R., 2011. *Inside Jokes*. MIT Press, Cambridge.

Husserl, E., 2001. *Logical Investigations, vol. I*. Routledge, New York. trans. J. N. Findlay.

Ismael, J., 2007. *The Situated Self*. Oxford University Press, Oxford.

Ismael, J., 2016. *How Physics Makes Us Free*. Oxford University Press, Oxford.

Israel, J., 2001. *Radical Enlightenment*. Oxford University Press, Oxford.

Israel, J., 2006. *Enlightenment Contested*. Oxford University Press, Oxford.

Ivanhoe, P. J. and Van Norden, B., editors, 2005. *Readings in Classical Chinese Philosophy*. Hackett, Indianapolis, 2nd edition.

Jarry, A., 1996. *Exploits and Opinions of Dr. Faustroll, Pataphysician*. Exact Change, Boston. Translated by Simon Watson Taylor.

Johnston, M., 1987. Human beings. *Journal of Philosophy*, *84(2)*, 59–83.

Johnston, M., 2007. 'Human beings' revisited: My body is not an animal. *Oxford Studies in Metaphysics*, *3*, 33–74.

Kahneman, D., 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Keefe, R. and Smith, P., 1996. *Vagueness: A Reader*. MIT Press, Cambridge, Mass.

Kelly, D., 2011. *Yuck! The Nature and Moral Significance of Disgust*. MIT Press, Cambridge.

Kierkegaard, S., 1968. *Attack Upon 'Christendom'*. Princeton University Press, Princeton, 2nd edition. Translated by Walter Lowrie.

Klein, P. and Mozes, S. unpublished. Optimization algorithms for planar graphs. Draft manuscript.

Klein, S. and Nichols, S., 2012. Memory and the sense of personal identity. *Mind*, *121(483)*, 677–702.

Kolmogorov, A. N., 1956. *Foundations of the Theory of Probability*. Chelsea Publishing, New York, 2nd edition. Translated and edited by Nathan Morrison.

Korsgaard, C., 2009. *Self-Constitution*. Oxford University Press, New York.

Kripke, S., 2024. The question of logic. *Mind*, *133(529)*, 1–36.

Kuhn, T., 1970. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

Kusch, M., 1995. *Psychologism*. Routledge, New York.

Landesman, C., 1995. When to terminate a charitable trust? *Analysis*, *55(1)*, 12–13.

Latour, B., 1988. *Science in Action*. Harvard University Press, Cambridge.

Lawlor, K., 2003. Elusive reasons: A problem for first-person authority. *Philosophical Psychology*, *16(4)*, 549–564.

Lear, J., 1990. *Love and Its Place in Nature*. Farrar, Straus and Giroux, New York.

Lenman, J., 2000. Consequentialism and cluelessness. *Philosophy & Public Affairs*, *29(4)*, 342–370.

Lewis, D., 1983. Scorekeeping in a language game. In *Philosophical Papers, Volume I*, pages 233–249, Oxford University Press, Oxford.

Locke, J., 1979. *An Essay Concerning Human Understanding*. Oxford University Press, Oxford. Edited by Peter Nidditch.

Luce, R. D. and Raiffa, H., 1957. *Games and Decisions*. John Wiley and Sons, New York.

MacFarlane, J., 2000. *What Does it Mean to Say that Logic is Formal?* PhD thesis, University of Pittsburgh.

MacIntyre, A., 1981. *After Virtue*. University of Notre Dame Press, Notre Dame.

MacIntyre, A., 1990. *Three Rival Versions of Moral Enquiry*. Notre Dame University Press, Notre Dame.

Madell, G., 2015. *The Essence of the Self*. Routledge, New York.

Mangen, A., Walgermo, B., and Brønnick, K., 2013. Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, *58*, 61–68.

McDowell, J., 1998. *Mind, Value, and Reality*. Harvard University Press, Cambridge, Mass.

McGeer, V., 1996. Is 'self-knowledge' an empirical problem? Renegotiating the space of philosophical explanation. *Journal of Philosophy*, *93(10)*, 483–515.

McGeer, V., 2007. The moral development of first-person authority. *European Journal of Philosophy*, *16(1)*, 81–108.

Medawar, P., 1967. *The Art of the Soluble*. Methuen, London.

Meiland, J., 1980. What ought we to believe? or the ethics of belief revisited. *American Philosophical Quarterly*, *17(1)*, 15–24.

Mercier, H. and Sperber, D., 2019. *The Enigma of Reason*. Harvard University Press, Cambridge.

Mill, J. S., 1967–1991. *Collected Works of John Stuart Mill*. University of Toronto Press/Routledge and Kegan Paul, Toronto/London.

Miller, G., 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Millgram, E., 1991. Harman's hardness arguments. *Pacific Philosophical Quarterly*, *72(3)*, 181–202.

Millgram, E., 1997. *Practical Induction*. Harvard University Press, Cambridge, Mass.

Millgram, E., 2004. On being bored out of your mind. *Proceedings of the Aristotelian Society*, *104(2)*, 163–184.

Millgram, E., 2005. *Ethics Done Right: Practical Reasoning as a Foundation for Moral Theory*. Cambridge University Press, Cambridge.

Millgram, E., 2008. Specificationism. In Adler, J. and Rips, L., editors, *Reasoning*, pages 731–747, Cambridge University Press, Cambridge.

Millgram, E., 2009a. *Hard Truths*. Wiley-Blackwell, Oxford.

Millgram, E., 2009b. John Stuart Mill, determinism, and the problem of induction. *Australasian Journal of Philosophy*, *87(2)*, 181–197.

Millgram, E., 2009c. The persistence of moral skepticism and the limits of moral education. In Siegel, H., editor, *Oxford Handbook of Philosophy of Education*, pages 245–259, Oxford University Press, New York.

Millgram, E., 2010. Virtue for procrastinators. In Andreou, C. and White, M., editors, *The Thief of Time*, pages 151–164, Oxford University Press, New York.

Millgram, E., 2011. Critical notice: Christine Korsgaard, *Self-Constitution* and *The Constitution of Agency*. *Australasian Journal of Philosophy*, *89(3)*, 549–556.

Millgram, E., 2013. Relativism, coherence, and the problems of philosophy. In O'Rourke, F., editor, *What Happened in and to Moral Philosophy in the Twentieth Century?*, pages 392–422, Notre Dame University Press, Notre Dame.

Millgram, E., 2015. *The Great Endarkenment*. Oxford University Press, Oxford.

Millgram, E., 2023a. Practical nihilism. *Philosophies*, *8(2)*.

Millgram, E., 2023b. *Why Didn't Nietzsche Get His Act Together?* Oxford University Press, Oxford.

Millgram, E., 2025. Some uses of reflective equilibrium. In Efron, N. and Furstenberg, A., editors, *'The View from Within' from Without*, pages 27–44, Mohr Siebeck, Tübingen.

Millgram, E., forthcominga. From policy creep to pataphysics (in higher education). *Social Philosophy and Policy*, *()*, .

Millgram, E., forthcomingb. Hilary Putnam and the trajectory of classical pragmatism. In Marchetti, G., editor, *Interpreting Putnam*, pages 191–213, Routledge, London.

Millgram, E. and Thagard, P., 1996. Deliberative coherence. *Synthese*, *108(1)*, 63–88.

Millgram, H., 2008. *Four Biblical Heroines and the Case for Female Authorship*. McFarland, Jefferson, NC.

Moore, G. E., 1903/1960. *Principia Ethica*. Cambridge University Press, Cambridge.

Moran, R., 2001. *Authority and Estrangement*. Princeton University Press, Princeton.

Morton, A., 1990. *Disasters and Dilemmas*. Blackwell, Oxford.

Mosdell, M., 2012. *Representing Ourselves as Rational Agents*. PhD thesis, University of Utah.

Mosdell, M., 2015. When to think like an epistemicist. *Canadian Journal of Philosophy*, *45(4)*, 538–559.

Moulin, H., 1981. *Game Theory for the Social Sciences*. New York University Press, New York.

Mycielski, J., 1986. Locally finite theories. *Journal of Symbolic Logic*, *51(1)*, 59–62.

Nagel, T., 1978. *The Possibility of Altruism*. Princeton University Press, Princeton.

Nell, O., 1975. *Acting on Principle: An Essay on Kantian Ethics*. Columbia University Press, New York.

Newton, A., 2014. Kant on testimony and the communicability of empirical knowledge. *Philosophical Topics*, *42(1)*, 271–290.

Nguyen, C. T., 2019. Games and the art of agency. *Philosophical Review*, *128(4)*, 423–462.

Nguyen, C. T., 2020. *Games: Agency as Art*. Oxford University Press, New York.

Nichols, S. and Stich, S., 2003. *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Clarendon Press, Oxford.

Nietzsche, F., 2000. *Basic Writings of Nietzsche*. Modern Library/Random House, New York. Edited and translated by Walter Kaufmann; introduction by Peter Gay.

Nozick, R., 1981. *Philosophical Explanations*. Harvard University Press, Cambridge, MA.

Nozick, R., 1997. *Socratic Puzzles*. Harvard University Press, Cambridge.

Nozick, R., 2001. *Invariances*. Harvard University Press, Cambridge, Mass.

Nussbaum, M., 2001. The *Protagoras*: A science of practical reasoning. In Millgram, E., editor, *Varieties of Practical Reasoning*, MIT Press, Cambridge, Mass.

Oaksford, M. and Chater, N., 1998. *Rationality in an Uncertain World*. Routledge, London.

Olson, E., 1997. *The Human Animal*. Oxford University Press, Oxford.

O'Neill, O., 1989. Consistency in action. In *Constructions of Reason*, pages 81–104, Cambridge University Press, Cambridge. Reprinted in E. Millgram, ed., *Varieties of Practical Reasoning*.

O'Neill, O., 2002. *A Question of Trust*. Cambridge University Press, Cambridge.

Oreskes, N. and Conway, E., 2010. *Merchants of Doubt*. Bloomsbury Press, New York.

Orwell, G., 1950. *1984*. Signet Classics, New York.

Parfit, D., 1984. *Reasons and Persons*. Oxford University Press, Oxford.

Pearl, J., 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Politis, V., 2004. *Aristotle and the* Metaphysics. Routledge, London.

Pollock, J., 1987. Defeasible reasoning. *The Philosopher's Annual*, *10*, 139–176.

Popper, K., 1977. *The Logic of Scientific Discovery*. Hutchinson, London.

Pradeu, T., 2012. *The Limits of the Self: Immunology and Biological Identity*. Oxford University Press, New York.

Proust, M., 1988. Sainte-Beuve and Baudelaire. In *Against Sainte-Beuve and Other Essays*, pages 34–55, Penguin, New York. Translated by John Sturrock.

Putnam, H., 1975a. The meaning of 'meaning'. In *Mind, Language and Reality*, pages 215–271, Cambridge University Press, Cambridge.

Putnam, H., 1975b. The mental life of some machines. In *Mind, Language and Reality*, pages 408–428, Cambridge University Press, Cambridge.

Putnam, H., 1975c. Minds and machines. In *Mind, Language and Reality*, pages 362–385, Cambridge University Press, Cambridge.

Quine, W. V., 1936. Truth by convention. In Lee, O. H., editor, *Philosophical Essays for A. N. Whitehead*, pages 90–124, Longmans, Green and Co., New York.

Quine, W. V., 1960. Carnap and logical truth. *Synthese*, *12*, 350–374.

Quine, W. V. and Carnap, R., 1990. *Dear Carnap, Dear Van*. University of California Press, Berkeley. Edited by Richard Creath.

Quinn, W., 1993. *Morality and Action*. Cambridge University Press, Cambridge.

Rawls, J., 1955. Two concepts of rules. *Philosophical Review*, *64(1)*, 3–32.

Reutlinger, A., Schurz, G., and Hüttemann, A., 2017. Ceteris paribus laws. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University.

Richardson Lear, G., 2004. *Happy Lives and the Highest Good*. Princeton University Press, Princeton.

Richerson, P. and Boyd, R., 2004. *Not By Genes Alone: How Culture Transformed Human Evolution*. Chicago University Press, Chicago.

Ringer, F., 1990. *The Decline of the German Mandarins*. Wesleyan University Press, Hanover.

Rödl, S., 2007. *Self-Consciousness*. Harvard University Press, Cambridge.

Ross, L., 1977. The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology*, pages 173–220, Elsevier.

Roush, S., 2003. Copernicus, Kant, and the anthropic cosmological principles. *Studies in History and Philosophy of Modern Physics*, *34*, 5–35.

Rovane, C., 1990. Branching self-consciousness. *Philosophical Review*, *99(3)*, 355–395.

Rupert, R., 2009. *Cognitive Systems and the Extended Mind*. Oxford University Press, Oxford.

Ryle, G., 1966. *Plato's Progress*. Cambridge University Press, Cambridge.

Ryle, G., 1967. *The Concept of Mind*. Barnes and Noble, New York.

Safranski, R., 2001. *Ein Meister aus Deutschland: Heidegger und Seine Zeit*. Fischer Taschenbuch Verlag, Frankfurt a.M.

Schechtman, M., 2017. *Staying Alive*. Oxford University Press, Oxford.

Schwenkler, J., 2011. Perception and practical knowledge. *Philosophical Explorations*, *14(2)*.

Schwenkler, J., 2015. Understanding 'practical knowledge'. *Philosopher's Imprint*, *15(15)*.

Searle, J., 2005. Desire, deliberation and action. In Vanderveken, D., editor, *Logic, Thought and Action*, pages 49–78, Springer, Dordrecht.

Shapiro, S., 2006. *Vagueness in Context*. Clarendon Press, Oxford.

Shoemaker, S., 1963. *Self-Knowledge and Self-Identity*. Cornell University Press, Ithaca, New York.

Shoemaker, S., 1984. Personal identity: A materialist's account. In Shoemaker, S. and Swinburne, R., editors, *Personal Identity*, pages 69–132, Basil Blackwell, Oxford.

Shoemaker, S., 1996. *The First-Person Perspective and Other Essays*. Cambridge University Press, Cambridge.

Shubik, M., 1985. *Game Theory in the Social Sciences*. MIT Press, Cambridge.

Simon, H., 1957. *Models of Man*. John Wiley and Sons, New York.

Simon, H., 1967. Motivational and emotional controls of cognition. *Psychological Review*, *74(1)*, 29–39.

Sinclair, U., 1935. *I, Candidate for Governor and How I Got Licked*. Self-published, Pasedena.

Small, W., 2012. Practical knowledge and the structure of action. In Abel, G. and Conant, J., editors, *Rethinking Epistemology, vol. 2*, pages 133–228, de Gruyter, Berlin.

Smith, A., 1976. *A Theory of the Moral Sentiments*. LibertyClassics, Indianapolis.

Smith, M., 1987. The Humean theory of motivation. *Mind*, *96(381)*, 36–61.

Sorensen, R. A., 2001. *Vagueness and Contradiction*. Oxford University Press, Oxford.

Sperber, D., 1985. Anthropology and psychology: Towards an epidemiology of representations. *Man (N.S.)*, *20*, 73–89.

Sterelny, K., 2004. Externalism, epistemic artefacts and the extended mind. In Schantz, R., editor, *The Externalist Challenge*, pages 239–254, de Gruyter, Berlin.

Sterelny, K., 2010. Mind: Extended or scaffolded? *Phenomenology and the Cognitive Sciences*, *9(4)*, 465–481.

Strawson, G., 2009. *Selves*. Oxford University Press, Oxford.

Strawson, P. F., 1971. *Individuals*. Methuen, London.

Sunstein, C. R., 2023. *Decisions About Decisions: Practical Reason in Ordinary Life*. Cambridge University Press,, Cambridge.

Thagard, P., 1985. Computational tractability and conceptual coherence: Why do computer scientists believe that $P \neq NP$? *Canadian Journal of Philosophy*, *82(7)*, 335–349.

Thompson, M., 2006. What is it to wrong someone? A puzzle about justice. In Wallace, R. J., Pettit, P., Scheffler, S., and Smith, M., editors, *Reason and Value*, pages 333–384, Oxford University Press, Oxford.

Thompson, M., 2008. *Life and Action*. Harvard University Press, Cambridge.

Tooby, J. and Cosmides, L., 1992. The psychological foundations of culture. In Barkow, J., Cosmides, L., and Tooby, J., editors, *The Adapted Mind*, pages 19–136, Oxford University Press, Oxford.

Townsend, R., 1990. *Financial Structure and Economic Organization*. Basil Blackwell, Cambridge, Mass.

Vardar, C., 2024. *Zitelosa of Benevolent Gravity: Uncontrollable Inertial Actions*. Master's thesis, University of Utah.

Velleman, J. D., 1989. *Practical Reflection*. Princeton University Press, Princeton.

Velleman, J. D., 2000. *The Possibility of Practical Reason*. Oxford University Press, Oxford, 1st edition.

Vlastos, G., 1994. The Socratic elenchus: Method is all. In Burnyeat, M., editor, *Socratic Studies*, pages 1–37, Cambridge University Press, Cambridge.

Vogler, C., 1998. Sex and talk. *Critical Inquiry*, *24*, 328–365.

Vogler, C., 2002. *Reasonably Vicious*. Harvard University Press, Cambridge, Mass.

von Mises, R., 1957. *Probability, Statistics and Truth*. Macmillan, New York, 2nd revised edition. Prepared by Hilda Geiringer.

von Neumann, J. and Morgenstern, O., 1944. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton.

Weber, M., 2004. Politics as a vocation. In Owen, D. and Strong, T., editors, *The Vocation Lectures*, pages 32–94, Hackett, Indianapolis. Translated by Rodney Livingstone.

Whewell, W., 1847. *The Philosophy of the Inductive Sciences*. John W. Parker, London.

Wickens, C. and Hollands, J., 2000. *Engineering Psychology and Human Performance, third edition*. Prentice-Hall, Upper Saddle River, NJ.

Wiggins, D., 1980. *Sameness and Substance*. Harvard University Press, Cambridge.

Wiggins, D., 1991. A sensible subjectivism? In *Needs, Values, Truth (second edition)*, pages 185–214, Blackwell, Oxford.

Wilkes, K., 1993. *Real People*. Oxford University Press, Oxford.

Williams, B., 1973a. Personal identity and individuation. In *Problems of the Self*, pages 1–18, Cambridge University Press, Cambridge.

Williams, B., 1973b. *Problems of the Self*. Cambridge University Press, Cambridge.

Williams, B., 1973c. The self and the future. In *Problems of the Self*, pages 46–63, Cambridge University Press, Cambridge.

Williams, B., 1978. *Descartes: The Project of Pure Enquiry*. Penguin, Harmondsworth.

Williams, B., 1981. *Moral Luck*. Cambridge University Press, Cambridge.

Williams, B., 2002. *Truth and Truthfulness*. Princeton University Press, Princeton.

Williamson, T., 1994. *Vagueness*. Routledge, New York.

Williamson, T., 2006. Must do better. In Greenough, P. and Lynch, M., editors, *Truth and Realism*, pages 177–187, Oxford University Press, Oxford.

Wilson, M., 2006. *Wandering Significance*. Oxford University Press, Oxford.

Wilson, M., 2017. *Physics Avoidance*. Oxford University Press, Oxford.

Wimsatt, W., 2007. *Re-Engineering Philosophy for Limited Beings*. Harvard University Press, Cambridge.

Winograd, T., 1972. *Understanding Natural Language*. Edinburgh University Press, Edinburgh.

Wootton, D., 2007. *Bad Medicine*. Oxford University Press, Oxford.