Knowing by Example: A Social-Cognitive Approach to Epistemology

by

Jeremy A. B. Lent

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in the University of Michigan
2016

Doctoral Committee:

Professor Elizabeth S. Anderson, Co-Chair
Professor James M. Joyce, Co-Chair
Associate Professor Maria Lasonen-Aarnio
Professor Richard L. Lewis
Professor Peter A. Railton

TABLE OF CONTENTS

CHAPTER

CHAPTER I: Mysterious Knowledge

1.1 A Puzzle

The verb “know” and its cognates are among the most commonly used words in the English language: According to the Corpus of Contemporary American English, “know” is the 49th most common English word, the 10th most common verb, and the most common mental-state verb (Davies and Gardner 2010). Yet some of our “know”-related linguistic habits are quite puzzling—in particular, our habits surrounding “know-that” constructions, such as “I know that ‘know’ is the 10th most common English verb.”[1] A particularly striking set of puzzles became widely recognized—at least among philosophers and linguists—in the wake of Edmund Gettier’s 1963 paper, “Is Justified True Belief Knowledge?” Until that point, not many philosophers or linguists had put much effort into charting how we use the word “know” in everyday life, let alone developing a general theory about those uses. Gettier took as his foil the theory that when we think about knowledge, we think of something roughly like “a true belief for which the believer has sufficient justification”.[2] (By calling a belief “justified”, Gettier and

[1] “Know-that” is to be distinguished from “know-how” (e.g., “I know how to search the Corpus of Contemporary American English (COCA)”) and “know-of” (e.g., “I know of the COCA,” or “I know the editors of the COCA”). Except where otherwise noted, when I use “know” and its cognates, I’ll be referring strictly to “know-that”.

[2] Despite the lore among epistemologists, it’s not clear how many pre-Gettier philosophers actually endorsed such a theory. For instance, Russell (1948: pp. 170-71) shows clear signs of disagreeing with the “justified true belief” analysis.

other philosophers seemed to imply their belief that they would have held the very same belief, had they been in the believer's shoes.)

To demonstrate that knowledge is *not* equivalent to justified true belief, Gettier posed two now-infamous hypothetical scenarios. In each of them, a believer has a clearly well-justified true belief, yet many of us (Gettier included) are strongly inclined to deny that this believer *knows* the fact in question. Here, for the sake of tradition, are Gettier's two cases:

> Smith and Jones have each applied for a certain job. Smith has strong evidence that Jones will get the job and that Jones has ten coins in his pocket. (Suppose that the president of the company assured Smith that Jones would be selected, and that Smith recently counted the coins in Jones's pocket.) Smith infers the following proposition from the latter evidence: "The person who will get the job has ten coins in his pocket." But unknown to Smith, he and not Jones will get the job. Also, unknown to Smith, he himself has ten coins in his pocket. (ibid., p. 122)
>
> Smith has strong evidence that Jones owns a Ford: Jones has at all times in the past within Smith's memory owned a car, and always a Ford. Smith has another friend, Brown, of whose whereabouts he is totally ignorant. Smith thinks of the city Barcelona quite at random, and constructs the following proposition: "Either Jones owns a Ford, or Brown is in Barcelona." Smith realizes the entailment of this proposition from his belief that Jones owns a Ford, and he proceeds to accept it on the basis of this entailment. But in fact, Jones does *not* own a Ford: He is merely renting one right now. And as it happens, Brown is in fact in Barcelona. (ibid., pp. 122-23)

Epistemologists promptly busied themselves trying to give an account of knowledge—that is, a set of necessary and sufficient conditions for a belief to count as knowledge—which excludes both of Smith's (justified, true) beliefs but which doesn't exclude any beliefs (real or hypothetical) which we intuitively *do* regard as knowledge. (So, for instance, the account shouldn't deny that we can know what our own names are, or that the derivative of sin(x) is cos(x), or that Ann Arbor is located in the Eastern Time Zone.)

Many accounts were proposed which did apparently exclude each of Smith's beliefs from counting as knowledge. But for each one, some philosopher devised a hypothetical counterexample, involving either (1) a belief which satisfies all the account's conditions but which doesn't seem to be knowledge, or (2) a belief which fails one or more of the conditions but which *does* seem to be knowledge. In the course of this "Gettier game," dozens of hypothetical cases were constructed involving a true belief which is quite plausibly justified, but which many of us would hesitate to label as knowledge. And today, more than 50 years since Gettier's paper, there is no account of knowledge which has garnered anything near a consensus among philosophers. In fact, some have abandoned the project of giving non-circular necessary and sufficient conditions as hopeless. For instance, Williamson (2000) proposes that our concept *knowledge* corresponds to a mental state whose exact nature cannot be expressed in any more-basic terms (at least not of our language). Meanwhile, we're left with a mountain of often-fanciful hypothetical cases where our willingness (or not) to ascribe knowledge to a particular believer doesn't seem to fit any detectable, exceptionless pattern.

1.2 A Hypothesis

This state of affairs raises quite a puzzle: Why do we humans use the words "know" and "knowledge" in the exact ways that we do? Why, for instance, does it seem wrong to ascribe knowledge to the protagonists in Gettier's cases and others like them?[3] If some account proposed in Gettier's wake had succeeded in capturing all of our judgments about knowledge, then perhaps we'd have a good lead on answering the puzzle: We could take the conditions for knowledge in the successful account and, if it weren't already obvious, try to piece together why humans might have developed a word sensitive to those conditions. However, no account

[3] Several recent empirical studies (e.g. Starmans & Friedman 2012) have provided evidence that the apparent philosophical consensus on Gettier's cases, and others like them, is not shared by a significant proportion of non-philosophers. In Chapter 5, I'll return to these apparent differences in knowledge ascription patterns between philosophers and non-philosophers, and propose an explanation for them.

seems to have succeeded in capturing all of our judgments (or at least those of philosophers) about knowledge. Thus, even if we can determine why humans might have developed a word sensitive to the conditions of an existing account, it will still be a puzzle why we do (or don't) ascribe knowledge in the cases which the account doesn't account for.

To illustrate: Suppose, counterfactually, that our concept of knowledge was fully captured by the "justified true belief" account. Then we'd still want to ask why humans developed a word specifically to alert each other to the presence of *justified* true belief, rather than just true belief. Presumably, it's important to recognize when other people have true beliefs: for purposes of social coordination, for predicting others' actions (and the consequences of them), etc. But what's the added value in recognizing that (roughly) we would have believed the same way someone else did had we been in their shoes? What are we thereby at least potentially alerting each other to? Perhaps the reader can begin to see a possible explanation here. But now remember that the "justified true belief" account *doesn't* fully capture our knowledge ascription practices. Thus, even if we could determine why humans might have developed a word synonymous with justified true belief, that wouldn't answer the puzzle we actually have before us.

In fact, I believe I've found a way to explain why our word "know" functions in all the apparently puzzling ways it does. In essence, my hypothesis is this: Our evolutionary ancestors developed a word which, through linguistic or "memetic" evolution, came to be understood (at least implicitly) by speakers and listeners to convey at least three messages: (1) a given person K is certain of the truth of a particular proposition *p* (e.g., "The new berries we found are safe to eat"); (2) the speaker is certain of the proposition's truth too; and (3) the speaker would advise anyone in a position *similar* to K's to do the same sort of thing K did in being certain that *p* (roughly, to "follow K's example").[4] My proposal is that our modern-day verb "know" is the descendant of this ancestral word, and that an understanding of the messages it conveys has been transmitted, at least implicitly, through each generation. I'll spell out the details of this hypothesis in (hopefully not too excruciating) detail in Chapter 2. (For now, I offer this brief statement of the hypothesis as a compass for my reader, so he or she can have a sense of where we're headed.) After laying out the hypothesis, I'll spend the rest of Chapter 2 showing how it can account for several potentially puzzling features of our knowledge ascription practices—that is, our uses of the word "know"—among both philosophers and non-philosophers.

In Chapters 3 and 4, I'll demonstrate in unabashedly excruciating detail how my hypothesis can explain the apparent verdict of the contemporary philosophical community on every sort of hypothetical case of belief—including Gettier's cases—which has been discussed in the epistemology literature.[5] (I'll discuss three types of "community verdict": a unanimous ascription of knowledge, a unanimous denial of knowledge, and a split or ambivalent verdict.) More specifically, I'll use the idea that speakers today at least implicitly understand the phrase "K knows that *p*" to convey messages (1)-(3) (stated above) in order to account for the philosophical community's verdict on a very wide range of cases in the literature. And I'll propose that on this basis, we should be confident that my hypothesis can account for the

[4] Throughout, I'll use 'K' to designate a generic candidate 'K'nower—that is, someone (real or hypothetical) for whom we consider whether or not a particular belief of theirs counts as knowledge.

[5] I believe the excruciating detail is necessary, in order to fully convince myself and others that my hypothesis really does account for *all* of the philosophical community's verdicts on hypothetical cases. The hypothetical cases proposed in the literature since Gettier's paper are many and varied, so it won't do to show that the hypothesis can explain the verdicts on just a few cases and then extrapolate to the rest.

community verdict on all existing cases which I don't have time or space (or patience) to consider individually, as well as any new case proposed in the future.

Then, in Chapter 5, I'll go on to show how the evolutionary hypothesis, along with several plausible auxiliary hypotheses, can account for a few puzzling patterns of philosophers' knowledge intuitions not covered in Chapters 3 and 4, as well as some puzzling results of several recent experiments conducted by epistemologists to test the reactions of non-philosophers to the sorts of hypothetical cases discussed in the epistemology literature.[6] Finding that my hypothesis offers plausible explanations for such a huge array of observations, and that none of its predictions seems to be contradicted by existing data, I'll propose in Chapter 6 that we at least provisionally take ourselves to have solved the puzzle of why we use the word "know" in all the exact ways we do. That is, we should provisionally accept my hypothesis as providing the truth about how our knowledge ascription practices came to be what they are. (I say "provisionally" because I can't *guarantee* that there won't be new observations which my hypothesis can't account for, or which contradict its predictions—or that there won't be *better* explanations devised in the future.) Since at least some epistemologists have paid some attention to the puzzle of *explaining* our knowledge ascription practices—and not simply capturing them in an abstract account—those epistemologists can cross this puzzle off their "to-know" lists. Just to provide some disciplinary context: It appears that at least initially after Gettier published his paper, the motivation of most or all

[6] Following the lead of other epistemologists, I use the term "knowledge intuitions" to describe the judgments that philosophers and others immediately have—seemingly without conscious critical reflection—about whether or not the protagonist of a given hypothetical case has knowledge as opposed to "mere" true belief. I include these intuitions within the scope of our knowledge ascription practices.

epistemologists was simply to find an exceptionless pattern in our knowledge judgments, and not necessarily to explain why we have the word "know"—in all its puzzling glory—in the first place. And the thrust of more recent work has been towards examining the nature of knowledge as a metaphysical *entity*—not just a concept—and exploring its relevance to various normative facts such as what it's appropriate for a given person to do or assert.[7]

1.3 Previous Efforts

Why the need for a wholly new hypothesis to account for our knowledge ascription practices? What has previous work left to be desired? As I indicated above, no extant account of the conditions for the instantiation of knowledge (that is, for a belief to count as being of the kind "knowledge") has aligned with the apparent community verdict among philosophers on *all* cases posed in the literature. For instance, after Gettier roundly demonstrated the inadequacy of the "justified true belief" account, Goldman (1967) proposed the following "causal" account of knowledge:

> S knows that *p* if and only if the fact *p* is causally connected in an "appropriate" way with S's believing *p*. "Appropriate" knowledge-producing causal processes include the following: (1) perception [that *p*]; (2) memory [that *p*]; (3) a causal chain [which includes the fact that *p* and] which is correctly reconstructed by inferences, each of which is warranted …; (4) combinations of (1), (2) and (3). (pp. 369-70)[8]

Goldman's theory appears to capture the philosophical verdict on Gettier's first case (p. 2 above) (and, by similar reasoning, the second case as well) as follows: Smith's belief that the person who will get the job has ten coins in his pocket is *not* caused by the *fact* that the person

[7] Williamson (2000) seems to be the progenitor of this more recent line of epistemological inquiry.

[8] Goldman offers this as an account of *empirical* knowledge—that is, knowledge of contingent propositions such as "Edmund Gettier published a short but influential paper in 1963", as opposed to "necessary truths" such as Fermat's last theorem.

who will get the job has ten coins in his pocket. After all, the latter fact is the fact of *Smith*, not Jones, having ten coins in his pocket—and Smith's belief is not caused by this fact, but rather by the fact of *Jones* having ten coins in his pocket. However, Goldman's account was quickly rebutted by Collier (1973), who offered the following counterexample:

> Suppose that unbeknown to Smith I administer an hallucinogenic drug to him. Since he doesn't realize that he has been drugged, he believes that his hallucinations are real. But one of the hallucinations is that I gave him the drug that I, in fact, gave him, and in particular, he believes that his hallucination is real. (p. 350)

Collier, apparently with the agreement of every other philosopher (including Goldman), takes it that Smith doesn't *know* that he's been drugged by the protagonist. Yet Goldman's theory seems to return the verdict that Smith *does* know, since Smith's belief is caused by the fact of his being drugged, and Smith seems to have "warrant"—given the verisimilitude of his hallucinations—for believing that this causal chain occurred.

Epistemologists promptly developed further counterexamples to Goldman's account, and also promptly developed new accounts which seemed to withstand the counterexamples—only for those new accounts to be convincingly "counterexampled" themselves. I won't try my reader's patience (or my own) with a full recounting of this intellectual tennis match; suffice it to say that no account appeared to align with *all* verdicts of the philosophical community. Yet even if some future account does appear to capture the philosophical verdict on every hypothetical case in the literature, it still might not be clear *why* we humans developed a word to refer to the exact conditions given by the account. (This is the puzzle I mentioned above.) If the successful account looks anything like Goldman's (which is fairly representative of other proposed accounts in its appeal to such metaphysical notions as causal chains), it won't be a trivial matter to see why we might have developed a word for *that*.

For instance, suppose counterfactually that Goldman's causal theory really did track the philosophical verdict on all real and hypothetical cases of belief. Then we'd still be left wondering about the process by which our linguistic ancestors came to have a word to alert each other to (roughly) whether or not someone's belief that *p* was caused by the fact that *p*. Why not just have a word for "true belief" and leave it at that? At least speaking for myself, I find it hard to believe that there is no connection—even if just an oblique one—between, on the one hand, philosophers' present-day denial of knowledge in Gettier's cases, and on the other, some of the recognizable needs, values, and/or goals of our species. That is, it's hard to believe that there isn't some relevant "upshot" to knowledge ascriptions and denials, even in Gettier's cases. What sort of valuable information are we conveying when we deny that the protagonists in Gettier's cases have knowledge?

One philosopher who explicitly tried to address this question was Lewis (1996). His "contextualist" account of knowledge departed significantly from previous proposals by entailing that the truth conditions for the statement "K knows that *p*" can change based on the conversational context in which it's uttered. Lewis's account is as follows: "K knows that *p* iff K's evidence eliminates every possibility in which ~*p*—Psst!—except for those possibilities that we are properly ignoring" (p. 554). K's evidence (that is, her perceptual experiences and memories) "eliminates" possibility W just in case if W were in fact the actual situation, K's evidence would be different than it actually is (p. 553). So, for instance, my current evidence—my visual perceptions as of being in Ann Arbor—eliminate the possibility that I'm currently at the beach and having visual perceptions as of being at the beach. In general, for nearly any proposition '*p*', no one's evidence will eliminate *all* possibilities in which ~*p*: For instance, my current evidence doesn't eliminate the possibility that I'm at the beach but being fed visual perceptions as of being in Ann Arbor by some evil neuroscientist who's hooked up my brain to a supercomputer. (Alas, epistemologists do at times consider such possibilities.)

However, that doesn't yet mean that it's never true to assert "K knows that *p*", since we may be in a context in which we're properly ignoring certain—and even many—uneliminated possibilities. Lewis proposes a series of rules about what possibilities are and are not properly ignored in any given context. The three most crucial rules, for our purposes, are as follows:

> *Rule of Actuality*: The possibility that actually obtains is never properly ignored. (p. 554)
>
> *Rule of Resemblance*: Suppose one possibility saliently resembles another. Then if one of them may not be properly ignored, neither may the other. (p. 556)
>
> *Rule of Attention*: [A] possibility not ignored at all is *ipso facto* not properly ignored. (p. 559)

Lewis uses the rule of attention to explain why, in certain unusual contexts, many of us tend to agree that we or others *don't* know certain things which we typically do take ourselves to know. For instance, I do feel some pressure, upon considering the possibility that I'm currently at the beach but am also at the whims of an evil neuroscientist, to say I don't absolutely *know* that I'm currently in Ann Arbor. (As we'll see in Chapter 5, many people—philosophers and non-philosophers alike—appear to have similar hesitation upon considering similarly "skeptical" possibilities.) Lewis explains my hesitation to claim knowledge as follows: Since I'm now considering an uneliminated possibility in which I'm not in Ann Arbor, by the Rule of Attention that possibility is not properly ignored (because not ignored), so it's true that I don't currently know I'm in Ann Arbor. However, in more ordinary contexts in which evil neuroscientist possibilities aren't being mooted (that is, in most every context other than some involving epistemologists), those possibilities are properly ignored and so don't threaten our ordinary claims to knowledge.

Lewis uses the Rules of Actuality and Resemblance to explain the philosophical consensus that the protagonists in cases like Gettier's don't have knowledge. For instance, take Gettier's first case: There's arguably a possibility closely resembling the actual one in which Smith will indeed get the job (rather than Jones), but where Smith has only (say) nine coins in his pocket. This possibility is not eliminated by Smith's evidence (which consists of his recollection of hearing that Jones will get the job, and of counting ten coins in Jones's pocket), and since it saliently resembles the actual scenario, it is not properly ignored (that is, by those of us judging whether or not Smith has knowledge). Since this not-properly-ignored possibility is one in which the proposition in question ("The man who will get the job has ten coins in his pocket") is false, it's true that Smith doesn't know it in the actual scenario.

Lewis's contextualist account of knowledge does seem to show promise of agreeing with the philosophical verdict on many or all hypothetical cases in the literature. It hasn't attracted a general consensus among philosophers mostly because of its entailment that the truth conditions of knowledge possession can vary by conversational context—whereas many epistemologists find that to be an unacceptable result, and insist that the truth conditions are invariant across all contexts. (See, for instance, Hawthorne 2004, Stanley 2005, Nagel 2010, and Weatherson 2011.) However, my concern here is not with the "contextualism vs. invarantism" debate, but rather with the question, Why have a word which abides by (say) the Rule of Resemblance? It's not immediately clear why, given our species' needs and goals, our linguistic ancestors would have developed a word sensitive to the fact that there are similar possibilities to Smith's actual situation, uneliminated by his evidence, in which the proposition he believes is false. Why not just have a word for "true belief" and leave it at that?

Lewis, as I mentioned earlier, does briefly address this concern: He suggests that if his account is correct, our word "know" is a way of conveying to ourselves and others that our evidence has eliminated all the possibilities—or rather, those that we can't properly ignore—in which what we believe is false, and hence that we "can't" be wrong. Lewis concedes that this is a "very sloppy way of conveying very incomplete information about the elimination of possibilities", since there will virtually always be some possibilities we're (properly) ignoring. (That is, there's always another sense in which we very well *might* be wrong.) Is there a reason why it's at least sometimes helpful to sloppily convey this incomplete information? Perhaps saying "I know that *p*", if Lewis's account is correct, communicates to others that we're not likely to change our mind about whether or not *p*, since our evidence is inconsistent with every possibility (within our purview) in which $\sim p$. It's not just that we feel *certain* that *p*—it's also that we don't currently see any way in which it could be the case that $\sim p$. (Lewis doesn't explicitly draw the connection between his account and an indication of "I'm not likely to change my mind", but I think he'd find it plausible.)

Now what about third-person uses of "know", such as "Smith doesn't know that the man who will get the job has ten coins in his pocket"? Perhaps, even though Smith himself isn't aware that there's an uneliminated possibility which is similar to his actual situation and in which the proposition he believes is false, it's helpful to communicate this information about Smith to others to warn them that he might soon *lose confidence* in the proposition. After all, he might soon discover that he, not Jones, will get the job, and he would thereby lose confidence about the number of coins in the job-getter's pocket (at least momentarily). However, this explanation for the utility of third-person knowledge ascriptions and denials (if they do indeed hew to Lewis's account) won't work in general, since there will clearly be times when the third party isn't at any risk of losing confidence despite being in a situation like Smith's. For instance, suppose that although Smith forms his belief on the basis of counting the coins in Jones's pocket, he also has a strange belief that if he should somehow end up getting the job instead of Jones, he'll suddenly and magically have the exact same number of coins in his pocket as Jones does (if the numbers don't already match, that is). I'm quite sure all philosophers will deny, as in the original case, that Superstitious Smith knows that the man who will get the job has ten coins in his pocket. But now the explanation we're considering for the utility of Lewisian knowledge ascriptions and denials doesn't explain why philosophers would deny that Superstitious Smith has knowledge, since he *isn't* at all likely to lose confidence in his belief. Thus, there's clearly more work to be done in explaining our knowledge ascription patterns, even if Lewis's account turns out to align with the philosophical verdict on all cases. [9]

However, I do appreciate Lewis's attention to the question of what our word "know" might do for us—a question which very few epistemologists prior to Lewis explicitly considered. One who did was Craig (1990), who in his book *Knowledge and the State of Nature* departed from the traditional methodology of proposing necessary and sufficient conditions for the instantiation of the property **knowledge**. Instead of treating our concept of knowledge as an

[9] As we'll see later, my own hypothesis does offer an explanation for why our word "know" is apparently always sensitive to whether there are (in Lewis's terms) saliently similar possibilities uneliminated by a believer's evidence in which ~*p*. However, my point here is that the reason isn't obvious: We still have work to do to satisfactorily explain our knowledge ascription practices.

ahistorical given, Craig suggests that our knowledge ascription practices have become what they are at least partly due to the historical exigencies of our species' social environment: We rely on other people for most of our beliefs about the world, and we need to coordinate with them to carry out any moderately complex inquiry. With this plausible idea as motivation, Craig proposes a hypothetical "state of nature" story to account for some facets of our knowledge ascription practices. In particular, he suggests that our ancestors would have had good reason to develop a word to indicate whether the speaker thought that some other person was a "reliable informant" about a certain matter. Suppose this word was the ancestor of our word "know". Then the construction "K knows whether or not *p*" (or rather its equivalent in our ancestors' language) would have been appropriate to assert if and only if the speaker thought it highly likely (at least, likely enough for the present practical purposes of her listeners) that K had a true belief about whether or not *p* (ibid., p. 85). Craig implies that our modern-day uses of the phrase "K knows whether or not *p*" at least implicitly follow something like this same norm—and if so, perhaps we can understand why we use the construction "knows-whether" in the ways that we do.

I deeply appreciate Craig's attempt to draw a clear connection between some of our uses of "know" and some of our recognizable human needs and goals. (Indeed, we do very often have a need to determine who is a good person to ask about a particular matter.[10]) And as will soon become clear, I've taken inspiration from his "evolutionary" approach to the mystery of our knowledge ascription practices. (My hypothesis, like his, begins with some plausible suggestions about the needs and goals of our early human ancestors, as well as the

[10] Or, at least, our pre-Google ancestors did.

environmental pressures they faced.) Alas, his work doesn't offer to explain *all* of our uses of the word "know", since we use "know-that" constructions—"K knows that *p*"—as well as "know-whether" (in fact, much more often than the latter, it seems). Craig does not attempt to expand his evolutionary hypothesis to account for our uses of "know-that". And it's not immediately clear how to do so: If we suppose that "K knows that *p*" conveys simply that K is a reliable informant about whether or not *p*, then we'll assert "K knows that *p*" whenever we ourselves believe that *p* and that K believes that *p*. Yet as we saw above in Gettier's two cases, there are clearly times when at least many of us *refuse* to assert "K knows that *p*" even when it's clear that *p* and that K believes it. Thus, we still don't have a full account of our "know-that" practices.

Perhaps the most promising extant attempt to build on Craig's proposal, and to extend it convincingly to "know-that", is Pritchard's (2010) "anti-luck virtue" account of knowledge, as follows:

> Knowledge is safe belief that arises out of the reliable cognitive traits that make up one's cognitive character, such that one's cognitive success is to a significant degree creditable to one's cognitive character. (p. 54)

Elsewhere, Pritchard defines "safe belief" as follows: For all agents and all propositions ϕ, an agent's belief that ϕ is "safe" if and only if:

> … in nearly all (if not all) nearby possible worlds in which she forms the belief about ϕ in the same way as she forms her belief in the actual world, that agent only believes that ϕ when ϕ is true. (Pritchard 2005, p. 163)

(Pritchard defines "possible worlds" as fully-specified sets of circumstances, where "'distant' possible worlds are very unlike the actual world, whilst 'nearby' possible worlds are very alike the actual world" (p. 128).)[11] I'll explain the connection Pritchard draws to Craig's ideas in a moment. But first, let's note how Pritchard's account fares on the prerequisite task of agreeing with the community philosophical verdict on all cases in the literature. The anti-luck virtue account rules out Smith's "ten coins" belief as being knowledge due to the "safety" requirement: It seems that there are very similar alternative worlds in which Smith has only (say) nine coins in his pocket, yet he still forms the belief "The man who will get the job has ten coins in his pocket" by counting the coins in Jones's pocket, thereby ending up with a false belief.

The "virtue" requirement is meant to handle such cases as the following one offered by Pritchard:

> Imagine that our agent—let's call him 'Temp'—forms his beliefs about the temperature in his room by consulting a thermometer on the wall. Unbeknownst to Temp, however, the thermometer is broken and is fluctuating randomly within a given range. Nonetheless, Temp never forms a false belief about the temperature by consulting this thermometer since there is a person hidden in the room, next to the thermometer, whose job it is to ensure that whenever Temp consults the thermometer the temperature in the room corresponds to the reading on the thermometer. (Pritchard 2010, p. 49)

Here, Temp's beliefs about the temperature are stipulated to be "safe", because there are seemingly no similar alternative scenarios where he forms a false belief—thanks to the person ensuring correspondence between the thermometer's reading and the actual temperature whenever Temp looks. Yet, Pritchard suggests, the truth of Temp's beliefs is not "to a significant degree creditable" to Temp's cognitive character, since the truth of those beliefs is entirely due to the hidden person in the room. Thus, the anti-luck virtue account rules that Temp's beliefs

[11] Note that this safety condition, when satisfied, entails that the proposition ϕ is true in the actual world. Hence, Pritchard's account says that we can know only true propositions—something which, as we'll later see, almost everyone accepts, philosophers and non-philosophers alike.

are not knowledge. And it appears that the philosophical community agrees with Pritchard's intuition that Temp's beliefs indeed do not count as genuine knowledge.

Let's suppose for the sake of argument that Pritchard's account aligns with all verdicts of the philosophical community (although I'll suggest in Chapter 4 that it doesn't). Still, we'll want an explanation for why we humans would have a special word to designate safe ("non-lucky") virtuous belief.[12] Pritchard is sensitive to this matter: Taking after Craig (1990), Pritchard appears to suggest that the verb "know" came to convey "safe virtuous belief" because of humans' need to identify informants who are reliable not only about the particular matter in question (that is, whether or not *p*), but also about *similar* matters, either presently or in the future (Pritchard 2010, p. 60). Surely, identifying reliable informants for a general subject area, and not just for one particular question, is a genuine human need. And I agree with Pritchard that if someone's true belief is "safe" but not "virtuous"—say, "an agent with poor mathematical skills who is trying to work out a series of mathematical problems, but who is unbeknownst to him being helped by a wizard who ensures that all his beliefs formed on this basis are true" (p. 61)—then typically we won't want to rely on that person's beliefs about similar matters in the future.

But suppose that Temp, and the just-mentioned wizard-assisted mathematician, are in no danger of losing their "assistance": The person hidden in the room with Temp will be there

[12] Perhaps there wouldn't be such a glaring need for explanation if Pritchard's account was that knowledge is simply "virtuous" belief (as defined), and did not require "safety" as well. In fact, some philosophers have proposed just this sort of account (e.g., Greco 2007 and Sosa 2007). After all, we do seem to value getting good things—like true beliefs—at least partly because of our own character traits, such that our success is in some sense a personal achievement. However, as Pritchard carefully argues, doing away with the safety condition renders the account incapable of matching the philosophical verdict on all cases in the literature. Thus, we'd still be in the dark about why we do or don't ascribe knowledge in those cases which the virtuous belief account doesn't account for.

as long as Temp is (or, even better: suppose that a highly accurate computer program is ensuring the correspondence whenever Temp looks at the thermometer, against unbeknownst to Temp), and that the wizard will be discreetly aiding our hapless mathematician his whole life. I'm quite sure that most any philosopher would still deny that Temp's belief and the mathematician's beliefs fail to constitute knowledge—yet now both of these protagonists look to be reliable informants *par excellence* about the subject areas in question.[13] Perhaps Pritchard could reply that even if Temp and the mathematician happen to be good "subject area" informants due to quirks of their situation, it's not a good idea to rely on *similar* people in *similar* (but less quirky) situations. *In general*, we don't want to rely on someone with poor mathematical skills to tell us truths about mathematics, and we don't want to rely on someone in a room with a broken thermometer (but which they don't realize is broken) to tell us what the temperature is.

Pritchard doesn't himself suggest this extension of his idea about knowledge ascriptions signaling reliable subject-area informants. But I believe this extension is worth pursuing: What if "K knows that *p*" conveys, at least implicitly, that people *like* K are very likely to be right about matters *like* whether or not *p* in future situations *similar* to K's? This is quite close to my own hypothesis about what we implicitly understand "K knows that *p*" to convey (see § 1.2 above). And I'll spend the next four chapters fleshing out the hypothesis, and proposing how this state of affairs—about what we implicitly understand "K knows that *p*" to convey—might have

[13] A similar concern applies to Gibbard's (2003) proposal that ascribing knowledge to K involves, at least implicitly, making a "plan to defer, in certain conditions, to judgments like [K's]" (p. 248). It seems we'd readily plan to defer to judgments like Temp's and the hapless mathematician's if we were certain that their secret help would continue indefinitely. Yet we don't agree that they have knowledge.

arisen. Thus, while there's clearly more work to be done, Pritchard's and Craig's ideas are helpful starting points. I'm grateful for their precedent of paying attention to the *social* context of our knowledge ascriptions, and for hypothesizing that "know" might serve more than a merely descriptive function: It might also convey normative suggestions to others.[14]

As we'll soon see, my own proposed explanation for our knowledge ascription practices does not take off from a particular set of necessary and sufficient conditions for the instantiation of knowledge. However, I want to be clear at the outset that my hypothesis is fully *consistent* with the existence of some set of truth conditions for knowledge-possession, and hence for the objective truth or falsity of knowledge ascriptions. For instance, perhaps Pritchard is correct that knowledge is instantiated precisely when a belief is both safe and "virtuous". (His account does seem to have gained some currency among philosophers in the last few years, although certainly not a consensus.) Thus, my project might be tangential to the search for an account of knowledge. In Chapter 6, however, I'll argue that what I'm doing here still very much counts as philosophy.

[14] Likewise, we could view my own hypothesis as an extension of Gibbard's (2003) proposal (see footnote 13), where the "conditions" in which we plan to defer to judgments "like K's" can vary in significant ways from K's own conditions. However, the point remains that there's still work to be done: in spelling out what these "conditions" are exactly, and in figuring out how our word "knowledge" might have become sensitive to such plan-making considerations.

CHAPTER II: Foundation of Trust

2.1 Our Knowledgeable Ancestors

In this chapter, I'll propose how evolution—both natural and social—might have led our early human ancestors to develop a word similar in many ways to our modern-day verb "know-that". After sketching my evolutionary hypothesis in this and the following section, I'll show in §§ 2.3 and 2.4 that it offers explanations for some of the most general and pervasive features of our knowledge ascription practices. Then, in § 2.5, I'll suggest that the hypothesis can also help explain why at least some speakers today tend to deny that "K knows that *p*" in certain cases like Gettier's (see Chapter 1) where K's belief is true and by all accounts "justified". I'll delay consideration of whether my hypothesis is, on the whole, physically plausible until Chapter 5, at which point we'll have a fuller understanding of just what the hypothesis predicts about our cognitive capacities.

Before we begin, I want to anticipate a concern my reader might develop: The evolutionary history I propose in this chapter is speculative, since I don't have any direct evidence for it. (Alas, we don't have any way to personally witness how natural and social selection operated over the generations of our early human ancestors.) However, once I present my full hypothesis about our evolutionary history, I'll spend the rest of the dissertation arguing that the hypothesis (or at least something in its vicinity) is likely true, because it explains so many features of our modern-day uses of the word "know", and because it doesn't seem to make any false predictions. Thus, I ask my reader to bear with me through the development of a hypothesis which might *at first* be rather hard to believe.

To begin, then: From the perspective of natural selection, we can imagine why early humans might have developed an innate and primarily subconscious "trust heuristic", which caused them to trust that certain propositions are true upon having mental states of particular sorts.[15] When I say that K "trusts that *p*", I mean that (1) K does not even subconsciously take into account the possibility that ~*p* when deciding what to do, and (2) K has no disposition to form a contingency plan—consciously or otherwise—for the event that ~*p* should she plan or undertake an action whose success depends on the truth of '*p*'.[16] To "take into account the possibility that ~*p*", I mean that K at least subconsciously predicts the valence of the consequences of her taking action A and it being the case that ~*p* (rather than *p*), such that she might decide not to take action A after all. Is it plausible that our ancestors capable of such apparently sophisticated calculation, even if only subconsciously? In fact, Seligman et al. (2013) review the growing body of evidence that in both humans and non-human mammals, certain neuron clusters very closely track (up to linear transformation) the values yielded by normative theories of expected utility for a given action under consideration (at least in carefully controlled environments).

[15] I mean to be neutral about the exact physical realization of this heuristic in the brain. In particular, I don't mean to say that the cognitive apparatus involved is necessarily a particular localized brain region, whose only function is to produce trust (as defined below). The system I hypothesize here might be realized in various different ways—perhaps most plausibly as a neural network involving many multi-use clusters of neurons spread over different brain regions. While I'm not aware of direct evidence for any particular physical realization of something like a trust heuristic (as I define it), I'll argue in Chapter 5 that it's not implausible that there is *some* such realization.

[16] By an action "succeeding", I mean that it both (1) brings about the goal or goals (if any) which K intends (consciously or otherwise) to accomplish by taking the action, and (2) does not bring about any consequences which K does not expect to cause and which K would be dismayed at having caused.

As for component (2) of my definition of trust, a “contingency plan” for the event that *q* is a representation of “what to do”—or at least the beginnings of what to do—in the event that we become subjectively certain that *q* (for some proposition ‘*q*’) while performing some action or sequence of actions which would be affected by the truth of ‘*q*’. The plan might include certain physical actions, or it might simply be “affective”: a plan for how to react emotionally. (For instance, if I’m walking home with my umbrella in tow, my contingency plan for the event of rain is to open up my umbrella. If I’m walking home through the countryside without an umbrella, my contingency plan for rain might simply be to smile and not get too perturbed—assuring myself that I’ll get dry eventually.) I think it’s plausible, just by reflecting on our daily experience, that we *do* form such contingency plans on a regular basis (even if only half- or sub-consciously). And there is indeed plenty of empirical evidence that both humans and non-human mammals continuously form and update cognitive representations of “what to do” in possible future scenarios (Seligman et al. 2013).[17]

Note that component (1) of trusting that *p* does not strictly entail component (2), since we might completely disregard the possibility that ~*p* when deciding *which* action to take, yet still be *prepared* for the event that ~*p* when performing the action we decide to take. What I’m proposing, then, is that our ancestors evolved to have an innate trust heuristic which, when a particular salient proposition ‘*p*’ seemed to them (at least subconsciously) more likely than not

[17] Note that as defined, trusting that *p* is compatible with cognitively representing some non-zero probability for the event that ~*p*. (The crucial question is whether K takes that probability into account when planning, and whether she’s disposed to form contingency plans for ‘~*p*’.) Moreover, note that having a contingency plan for ~(*p* & *q*) does not necessarily entail having a contingency plan for both ~*p* and ~*q* (or even for just one of these). For instance, consider Effie the mathematician (whom we’ll encounter again in Chapter 3), who has just edited a draft of her new textbook. Plausibly, she can have a contingency plan for the event of at least one of her claims being false (for instance, she might put a caveat in the book’s introduction saying that she may have missed a few errors despite her best efforts) without having a contingency plan for any particular claim being false.

to be true,[18] either (a) ensured that K took into account the possibility that $\sim p$ when planning what to do, and fostered or maintained contingency plans for the event that $\sim p$, or (b) disposed K to disregard the possibility that $\sim p$ in planning, and automatically got rid of any contingency plans already in place for the event that $\sim p$ (and inhibited subsequent formation of such plans). I've called this a "heuristic" because surely there's almost always *some* likelihood, from our perspective, that a proposition which seems very likely to us is false (e.g., "There won't be a massive flood tomorrow"). Thus, the heuristic I've proposed involves *simplification*: We could take more into account when planning what to do, and make more preparations for relatively unlikely possibilities, but the trust heuristic causes us to act as if there were *absolutely no* reason to do so.

What were the advantages for our ancestors of having a trust heuristic, such that they evolved to have one? Consider first that had our ancestors *always* taken into account low-probability possibilities when deciding what to do, they might well have become too risk-averse for their own good (that is, they might have died earlier and/or reproduced less than they could have). For instance, had they always taken into account the possibility of being attacked by a wild animal during the night (despite being aware that such attacks happened hardly ever, if at all), they might never have been willing to take multi-day journeys (say, for hunting) that would require sleeping away from their protected home area. Or, they might never have tried any new kind of berry, for fear that it might cause instant death. (Low-probability possibilities can

[18] Seligman et al. (2013) review the growing body of evidence that specific neuron clusters in both human and non-human mammals track likelihoods of salient propositions, based on current mental states (sensory perceptions, memories, beliefs, etc.). Thus, there's good reason to think that a trust heuristic of the sort I'm describing could have "piggybacked" on the systems already in place for tracking likelihoods, taking as input whether or not a given proposition appeared to be more likely than not to be true.

derail action when the potential consequences seem bad enough.) Next, consider that our ancestors' brains—like ours—had a limited storage capacity. Thus, storing a representation of "what to do" for *every* action-relevant possibility they could think of (consciously or otherwise) consistent with their sensory experiences surely wasn't feasible. Without a way to avoid "comprehensive" contingency planning, our ancestors' brains would have used a large amount of storage space for contingency plans for very unlikely events. (Consider what would happen if our ancestors didn't take for granted that they would wake up in the same place where they went to sleep, that the sun would rise the next morning, that the food and water they were accustomed to consuming would continue to satisfy their hunger and thirst, that their friends and relatives wouldn't suddenly turn against them for no reason,…) Plausibly, then, survival and reproductive success would have required either not making *any* contingency plans, or having some way to *limit* their proliferation.

But storing *no* contingency plans (or only a few) surely wasn't a promising route, since then our ancestors would have been cognitively unprepared for a great many scenarios which appeared *relatively* unlikely but in fact occurred.[19] (For instance, one of our ancestors might not have gotten ill *immediately* after eating a few of the new type of berry she just found, but it was still perhaps better to wait a while before eating more.) Thus, while we can expect that early humans who readily formed contingency plans generally outcompeted those who didn't

[19] One might worry here that cognitively storing any contingency plans at all would quickly become overwhelming— for us or our ancestors—given the remarkable number of propositions on which the success of our actions on any given day depends, and which appear to have at least some likelihood of being false. ("I'll get to work on time if I wake up at 6:30", "There won't be a traffic standstill on the highway", "My coworker won't suddenly be out sick today", …) However, we can surely imagine that our brains get rid of contingency plans for the event that *q* once the matter of whether or not *q* is settled. (For instance, we don't retain our contingency plan for the event of heavy highway traffic once we've arrived at work on time.)

(that is, they lived longer and reproduced more), we can also expect that among those who did, some had competitive advantages over others. In particular, those who had a subconscious heuristic to trust that *p*—and so prune or inhibit contingency plans for the event that ~*p*—roughly only when it was *very* unlikely from their perspective that ~*p* were generally better off than those who formed and lost contingency plans either haphazardly or otherwise without much regard to likelihoods.[20] We can thus expect natural selection to have calibrated our ancestors' trust heuristics, over many generations, to cause them to trust that *p* at time t roughly only when it was *very* unlikely, given their mental states at t, that ~*p*. (I'll make this statement of the "evolved conditions for trust" more precise in Chapter 5, and also discuss some plausible exceptions. But for now, this statement will suffice.)

Assuming that our early human ancestors had a robust consciousness like ours—a domain of self-awareness—it wouldn't be surprising if the trust heuristic also evolved to remove or prevent any *conscious* subjective feeling of doubt regarding a proposition which our ancestors trusted (should such a proposition rise to conscious consideration). That is, there plausibly would have been fitness-related advantages to not feeling any sense of uncertainty upon consciously considering a proposition which one's trust heuristic had designated as trustworthy. Had our ancestors felt anything other than a "default" or "mundane" or "prosaic" attitude toward the thought that *p* (for some proposition '*p*' which they subconsciously

[20] Implicit in this statement is the proposal (which I very much intend) that actions of the trust heuristic can be "undone"—whether by the trust heuristic itself, or by some other cognitive system—if subsequent mental states indicate that it's not so unlikely that ~*p* after all. In fact, it's plausible that the trust heuristic (if it exists as I've defined it) includes a single "all-purpose" contingency plan for the event of receiving evidence which significantly tells against a proposition that one began to trust at some previous time. For instance, this contingency plan might include an affective conscious-level feeling of surprise (a feeling which we do often experience upon receiving countervailing evidence), so as to ensure that one won't continue proceeding on the assumption that *p* in one's conscious deliberations and planning.

trusted), they might well have begun *consciously* making contingency plans for the event that ∼*p*, or *consciously* taking the possibility that ∼*p* into account when deliberating about what to do. But if the trust heuristic not only got rid of *subconscious* sensitivity to the possibility that ∼*p*, but also made unlikely any subsequent *conscious* overriding, our ancestors would have enjoyed the advantages of not unwittingly undoing the adaptive work of their trust heuristic. Had our ancestors sometimes consciously felt uncertain about whether in fact *p* after their trust heuristic had already produced trust that *p*, our ancestors would not have enjoyed the full time-saving and cognitive-space-saving benefits of the heuristic.

Let's suppose then, as seems at least possible, that a likelihood-calibrated trust heuristic, producing not only trust (as defined) but also an absence of any subjective doubt, conferred fitness advantages on those of our ancestors who had one. Consequently, by the time of significant human society-formation and the development of language, we can expect that nearly every human had such a trust heuristic, and consequently had experience with both subjective feelings of doubt upon considering certain propositions, and an absence of doubt (if not an *active* feeling of confidence) upon considering other propositions.[21] We can also expect that early humans recognized or inferred that they shared certain features of their conscious mental lives. And so we can expect that through a process of social trial-and-error, they coordinated on words to signify at least some of those shared features. The sort of social trial-

[21] Plausibly, had our ancestors felt an active feeling of confidence about *every* proposition which they trusted and which entered their consciousness ("The sun is up!"; "That's my friend!"; "These berries are safe!"), they would have been distracted, without any compensating advantage. A default state of "no doubt" in most cases would have conferred all the benefits of mentioned above (that is, not undoing the work of the trust heuristic) without the distraction.

and-error I have in mind here is what some have called "memetic evolution".[22] It wouldn't be surprising, then, if early humans developed a word or words to refer to the "default" subjective state associated with the trust heuristic. My hypothesis is that our English words "certainty" and (in at least some of its uses) "belief" are the descendants of early human words for the no-doubt state associated with the trust heuristic. And similarly, of course, for the counterparts of "certainty" and "belief" in other languages, modern and ancient. [23]

With these words in place, our ancestors could report on their own and others' certainty, which would have been very useful information for the sake of social coordination. But it wouldn't be surprising if our ancestors then also developed a word or words—again through a process of memetic evolution—to categorically *agree* or *disagree* with the certainty of others. After all, any given one of our ancestors would have often found themselves in a situation where (1) someone else, K, was certain that *p*, and (2) they were also certain that *p*. In such a situation, our ancestor would have wanted a word to express—to K or to others—that they *agreed* with K's certainty: They were on the same page as K. My hypothesis is that our words "correct" and "right"—when applied to the beliefs of others—are the descendants of the

[22] Memetic evolution is the hypothesized (and well-evidenced) process by which certain cultural items propagate through societies and gain dominance over alternatives—due to perceived advantages and/or superior status—and acquire widely recognized conventional meanings. Importantly, just like in natural selection, there needn't (and usually isn't) any single person or group who orchestrates the process. Rather, certain items catch on—and others don't—for reasons that may not be localized to any particular agent. (And indeed, the process may not seem perfectly "rational" in hindsight from the perspective of a single agent). Dawkins (1989), who coined the term "memetic evolution", offers the following description: "Examples of memes are tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches. Just as genes propagate themselves in the gene pool by leaping from body via sperms or eggs, so memes propagate themselves in the meme pool by leaping from brain to brain via a process which, in the broad sense, can be called imitation" (p. 192).

[23] My hypothesis here is inspired by the work of Peter Railton (2014 and ms.), who very convincingly argues that our concept "belief" correlates with an affective state of trusting that *p* (which he defines similarly to how I've defined it here). He doesn't extend his thesis to our concept of knowledge, as I attempt below; but without his precedent, I wouldn't have seen the possibility of the extension.

words our ancestors developed to agree with others' certainty. (For instance, "K is correct that there isn't any good wood around to start a fire".) Further, any one of our ancestors would have sometimes found themselves in a situation where (1) K was certain that *p*, and (2) they were certain that ~*p*. In such a situation, our ancestor would have wanted a word to express—to K or to others—that they *categorically disagreed* with K's certainty. (I say "categorical" to distinguish the non-categorical disagreement we have with someone else's certainty when we ourselves are unsure *whether or not* it's the case that *p*.) My hypothesis is that our words "incorrect and "wrong"—when applied to the beliefs of others—are the descendants of the words our ancestors developed to categorically disagree with others' certainty. (For instance, "K is *wrong* that there isn't any good wood around to start a fire—here's some right under this tree".)

So suppose our ancestors had developed words or phrases to convey "correct certainty" and "incorrect certainty". But now imagine one of our ancestors, "Certy", who found himself in a situation where (1) he was certain that *p*, but (2) some others around him were *not* fully certain that *p*. Moreover, suppose Certy remained certain that *p* on conscious reflection, and suppose he was certain that he was aware of all the relevant evidence possessed by the others around him. (By someone's "evidence at t relevant to whether or not *p*", I mean the propositions—true or not—which they trust at t and which seem, at least from our perspective as observers, to speak either in favor of or against trusting that *p*, either by themselves or conditional on other propositions.) Presumably, Certy would have wanted a way to *urge* his companions to be certain that *p*—in effect, “I advise that we all be certain that *p*!”[24] (We can imagine that Certy wouldn’t have wanted to convey this advice—at least not yet—if he wasn’t sure whether some of his potential listeners had evidence of which he wasn’t aware, and which might change his mind about the matter.) For instance, suppose that Certy was among a group preparing to go hunting, when several members said they weren’t certain that there was enough daylight left to make the trip worthwhile. But Certy, whose trust heuristic activated for the proposition “There’s enough daylight left for a decent chance of catching an animal”, would have wanted a way to say, “I advise being certain that there’s enough daylight left to make the trip worthwhile: Let’s get going now!”.

Likewise, imagine another of our ancestors, “Duby”, who found herself in a situation where (1) she was uncertain that *p*, but (2) some others around her *were* certain that *p*. Moreover, suppose Duby remained uncertain on conscious reflection, and suppose she was certain that she was aware of all the relevant evidence possessed by the others around her. Presumably, Duby would have wanted a way to urge that her companion be *un*certain that *p* (at least for the time being)—in effect, “I advise that we all be uncertain for now whether or not *p*!” (We can imagine that Duby wouldn’t have wanted to convey this advice—at least not yet—if she wasn’t sure whether some of her potential listeners had evidence of which she wasn’t aware, and which might change her mind about the matter.) For instance, suppose that Duby was among a group foraging for berries, and one member of the group found a new kind of

[24] If, as many philosophers have suggested (e.g. Alston 1988), we don’t have immediate conscious control over what we’re certain of at a given time (in my terminology, that we can’t consciously override our trust heuristic), wouldn’t any such advice be pointless? Not if the implicit message is, “Let’s at least *act* as though we’re certain that *p*”.

berry that looked much like another kind of berry the group frequently ate. Perhaps most of the group agreed that the new berries must be safe to eat, based on the similarity in appearance—but Duby felt uncertain (due to her trust heuristic not activating) that the new berry was indeed safe to eat. Thus, she would have wanted a way to say, "I advise us all being uncertain right now whether the new berries are safe to eat: Let's not eat a whole lot of them all at once, but rather test them slowly".

How might Certy and Duby have conveyed the messages they wanted to? The words for "correct certainty" and "incorrect certainty" wouldn't have sufficed. Certy's companions were not certain of the relevant proposition—"There's enough daylight left"—so neither "Your certainty is correct" nor "Your certainty is incorrect" would have applied. And Duby didn't want to say that her companions were *incorrect* that the new berries were safe to eat: She just wanted them to stop being certain for the time being. Perhaps Certy and Duby could have simply made do with reporting their own levels of certainty: "I'm not certain that the berries are safe; "I'm certain that there's enough daylight left". But the reference to "I" in these sentences might have seemed to give them less force: There's no direct implication about the speaker's view of his or her listeners' state of (un)certainty. The "I" also might give the implication that it was simply the speaker's *opinion* that it would be better for others to be (un)certain—but for all the speaker could tell, there was no *objective fact* of the matter about what would be best. What Certy, Duby, and others like them might have wanted, then, was a word that invoked *objective reasons* for being (un)certain, and which made no explicit reference to the speaker's own mental state—the implication then being, "It's not just *me* who advises us to be (un)certain—it's a *fact* that it's best for us to be (un)certain right now". If our ancestors had some word analogous to our normative "should", then perhaps they could have said (the equivalent of) "We should/should not be certain that *p*". But let's suppose they had not yet developed an all-purpose "should" analogue which could be appended to any other verb-phrase—or that they simply landed on a more economical word to express impersonal certainty advice. Let's use "gno" to denote this hypothesized, memetically-evolved normative word. Our ancestors would have used it as follows: "You/We gno there's enough daylight to make the hunt worthwhile [= We should be certain that there's enough daylight]"; "We don't gno right now that the new berries are safe to eat [= We shouldn't be certain that the berries are safe]".

If my hypothesis about memetic evolution is correct, "gno" developed as a way to invoke objective reasons—that is, something beyond the speaker's or anyone else's personal opinions—for being (un)certain, rather than *merely* to express the speaker's advice. ("I advise us to be (un)certain that *p*" sounds subjective, less forceful, and more open to debate.)[25] And we can expect that this implicit appeal to objective reasons had implications for when speakers were willing to use "gno" and "not-gno". After all, I think it's plausible that our implicit conception of objective reasons is of normative facts that apply *across place and time*, and not

[25] I mean to be silent here on whether there *are* any such objective reasons. (I leave that matter for the consideration of other epistemologists.) But let me note here that on my hypothesis, our ancestors who used "gno" weren't necessarily being *disingenuous* in invoking objective reasons for (un)certainty. After all, any ancestor who wanted to urge others to be (un)certain presumably had that desire partly because of the functioning of their own trust heuristic, which had already led them to trust (or not trust) that *p*. Moreover, the trust heuristic (if it exists as I've hypothesized it) surely must work by *abstracting away* from certain details of our current mental states in order to categorize our current position as being of a certain *type*—a type where trust of a proposition of the type '*p*' is either appropriate or not. (After all, we can't cognitively store a representation of "what to do" trust-wise for *every possible* combination of propositions and sets of fully-detailed mental states.) So if we all have at least an implicit sense of how our trust heuristic works (which I think is plausible), then we'll at least be disposed to regard our (un)certainty that *p* on any given occasion as being appropriate in light of certain *general* facts of our situation—facts which would call for analogous (un)certainty in other situations where those facts apply.

just to individual situations. That is, if we "should" do A in situation S, then if we're ever in some relevantly similar situation S′, it's best for us to do the action A′ which in S′ is relevantly analogous to A. (I'll say a lot more about these implicitly understood "similarity" and "analogue" relations in Chapters 3 and 4.) For instance, if our teacher tells us, "You should have studied harder for last week's exam", we naturally expect him to mean "I advise you to study harder than you did this time before the next exam". When we shove our friend in the sandbox and our mother or father exclaims, "You shouldn't do that! Please apologize to Caleb", we expect similar reproach should we do likewise in the future (although as we all know, that expectation doesn't *invariably* keep us from misbehaving again).[26]

If I'm right about our "generalized" interpretation of appeals to objective reasons, then those of our ancestors who used the phrase "We gno that *p*" would have at least implicitly expected their listeners to understand them to not only mean "I advise us to be (or remain) certain that *p*" but also to intend the following *prospective* message:

> (PROS+) "I would advise anyone, now or in the future, who's in a position similar to ours to do the same sort of thing as what we are doing or would do in being certain that *p*".[27]

[26] The philosophical literature provides further evidence that we naturally understand objective reasons to apply across relevantly similar situations. After all, the "general" nature of reasons is a consistent theme in ethics, metaethics, philosophy of action, and philosophy of law. For instance, Schauer (2009) posits that "a reason is almost always more general—broader in scope—than the result or decision that it is a reason for":

> When a physician says that she prescribed a statin drug because the patient had high cholesterol, she is saying that there is a reason … to prescribe a statin drug not just in *this* case of high cholesterol but in *all* similar cases of high cholesterol (p. 176).

[27] I won't propose a formal analysis of "doing the same sort of thing as K would do in being certain that *p*", since I'm not sure that any general analysis can be given for this (arguably) intuitive idea. However, thanks to the many examples in Chapters 3 and 4, it should become quite clear what "doing the same sort of thing" amounts to, given a specified "similar position".

Likewise, speakers who used the phrase "We don't gno that *p*" would have at least implicitly expected their listeners to understand them to not only mean "I advise us to not be (or remain not being) certain that *p*" but also to intend:

> (PROS-) "I would advise anyone, now or in the future, who's in a position similar to ours to do the same sort of thing as what we are doing or would do in not being certain that *p*".[28]

We can imagine that speakers like Certy and Duby at least implicitly saw an advantage in conveying these prospective advisory messages *along with* the primary message "I advise us to (not) be certain that *p*". For they likely would have seen the benefit in their listeners developing *dispositions* or *habits* to be certain (or not certain) in similar future situations. Certy presumably would want his listeners not only to be certain *now* that there's enough daylight left for the hunt, but also to be certain of similar things ("There's enough daylight left") in similar future situations ("We're thinking about leaving for a hunt, and the sun is just about at its highest point in the sky"). Duby presumably would have wanted her listeners not only to be uncertain *now* that the new berries are safe, but also to be uncertain of similar things ("These new berries we just found are safe) in similar future situation ("We've just found a new kind of berry, and no one's tried them yet").

In the next section, I'll begin exploring some further implications of the tacit understanding that "gno" conveyed a prospective advisory message.

[28] As I've parsed it here, "We don't gno that *p*" would have been ambiguous between "I advise us to be uncertain whether or not *p*" and "I advise us to be certain that ~*p*". However, speakers surely could have resolved this ambiguity by qualification or context. (For instance, it might already have been clear to listeners, due to previous assertions, that the speaker was uncertain whether or not *p* as opposed to certain that ~*p*, or vice versa.)

2.2 From “Gno” to “Know”

Equipped with the normative “gno”, our ancestors presumably would have begun applying it not only to their listeners (as in “We/You (don’t) gno that *p*”) but also to third parties. For instance, Duby might have wanted to convey that she would advise *all* members of the foraging group to be uncertain that the new berries were safe—not just those currently within earshot. She could then imply, “If you or I see anyone else, we should advise them not to be certain that the new berries are safe (if the matter is still relevant)”. It would have been natural, then, for Duby to say not only “We don’t gno that the berries are safe”, but also (of another member not within earshot), “She doesn’t gno that the berries are safe either, since I’m certain she’s never tried them before—I’ve been with her every time she’s gone foraging, and we’ve never found berries exactly like this before”. Note that Duby probably did *not* want to say that someone else “doesn’t gno that *p*” if she wasn’t sure whether this other person had relevant evidence which she did not (and which, if she found out about it, might change her mind about the matter). Thus, we can expect that our ancestors used “not-gno” in the third person only when they were certain that their target (that is, the third party in question) didn’t have *more* relevant evidence than they did.

How about third-person *ascriptions* of “gno”? Again, these surely would have been useful for our ancestors. For instance, Certy might have wanted to convey that he would advise *all* members of the hunting group to be certain that there was enough daylight left to make the trip worthwhile—not just those currently within earshot. He could then imply, “If you or I see anyone else, we should advise them to be certain that there’s enough daylight left (if the matter is still relevant)”. It would have been natural, then, for Certy to say not only “We don’t gno that there’s enough daylight left”, but also (of another member not within earshot), “He gnos that there’s enough daylight left, too. I’ve been on previous hunting trips with him where we left later than this, and we always caught at least one animal”.

Because “gno” also conveyed the prospective advisory message PROS+ (see § 2.1), there were surely times where one of our ancestors wanted to advise someone else to be certain that *p*, but where they would *not* have wanted to say “You/He/She gnos that *p*”. For instance, suppose that Duby and her foraging partner discovered a new type of berry several weeks ago. They prudently tried just a few at first, and didn’t experience any adverse side-effects. They preceded to eat a few more the next time they went foraging, and again felt fine afterwards. Duby thereby grew certain that the new berries were safe to eat. However, Duby and her partner hadn’t yet shared their discovery with the rest of the community, since they didn’t want to risk some of their rasher compatriots from gobbling up the new berries right away without a careful sequence of safety tests. But now that she’s certain that the new berries are safe, would Duby tell her partner, “Okay, now everyone gnos that the berries are safe”? Presumably not: Duby wouldn’t want her partner to take her to mean, “I would advise anyone who’s in a position similar to our compatriots’ current position to do the same sort of thing as what our compatriots would be doing in being certain *right now* that the new berries are safe”. The latter statement, I’ve suggested, is what Duby would have at least implicitly expected her listener to understand her to mean (in part) by “Everyone gnos that the berries are safe”. Further, it seems natural to interpret the “current position of our compatriots” in terms of *their* current relevant evidence (which includes no memories of anyone carefully testing the new berry), rather than (or not just) in terms of *Duby and her partner’s* relevant evidence. But Duby would *not* want her partner to think that she meant, “I would advise anyone who isn’t aware of anyone carefully testing a new kind of berry to be certain straightaway that the new berry is perfectly safe to eat”. After all, that might indicate to her partner that Duby has bad certainty-forming habits herself, such that her testimony shouldn’t necessarily be trusted in the future.[29]

Thus, thanks to the prospective advisory message conveyed by “gno” (due to the implicit appeal to objective reasons), we can imagine that our ancestors used the construction “You/He/She gnos that *p*” only when they would advise their target to be certain that *p even if their only relevant evidence were what their target’s evidence was*. And we can expect that they would have made this determination (at least implicitly) about their “counterfactual” willingness to advise by *simulating* having (what they took to be) all the relevant evidence of their target, and recognizing whether they themselves would be certain that *p* in that case. This latter recognition presumably would have been informed by the speaker’s own trust heuristic. (Gopnik and Meltzoff 1997, Goldman 2006, and Nagel 2012, among others, discuss the empirical evidence for our ability to simulate having (what we take to be) the mental states of others, and the incredible frequency and ease with which we do so.)

So, if my story up till now is correct, our ancestors would have said “K gnos that *p*” only when they recognized that they themselves would be certain that *p* even if their only relevant

[29] What if a speaker like Duby was not worried at all about her listeners wondering whether she herself had bad certainty-forming habits—say, because the speaker and listeners knew each other well? She might still have been concerned about “fourth” parties who didn’t know her personally (or not as well as her listeners) but who might hear her speech quoted, and who might thereby speculate that she has bad certainty-forming habits. Even aside from these concerns (which could have been either conscious or subconscious), speakers might simply have developed conventions of not *seeming* to convey a message which they wouldn’t endorse, even if they didn’t expect any bad effects to come of it. (We can imagine that there were fitness-related advantages for our ancestors who tended to develop such conventions—since they, as we, were not infallible about telling when their listeners would or wouldn’t think that they actually endorsed a message which they speech conveyed.)

evidence were K's evidence—that is, even if they were "in K's shoes". But because of this, it's plausible that in *most* cases where our ancestors said "K gnos that *p*", K was *already* certain that *p*.[30] After all, if I'm correct that our ancestors all had an evolved trust heuristic by the time of language-formation, we should expect that given the same evidence, most of them were in agreement on either trusting that *p* or not trusting that *p*. (Of course there would have been some individual variation—but natural selection should have yielded broad convergence, given early humans' common environments.) Thus, cases where a speaker wanted to advise K to be certain that *p*—and would do so even if their only evidence were K's evidence—but where K wasn't already certain that *p*, were probably the exception rather than the rule: Most uses of "You/He/She gnos that *p*" were intended as encouragement to *remain* certain that *p*. Because of this, some of our ancestors might have at least implicitly recognized the advantages of, as a default, ascribing "gno" *only* when their target was already certain that *p*.

What would the advantages have been of "K gnos" carrying the default implication "K is already certain"? First, this usage would have removed the ambiguity in third-person ascriptions of "gno" as to whether K wasn't currently certain that *p* (such that the speaker was hoping that K would *become* certain that *p*) or was already certain that *p* (such that the speaker was hoping that K would *remain* certain that *p*). But beyond this, our ancestors could then have used the short phrase "K gnos that *p*" to economically convey at least three messages at once: (1) "K is certain that *p*"; (2) "I'm certain that *p* (so I advise everyone else to be as well)"; (3) "At least so long as K doesn't gain any new evidence, she's very likely to *remain* certain that *p*". The

[30] At least subconsciously. That is, K might not have had the thought that *p* in mind *right that moment* when the speaker said "K gnos that *p*", but K was *disposed* to feel certain that *p* should the thought arise—and, moreover, she was disposed to *act* certain that *p* even if the thought that *p* wasn't conscious.

idea behind (3) is this: If our ancestors were willing to say “K gnos that *p*”, then (as I reasoned above) they most likely had recognized that they themselves would be certain that *p* if their only relevant evidence were K’s evidence. But this means, *ipso facto*, that they had recognized (at least implicitly) that if they were in K’s shoes, they would *remain* certain that *p even if they gave the matter a second thought*. (As we all recognize, we sometimes find ourselves certain of propositions which, on reflection, we become less-than-fully certain of.) Thus, those of our ancestors who were willing to say “K gnos that *p*” didn’t expect K to stop being certain that *p*—even if K gave the matter a second thought.

Combine now the potential communicative advantages (as described above) of “K gnos that *p*” implying “K is certain that *p*” with the fact that speakers rarely found themselves wanting to say “K gnos that *p* but isn’t currently certain that *p*”. Given these two factors, we shouldn’t be surprised if there was a further process of memetic evolution which resulted in our ancestors by default using and understanding “K gnos that *p*” to convey not only “K *should* be certain that *p*” but also “K is already certain that *p*”. “Gno” might thereby have come to compete with—and perhaps become preferred over—our ancestors’ words for “certainty” or “is correct”, since “K gnos that *p*” was more informative (in those cases where it applied) than “K is certain that *p*” or “K is correct that *p*”: The speaker could convey not only that *p* and that K was certain that *p*, but also that K was likely to remain certain that *p* (other things being equal). The latter piece of information often would have been useful for predicting others’ actions and coordinating with them. (For instance, Certy would not have wanted to embark on a hunting trip with a new partner if he suspected that this person might suddenly stop being certain that Certy was a friend and not a foe.)

So let's suppose, as seems plausible, that ascriptions of "gno" came to carry the default implication "K is (already) certain that *p*". If so, there would have then been an ambiguity in third-person *denials* of "gno" ("K doesn't gno that *p*"). First of all, consider cases where (1) a speaker was certain that *p*, but (2) she recognized that she *wouldn't* be certain that *p* if the only evidence she had were K's evidence. If she then said "K doesn't gno that *p*", would she be misunderstood to mean "I would advise K not to be certain that *p* (if she were here right now)"? Presumably not—so long as it was clear from the context, or from the speaker's verbal qualifications, that she was certain that *p* (and so *would* advise K to be certain that *p*). Thus, the speaker could almost surely expect her listeners to understand "K doesn't gno that *p*" to convey "I would advise K to be certain that *p*—but if my only evidence were K's evidence, I *wouldn't* advise her to be certain that *p*". (This could have been useful information for listeners, for it suggests that K might have a bad certainty-forming *disposition*. Listeners might then want to be cautious about trusting K's testimony in the future.)

Now, let's add into the mix that "K gnos that *p*" came to convey by default "K is (already) certain that *p*" as well as "I would advise K to be certain that *p* (even if my only evidence were K's evidence)". Then, in a context where it's clear that the speaker is certain that *p*, "K doesn't gno that *p*" would, by itself, be ambiguous between the following: (1) "K isn't certain that *p*, but I would advise her to be (even given just her evidence)"; (2) "K isn't certain that *p*, and I wouldn't advise her to be given just her evidence"; (3) "K is certain that *p*, but I wouldn't advise her to be given just her evidence".[31] Wouldn't this troublesome ambiguity have counted against

[31] In fact, there sometimes (although probably not often) would have been even more ambiguity than that between (1), (2), and (3). I'll discuss this matter at length in Chapter 4.

the memetic evolution of “gno” towards by default conveying “K is certain that *p*”? Actually, we can imagine that our ancestors easily handled the ambiguity by relying on context, tone, and/or qualifications. To convey (3), a speaker might say, “K is certain that *p* (for some reason), but she doesn’t *gno* that *p*” (with a tonal exaggeration of “gno” to convey disapproval of K’s apparent habits). Plausibly, (2) was the most typical case, requiring a simple “Unlike us, K doesn’t gno that *p*” (with no tone modulation of “gno”, because the speaker didn’t mean to convey any disapproval of K’s habits). Speakers who wanted to convey (1) might have said something like “Oh, K *gnos* that *p*—but for some reason she’s not certain”.[32] Finally, if it was already clear from context that K was certain that *p*, then (3) would have been the only available interpretation—no need for special effort on the speaker’s part to resolve an ambiguity.[33]

Let’s suppose, then, that the ambiguities of “not-gno” resulting from the semantic expansion of “gno” to further convey “K is certain that *p*” did not derail that expansion from catching on: There would have been ways to cope (as just detailed), and the communicative advantages of the expansion (as explained above) were probably worth the cost.

2.3 Knowledge of the Present

Here is an official statement of my evolutionary hypothesis developed in the last two sections, which I’ll call the “Trust Heuristic Yields Memetic Evolution” hypothesis, or “THYME”:

> (THYME) Our words “certainty” and “belief” (in at least some uses) are descendants of words that our early ancestors used to describe the subjective “no doubt” state arising from an evolved subconscious trust heuristic. Our word “know” is the descendant of a

[32] I’ll suggest in the next section that we today use “know” in a similar way in situations of type (1).

[33] A related ambiguity of “K doesn’t gno that *p*”—albeit one shared by “K isn’t certain that *p*”—is the exact status of K’s certainty relation towards ‘*p*’: Does K think that there’s at least some possibility that *p*? Is she *nearly* certain that *p*? Is she certain that ~*p*? Again, I think there are many ways our ancestors could have resolved these ambiguities—for witness how we today solve the same problems. For instance, we might say, “He doesn’t know for sure that we’re coming” to convey “He thinks it’s somewhat probable that we’re coming, but he’s not certain”.

> word our early ancestors used to urge others to be either certain or uncertain that *p*, and subsequently also to convey that third parties were already certain that *p*. Typically, our ancestors' desire to urge others in this way was due to their own evolved trust heuristics having produced (or not produced) trust that *p*.

I'll also formulate an auxiliary hypothesis to THYME, based on my observations about the likely effect of our ancestors instituting a word which not only conveyed advice but also invoked *objective reasons* for (un)certainty. I'll call this the "Signaling a Good Example" hypothesis (SAGE), to encapsulate the proposal that our modern-day knowledge ascriptions at least implicitly signal that K's certainty that *p* is a "good example to follow" for the future:

> (SAGE) Speakers today at least implicitly expect their listeners to interpret "K knows that *p* at t" to convey the prospective advisory message "I would advise anyone in a position similar to K's at t to do the same sort of thing K did in believing that *p*".[34]

As I'll show throughout this and the following chapters, THYME and SAGE offer plausible explanations for a wide range of otherwise puzzling features about our modern-day uses of "know". I'll say that hypothesis H "offers a plausible explanation" of observed results R if there is a physically possible causal chain which begins with the facts posited by H and ends with R.

My goals for the this and the next three chapters are as follows: (1) Show that THYME, along with several auxiliary hypotheses (including SAGE), offers explanations for a wide range of puzzling features of our modern-day knowledge ascription practices, among both philosophers and non-philosophers. (2) Whenever there is some doubt about the matter, argue that no existing empirical evidence clearly contradicts any predictions of THYME or of the auxiliary hypotheses. (I'll say that a "prediction" of hypothesis H is an event or state of affairs which we

[34] When I say that speakers "implicitly expect" their listeners to interpret knowledge ascriptions to convey "I would advise anyone in a similar position to do as K did", I don't mean that speakers subconsciously think of the *exact words* that I've put in quotation marks here. Rather, I mean only that speakers have some general tacit understanding *to the same effect* as what I've written.

would strongly expect to observe if H is true.) (3) Argue that, at least given the current state of research in psychology and cognitive science, THYME and the auxiliary hypotheses are physically *feasible*: They doesn't presuppose a level of cognitive functioning which we clearly aren't capable of. Accomplishing goals (1), (2), and (3) will establish that THYME offers plausible explanations for a wide range of otherwise puzzling features of our knowledge ascription practices. For lack of better explanations for these features, I'll propose in Chapter 6 that we provisionally accept THYME and the auxiliary hypotheses as at least something close to the truth about the development and functioning of our word "know".

Since there are areas where THYME and the auxiliary hypotheses yield both predictions and explanations, I won't separate the discussions of goals (1) and (2). Rather, I'll address both goals simultaneously as I move through various features of our modern-day knowledge ascription practices. (And I'll address goal (3) in Chapter 6, once we have a better sense of the exact nature of the presuppositions which THYME and the auxiliary hypotheses make about our cognitive capacities.) Let's begin, then, with the question of whether THYME and SAGE can explain the most general features of our modern uses of the verb "know". First of all, given THYME's posit that our word "know" descended from a word our ancestors used to convey advice about certainty, THYME straightforwardly explains why we today use "know" to convey just this sort of advice.[35] For instance: "Now, you don't *know* that he'll refuse if you ask him to the dance. (Why not try?)", or "You didn't know that we'd make it home with only what we had in the tank. (You really should stop at a gas station the next time the fuel gauge is near

[35] Here and throughout, I'll use "explain" as an abbreviation for "offers to explain", for convenience. Thus, when I say "THYME explains such-and-such", I don't (yet) mean to insist that the explanation offered is the *correct* or *true* explanation. (I'll get to that matter in Chapter 6.)

‘empty’)”, or “Well, I guess now we know that the fuel gauge isn’t very accurate. (Let’s not rely on it until we can get it fixed)”.

Next, THYME explains why typically we’ll apply the word “know” only when we’re certain that our target is certain of the proposition in question. Indeed, research has shown that nearly all humans, at least by ages 6-8, tend to be willing to say that someone “knows” something only when it’s clear (or at least expected) that the person feels certain of the matter (Montgomery 1992).[36] Moreover, philosophers at least since Gettier (1963) have explicitly supposed that subjective certainty is a necessary condition for knowledge. Or, to be more careful: Philosophers typically state that “belief” is the prerequisite. However, it’s clear that some uses of “belief” don’t entail certainty, such as “I believe she’s in her office, but I’m not sure—let me check”. Unger (1975), among other philosophers, have argued that the “belief” required for knowledge must be something like “subjective certainty”. After all, Unger proposes, the following sort of statement sounds quite odd: “I know that *p*, but I’m not certain that *p*”. THYME accounts for our modern-day “certainty” condition on knowledge by virtue of the memetic evolution I’ve proposed by which the ancestor of our “know” was semantically expanded to by default convey both that the speaker advised K to be certain that *p*, and that K was *already* certain that *p* (see § 2.2).

Then again, some philosophers have argued that certainty in fact *isn’t* required for knowledge (for instance, Stanley 2008), and there is some empirical evidence that non-philosophers are quite willing to attribute knowledge in some cases where someone is explicitly

[36] Admittedly, the research studies which Montgomery cites were conducted exclusively with English-speaking children. However, there seems to be no reason to expect substantially different results for the use of the counterparts of “know” among children who speak languages other than English.

described as not being certain of the proposition in question. For instance, Myers-Schulz and Schwitzgebel (2013) asked non-philosophers—who presumably didn't have prior explicit theoretical views about the matter—for their reactions to several anecdotes, among them a case where "Tim" has overwhelming evidence that his wife is cheating on him, yet tries to explain away all the evidence and so remains confident that his wife is faithful. Nearly 90% of the subjects in the study judged that Tim knows that his wife is cheating—and, moreover, nearly 70% attributed belief. Can THYME explain these "recalcitrant" data? I believe so: When it's clear to both speaker and (at least imagined) listener that K isn't certain that *p*, saying that K knows that *p* is most likely to be understood as "Given the sort of position K is in, I *would advise* her to be certain that *p* (although as we both see, she's not)". In fact, this sort of use of "know" fits perfectly with the original use of our ancestors' "gno", as proposed by THYME: to urge others to be (or remain) certain that *p* (see § 2.1). And in fact, it seems we do find ourselves using "know" in a "hortatory" way on occasion, when we want to advise others to *become* certain that *p*, or when we want to convey that we *would* have advised someone else to be certain that *p*: "C'mon, you know the Orioles won't win the pennant this year"; "Oh, she knew that I'd be angry if she didn't do her chores".

Let's move on to another feature of our modern use of "know", which might seem mundane but still bears explanation: It appears that we strongly tend to ascribe knowledge only when we judge (at least implicitly) that, given only K's mental states at the relevant time, we too would have been certain that *p*. That is, we would have done the same thing K did had we "been in her shoes". As we'll see in the next chapter, philosophers unanimously deny knowledge in cases where, conditional only on K's conscious mental states at the time of her belief, they themselves wouldn't have been certain that *p*. (For instance, consider Dale, who believes confidently that a fair coin will land tails on the next toss.) Philosophers tend to call such beliefs "unjustified". Moreover, although I'm not aware of any studies about the knowledge-ascription tendencies of adult non-philosophers in cases of (what philosophers would call) unjustified beliefs, there is some suggestive evidence about children: It appears that by around age 6, most children tend to deny that someone knows something unless that it's clear or expected that that person has some source of reliable information for their belief—most notably, visual access or verbal messages from others (Montgomery 1992). For instance, beginning around age 6, children tend to deny that someone knows where a hidden toy is if that person appears to be "guessing", without any visual or testimonial source for their belief. THYME and SAGE offer a simple explanation for the apparently widespread "justification" condition on knowledge[37]: Our word "know" derived from a word which any one of our ancestors would have used to report the certainty of others—but only when she recognized that she *would* have been certain that *p* had she possessed only her target's evidence (see § 2.2). Thus, we inherited a word for certainty reports which we at least implicitly understand should not be used when we ourselves wouldn't advise our target to be certain that *p* if we had only *her* current evidence. This accounts quite well for our apparent justification condition on knowledge.

[37] I'm using "justification", here and elsewhere, to mean simply that most or all of us would have believed that *p* had we possessed the exact same conscious mental states as K did at t. As I'll explain in Chapter 6, many philosophers require more than the latter condition in order for a belief to be "justified" in their preferred sense. Thus, think of my term "justification" as the most minimal sort of justification for beliefs.

But there's a third "traditional" condition on knowledge which philosophers unanimously posit: truth. It appears that every contemporary philosopher either insists or assumes that knowing that *p* requires that '*p*' be true. And according to psychological studies, nearly all humans by ages 4-6 will ascribe knowledge to someone only if they themselves are certain that the proposition in question is true (Montgomery 1992).[38] Why do we apply the word "know" only when we're certain that the believer's belief is true, given that we have the phrases "is correct that" and "believes truly that" at our disposal? Why not let "know" cover some non-true beliefs as well? THYME explains the truth condition as follows: Our "know" derived from a word our ancestors used to convey advice to others to be certain that *p*. Thus, even if their target clearly was certain that *p*, our ancestors would not ascribe "gno" unless they themselves were certain that *p* (see § 2.1). Thus, if our word "know" descended from our ancestors' "gno", we can see why we deny that K knows that *p* when we're not certain that *p*—even if we perceive K's belief to be justified.

Note that THYME thereby explains our unwillingness to ascribe knowledge that *p* not only in cases where we ourselves are certain that ~*p* but also where we're uncertain *whether or not* it's the case that *p*. Although I'm not aware of scientific evidence about non-philosophers' tendencies in such cases, nor of discussion about these cases among philosophers, I myself would strongly hesitate to assert "K knows that *p*" if I myself were uncertain whether or not *p*. Suppose, for instance, that my roommate Taku and I both see it reported on television that the local roads will be closed tomorrow morning for a road race.

[38] How to account for the "non-factive" uses of "know" common among children younger than 4? If THYME is correct, it's reasonable to expect there to be a period of calibration for children's concept of knowledge. In this calibration period, children may overgeneralize the uses of "know" which they hear in others' speech.

Taku then leaves the room, and soon after my roommate Peter says, “Actually, I just checked the weather, and now there’s a prediction of a morning thunderstorm. The race will be canceled if there’s lightning, and the roads will be open after all.” I then remember that Peter has sometimes in the past made false statements about the weather forecast in order to fool me. (It’s one of his many practical jokes.) In this situation, I wouldn’t say of Taku, “He knows that the road will be closed tomorrow morning”. (Although I also wouldn’t say “Taku *doesn’t* know”. Rather, at least until I could check the weather forecast myself, I’d hedge with some statement like “I’m not sure whether Taku knows that the road will be closed”.) THYME, for the reasons given above, explains why I won’t ascribe knowledge to Taku (at least pending further evidence-gathering on my part): “Know” developed from a word our ancestors used to report others’ certainty that *p*, but only when the speaker herself was certain that *p*.

Thus, THYME explains the three “minimal” conditions which philosophers have proposed for knowledge: belief (or certainty), justification (“I would do what K did if I were in her shoes”), and truth.[39] Let’s now move to some additional explanatory work that THYME can do. I believe it offers an interesting explanation for the modern linguistic priority of “knows” over “believes” and “is certain”. (As I mentioned in Chapter 1, “know” is the 8th most common verb in the English language, and the most common mental state verb. “Think”—a cousin of “believe”—clocks in at 12th.) THYME proposes that the ancestor of our word “know” gradually became the default verb used by our ancestors to indicate that someone was certain of a given proposition—becoming preferred to the ancestors of “believes” and “is certain” by virtue of its additional message that the person in question was likely (other things being equal) to *remain*

[39] I’ll devote Chapters 3 and 4 to cases where it appears that these three conditions do not *suffice* for knowledge.

certain of the proposition (see § 2.2). Thus, THYME offers to explain why we today favor "know" over its certainty-entailing cousins: The ancestor of "know" became favored over the ancestors of "believes" and "is certain" for certainty reports, and we've inherited this favoritism by dint of linguistic tradition-passing from one generation to the next.

Can THYME even explain why we appear to prefer "I know" over "I believe" and "I'm certain" to report our *own* certainty? Plausibly, our ancestors came to favor the ancestor of "know" for first-person certainty reports at least partly because of its popularity in second- and third-person certainty reports. (That is, it was already the "go-to" word.) But they might also have seen (at least implicitly) an advantage in the more authoritative *ethos* of "gno". By reporting to others "I gno that *p*"—rather than merely (the equivalent of) "I believe that *p*" or "I'm certain that *p*"—our ancestors presumably indicated (if THYME and SAGE are correct) that they didn't just simply *find* themselves not feeling any doubt that *p*. Rather, they had given the matter at least a second thought and thereby *approved of* their own certainty: "I advise myself (consciously) to remain certain that *p*—and I would advise anyone in a similar position to do likewise". (As we all recognize, we sometimes find ourselves certain of propositions which, on reflection, we become less-than-fully certain of.) Thus, THYME allied with SAGE offers to explain why we today prefer "know" for first-person certainty reports: Our ancestors perceived the ancestor of "know" more effective for engendering trust among their listeners, and we've inherited this rhetorical preference (and may indeed feel some of that preference ourselves).

Finally, THYME suggests an explanation for why English speakers use "know" to indicate not only propositional certainty, but also assurance about the traits or features of particular people and places, as well as about the proper sequence of actions for performing some task.

I'm referring here to our use of such phrases as "I know Detroit like the back of my hand", "Do you know how to dance?", and "Alas, poor Yorick! I knew him, Horatio".[40] Plausibly, our ancestors found themselves in a subjective "no doubt" state not only with regard to certain propositions (that is, their mental representations of possible states of affairs), but also with regard to their mental representations of certain people, places, and sequences of actions. Given the similar subjective experience of certainty across all these domains, we can imagine why memetic evolution might have yielded, among at least some groups of our ancestors, the same word to report on—and convey advice about—subjective certainty in any of the domains: propositional, personal, locational, task-oriented. I've offered only a sketch here (for the sake of space), but I think we can begin to see how THYME can account for why our English word "know" takes not only propositions as objects, but also people, places, and actions.

2.4 Fallible Knowledge

Here again, for reference, are the first two hypotheses I've proposed:

> (THYME) Our words "certainty" and "belief" (in at least some uses) are descendants of words that our early ancestors used to describe the subjective "no doubt" state arising from an evolved subconscious trust heuristic. Our word "know" is the descendant of a word our early ancestors used to urge others to be either certain or uncertain that *p*, and subsequently also to convey that third parties were already certain that *p*. Typically, our ancestors' desire to urge others in this way was due to their own evolved trust heuristics having produced (or not produced) trust that *p*.
>
> (SAGE) Speakers today at least implicitly expect their listeners to interpret "K knows that *p* at t" to convey the prospective advisory message "I would advise anyone in a position similar to K's at t to do the same sort of thing K did in believing that *p*".

[40] At least among the Romance languages, there are distinct verbs to indicate certainty about propositions and familiarity with people or places (for instance, "saber" vs. "conocer" in Spanish). It would be interesting to use linguistic data to develop explanations for why English differs from the Romance languages (at least) in this regard. I'll leave such theorizing for future work.

I’ve shown how THYME and SAGE can explain several general features of our modern uses of the word “know”. There are some additional “mundane” features of our knowledge ascription practices upon which THYME also sheds light—features which might actually start to seem puzzling (as some epistemologists have noted) once we consciously reflect on them. For instance, it might seem puzzling that we so readily ascribe knowledge to those who form beliefs by processes which we clearly recognize to be fallible, such as perception, memory, testimony, and inductive inference (that is, projecting patterns observed in the past into the future). Given that we have the phrase “believes truly”, why not reserve “know” for (say) beliefs which, given the believer’s mental states, have virtually *no* likelihood of being false—for instance, “I currently exist”, or “1+1=2”? THYME offers to explain our relative “leniency” in ascribing knowledge. First, consider that philosophers—and, in my experience, non-philosophers too—overwhelmingly agree that we generally come to know that some event E is occurring or ongoing (say, that there’s a rosemary plant in my garden) upon experiencing E ourselves. More carefully: We all tend to agree that we know E is occurring when E is the content of one or more of our sensory perceptions (sight, hearing, smell, touch, taste, or proprioception). We also agree that typically we know that some event E occurred in the past when we have a *memory* of having experienced E (that is, having experienced it through one or more sensory modalities).

Now recall THYME’s posit that our knowledge ascriptions are typically mediated by our own trust heuristics: Our word “know” originated in a word which speakers used to advise others to be certain that *p*, such that speakers readily used the word when they themselves were certain that *p* (typically due to the workings of their own trust heuristic). Consider that if our ancestors' trust heuristics did not by default produce trust in propositions which were the contents of their sensory perceptions or memories, our ancestors would have been prone to time- and energy-wasting double-checking. They would not have taken their sensory perceptions at face value in planning what to do, and they would have made contingency plans left and right for the possibility that any one of their sensory perceptions was misleading. ("I seem to see the sun high in the sky, but I'll plan for the event that it's actually almost night".) This would have been debilitating: Consider just how many sensory perceptions we receive in an hour—let alone a day—whose contents are stored (at least for a time) in our working or long-term memories, and whose contents are relevant to the actions we plan to take. Now consider memories: Surely our ancestors, like us, frequently acted in light of the contents of their memories. ("I'll return to the part of the forest where I found a whole bunch of berries yesterday".) Had they made contingency plans for the falsity of the content of each (or even just a fair number) of their action-relevant memories, they would have been significantly cognitively burdened, typically without compensating benefit. Thus, THYME explains our universal tendency to ascribe knowledge to others' evidently true beliefs formed through sensory perception and memory: Our knowledge ascriptions are typically mediated by our own trust heuristics, and our trust heuristics evolved to by default produce trust in propositions suggested by sensory perception and memory.

In similar fashion, THYME explains our near-universal willingness to ascribe knowledge when we ourselves are certain that *p* and we learn that someone else has come to believe that *p* via testimony (as opposed to personal experience). Both philosophers and non-philosophers (at least in my experience) readily agree that we generally come to know that some out-of-sight event E occurred if we've heard some other trustworthy person earnestly assert that E occurred—particularly when the assertion indicates or implies that the speaker either witnessed E themselves or heard about E from someone who witnessed it (or heard about E from someone who heard about E from someone who witnessed it, etc.). Moreover, there's plenty of empirical evidence that humans by default trust the autobiographical reports of others (e.g., Gilbert et al. 1990). (In fact, our default trust of testimony can become problematic in courtroom settings: Jurors have been shown to assign great weight to eyewitness testimony and confessions, often in preference to objectively more reliable forensic evidence—and even when given statistics on the surprisingly high rates of inaccurate witness testimony and false confessions.)[41]

Now consider what would have happened to any of our ancestors who did not by default trust apparently earnest assertions about plausible out-of-sight events made by fellow members of their community: They would not have taken the testimony of others at face value, but rather would have been prone to either personally check out things for themselves, or at least make contingency plans for the falsity of any given action-relevant assertion made by someone else. And indeed, a very large proportion of the action-relevant information we have at any given time comes from reports from other people, not directly from our own senses. (Surely the same held for our ancestors, at least once they formed communities and developed languages.) Thus, those of our ancestors who did not be default trust testimony were surely much less successful in navigating daily life. In addition, those of our ancestors who did not by default trust the reports of others probably were less likely to rise in their communities, let

[41] See, for instance, Loftus (1974) and Penrod and Cutler (1995).

alone remain welcome there: They probably revealed their lack of trust of others' testimony in their words and actions, and so were not seen as good collaborators by others. (Notice that we today tend not to trust those who don't appear to trust us. I suspect the same was true for our ancestors.) Thus, if we indeed have trust heuristics which evolved through natural selection, as THYME posits, then we can easily account for our ready ascriptions of knowledge to those—whether ourselves or others—who form evidently true beliefs via testimony.

Finally, THYME explains our near-universal willingness to ascribe knowledge to at least some beliefs formed by "inductive inference": inferring that hitherto reliable patterns will continue in the future. I trust that most of us take our beliefs that the sun will rise tomorrow, or that it will be colder on average in December than July this year, or that our house will still be standing when we return from work, as obvious instances of knowledge. I and those whom I know (both philosophers and non-philosophers) also frequently take ourselves to know that people will do what they say they'll do (so long as it seems quite likely that they will, given the circumstances), such as when someone tells us, "I'll be going to Scarborough Fair". Likewise, we frequently take ourselves to know that *we* will do what we firmly intend to do (again, so long as it seems quite likely that we will).

Plausibly, those of our ancestors who regularly remained uncertain whether stable patterns of nature would continue (for instance, that the sun would continue rising each day) or that stable social patterns would continue (for instance, that their community would continue accepting them as a member) became cognitively overwhelmed with contingency plans, became far too risk-averse for their own good, and risked losing (or never gaining) the trust of others in their communities. Thus, if we indeed have trust heuristics which evolved through natural selection, as THYME posits, then we can easily account for our ready ascriptions of knowledge to those—whether ourselves or others—who believe that past reliable natural and social patterns will continue into the future (that is, barring some specific reason to think they won't).

Let's collect the various plausible proposals I've made about the evolution of our trust heuristics into a single auxiliary hypothesis:

> (FALL) Due to the pressures of natural selection, our trust heuristics evolved to produce trust by default in (1) propositions which are the contents of sensory perceptions and memories; (2) plausible propositions earnestly asserted by others in our epistemic community; and (3) propositions to the effect that past reliable natural and social patterns will continue in the future.

THYME and FALL together offer a straightforward explanation for why we readily ascribe knowledge to those who form beliefs by generally reliable but still fallible processes: Owing to the origins of our word "know" in our ancestors' "gno", and our inheritance of their linguistic patterns due to the influence of each generation on the next, we're inclined to agree that K knows that *p* whenever (a) we're certain that *p*, and (b) we recognize (at least implicitly) that we *would* be certain that *p* if we had only K's mental states at the time in question. Thanks to the fitness-related advantages for our ancestors of by default trusting propositions suggested by sensory perception, memory, testimony, and inductive inference, condition (b) is almost always satisfied when we recognize that K formed her belief via one or more of those methods.

CHAPTER III: Trouble from Within

3.1 A Matter of Perspective

In the second half of Chapter 2, I demonstrated how my hypotheses THYME and SAGE offer explanations for a number of general and seemingly mundane features of our modern-day uses of the word “know”. In this chapter, I’ll begin using SAGE to offer explanations for the “knowledge verdicts” which philosophers tend to give on various hypothetical and decidedly non-mundane cases of belief that have been discussed in the literature since Russell (1948), and especially since Gettier (1963).[42] In many of these cases, a certain believer K believes that *p*, and it is indeed the case that *p*—yet a significant proportion (if not all) philosophers report the intuition that K’s belief does not constitute knowledge. (By “intuition”, I mean a spontaneous verdict that doesn’t seem to be preceded by conscious critical reflection. And from here on, I’ll use “belief” synonymously with “certainty”, as is common among epistemologists.)

In § 2.2, I suggested an explanation for why we today don’t *always* ascribe knowledge in cases where (1) K believes that *p*, and (2) we also believe that *p*. In particular, I discussed cases

[42] I’ll discuss the reactions of non-philosophers to these sorts of cases in Chapter 5. I restrict my attention to philosophers here for two reasons: First, there’s better available evidence (thanks to the epistemology literature) about the consensus verdict among philosophers than among non-philosophers on most of the cases in question. Second, several recent experiments have found notable group-level differences between the verdicts of philosophers and non-philosophers on several of these cases. Thus, I’ll wait to discuss non-philosophers’ verdicts until we can clearly contrast them (where they differ) from philosophers’ verdicts.

where a speaker, although herself certain that *p*, recognized (at least implicitly) that she would not be certain that *p* if her only evidence were K's relevant evidence. (Recall that I defined "evidence at time t" as the propositions of which an agent is at least implicitly certain at t and which seem—at least to an observer—to count as reasons for or against believing that *p* at t.) I explained that if my hypothesis about the origins of our ancestors' word "gno" is correct, then the speaker in this case likely would *not* have uttered "K gnos that *p*". For "gno" conveyed not only the message "I advise that K be certain that *p*" but also "I would advise anyone in a position similar to K's to do the same sort of thing K is doing in being certain that *p*". (This latter prospective message attached to ascriptions of "gno" because "gno" implicitly invoked objective reasons for certainty—reasons which were abstract and so presumably held across similar situations, not only in individual instances.) And our speaker in question would *not* have wanted to convey this prospective advisory message to her listeners, or to even let her listeners wonder whether she meant to convey it. For it's natural to understand K's "position" solely in terms of *K's* relevant evidence, and not (or not only) in terms of the *speaker and listeners'* evidence. And the speaker has recognized that she wouldn't be certain that *p* if her only evidence were K's evidence: In effect, she wouldn't do the same sort of thing K did if she were in K's "position" (on this understanding).

I used this line of reasoning in § 2.3 to explain why we today tend to deny that "unjustified" beliefs constitute knowledge. (Recall that for present purposes, I've defined "justification" in a very minimal sense: K's belief that *p* is justified just in case most or all of us would also have been certain that *p* had we been certain of all the same propositions relevant to whether or not *p* as K was at the time of her belief.) Crucial to my explanation was that our word “know” derived from our ancestors’ “gno”, and that we thereby inherited the linguistic habits of our ancestors, who declined to ascribe knowledge in cases where they wouldn’t have been certain that *p* had they been “in K’s shoes”.[43] But implicit in my entire line of reasoning was the hypothesis that it’s *natural* to conceive of K’s “position” exclusively in terms of *her* evidence, and not (or not only) in terms of all the relevant facts of which the speaker and her listeners are aware (if these exceed K’s evidence). In this section, I’ll make that hypothesis explicit, and also expand on it. Before I do so, here, for reference, are restatements of the two central hypotheses I developed in Chapter 2:

> (THYME) Our words “certainty” and “belief” (in at least some uses) are descendants of words that our early ancestors used to describe the subjective “no doubt” state arising from an evolved subconscious trust heuristic. Our word “know” is the descendant of a word our early ancestors used to urge others to be either certain or uncertain that *p*, and subsequently also to convey that third parties were already certain that *p*. Typically, our ancestors’ desire to urge others in this way was due to their own evolved trust heuristics having produced (or not produced) trust that *p*.
>
> (SAGE) Speakers today at least implicitly expect their listeners to interpret “K knows that *p* at t” to convey the prospective advisory message “I would advise anyone in a position similar to K’s at t to do the same sort of thing K did in believing that *p*”.[44]

[43] We can imagine how this linguistic habit was faithfully (and continues to be) transmitted from generation to generation: Young children hear their parents or other adults sometimes use “know” in a tone of either encouragement or disapproval, as in “You know that I’ll always come pick you up from after-school care—don’t get worried if I’m a few minutes late”, or “Now, you don’t know that Aisha will like all the same games you do, so be sure to ask her what *she’d* like to play”. Plausibly, children quickly come to implicitly grasp that “know” always at least potentially carries an advisory message.

[44] I should note that Schafer (2014) makes a proposal about our attributions of “rational belief” which resembles SAGE’s proposal about our attributions of knowledge. Schafer writes, “Judging that S’s belief that P is epistemically rational involves planning to believe that P when in S’s situation” (p. 2577). However, Schafer does not suggest an *explanation* for how our uses of the word “rational” came to correspond with our sense of prudent “doxastic planning”. In fact, he argues only that rationality attributions *commit* us to making such plans; he does not propose that we’re psychologically sensitive to this commitment (implicitly or otherwise) when we ascribe rationality. So although I’m very sympathetic with Schafer’s proposal, it leaves work left to be done (not least in the task of explaining our *knowledge* ascriptions)—and I believe that SAGE might fill the gap.

My intent in this chapter is to use SAGE in particular to explain the philosophical verdicts on hypothetical cases in the literature where (1) it appears that a significant number (if not all) philosophers deny that the believer K has knowledge at time t, and (2) it appears that for those philosophers who deny knowledge, their judgment is independent of any facts beyond the history of K's mental states up through time t. That is, even if these philosophers weren't told anything about the world outside of K's head, they still would confidently deny that K had knowledge.[45] I'll show that in all such cases, SAGE offers an explanation of the philosophical community's verdict—be it unanimous denial of knowledge or a split verdict (with some philosophers ascribing and others denying knowledge). I'll also, in a few places where there might be some doubt, argue that SAGE does not make false predictions. In particular, it doesn't predict that at least some philosophers will deny knowledge in cases where philosophers unanimously ascribe knowledge. Before I begin, let me note that many of the cases I consider in this chapter are among those where many philosophers deny that the belief is "justified", albeit in a different sense than the minimal conception of justification I'm working with here (see above). However, my primary focus is on explaining *knowledge* ascriptions and denials. Thus, I won't enter into a discussion of the philosophical debates over "justification" for beliefs (although I'll make a few remarks about them in Chapter 6).

Crucial to my explanation of these verdicts will be a further hypothesis—alluded to above—about a natural interpretation of K's "positon" in the context of the prospective (and implicitly understood) message "I would advise anyone in a position similar to K's to do just as K

[45] In Chapter 4, I'll consider cases—such as Gettier's famous examples—where philosophers' denial of knowledge *does* depend on facts "beyond K's head".

did (that is, to do whatever counts in their situation as ‘the same sort of thing K did’)”. I’ll say that an interpretation of “K’s position” is natural if it’s at least a *candidate* for becoming salient to a speaker who’s considering whether or not to ascribe knowledge to K. According to SAGE, when speakers consider whether to ascribe knowledge to a believer K, they are likely to at least implicitly consider whether they want to convey the advice “I would advise anyone in a similar position to do as K did”. I’ve already hinted that there might be *multiple* interpretations of K’s “position” that could become salient to such a speaker, given the relevant facts about K’s situation of which the speaker is aware—and particularly if the speaker is aware of more facts than K is or was. (By an interpretation being “salient”, I mean that that conception of K’s position is at least one of those which the speaker explicitly or implicitly has in mind—and thereby likely expects her listener to have in mind—when she refers to K’s “position” either directly or indirectly. If SAGE is correct, then making a knowledge ascription or denial typically entails referring to K’s “position” indirectly.)

However, I haven’t yet officially offered any hypothesis on which interpretations of K’s “position”, in any given case, are candidates for becoming salient. In principle, we could imagine that *any* subset of relevant facts about K’s situation of which the speaker is aware could strike the speaker as a plausible conception of K’s “position”—and not just the *full* set of facts. (For instance, as I’ve already suggested, at least one natural interpretation in any case might be just those facts of which K was aware—not including any facts of which the speaker is aware but K wasn’t.) However, it seems plausible, even in the abstract, that many of these subsets just won’t be *natural* interpretations for creatures like us, given the distinctly non-haphazard ways in which our minds categorize information.

Thus, I'll be concerned here and in the next chapter to argue that particular subsets of facts in any given case *are* natural interpretations of K's "position", and so are candidates for being salient to a speaker if he or she considers ascribing knowledge to K. My guiding idea will be the following: There are good evolutionary reasons for us to have dispositions to mentally categorize a given agent at a given time—and with regard to a given action—in *numerous* ways. These different ways will correspond with various cognitive functions that it was important for our evolutionary ancestors to perform. For instance, we can see the advantages of easily forming a representation of some agent which includes only the relevant conscious mental states of that agent at the time in question (at least those of which we're aware)—such that we can quickly consult that representation if we wish to (say) predict what she will say or do now *right now*. (For such predictive purposes, facts of which she's not currently aware—at least not yet, or because she forgot them—aren't directly relevant.) It might also be advantageous to easily form a representation of the agent which includes all of her relevant mental states which are currently conscious *or* which she could very easily recall (that is, bring into consciousness with hardly any "effort"). This representation might be helpful if we wish to (say) communicate with the agent and meanwhile be sensitive to what information she's very likely to recall in the course of our conversation, or to predict her behavior at some time in the future—at which point she may very well have recalled some of the mental states which aren't *currently* in her working memory.

In this section, I'll first offer four supporting claims which, taken together, support my first hypothesis about the natural interpretations of K's "position" in the context of considering whether to ascribe knowledge to K and thereby (if SAGE is correct) at least implicitly considering whether to convey the prospective message “I would advise anyone in a similar position to do as K did”. My first claim is that when we use the locution “If I’m ever in a similar position as you, I’ll do the same thing you did” or “If I were in your position, I would do B, not A”, we almost never mean “If I am (or were) similar to you in *every* respect that you are now…”. Rather, there’s almost always some (more or less vague) restriction that we intend on the mooted similarity relation between “your position” and “my (future or counterfactual) position”. Thus, when someone considers the statement “I would advise doing as K did for anyone in a similar position”, any natural interpretation of K’s “position” in the context will almost certainly include only *some* facts about K’s situation at t, such that being in a “similar position” doesn’t require being similar to K in *every* respect. Specifically, the facts about K’s situation which we have in mind are plausibly only those which seem relevant to the action(s) which K performed or is considering—that is, those which seem to us to count for or against the action(s) in question, either by themselves or conditional on one or more other facts.[46]

My second claim is that when we consider being in a position similar to someone else’s, many of us direct our attention specifically to the *mental states* of the person in question: the beliefs, memories, sensory perceptions, desires, etc. which she has experienced—consciously or otherwise—up until now.[47] In effect, we naturally consider having a similar *perspective* or

[46] That is, a fact might be “relevant” even if it doesn’t seem to be a reason *all by itself* for or against the action, but rather is a reason (say) against the action conditional on some other fact which *is* a direct reason for the action. For instance: Suppose that Brendan tells me that there’s currently free food on the top floor of our apartment building. But just before I venture upstairs, Varun tells me that Brendan has recently been playing a whole bunch of practical jokes. Varun’s testimony seems to count (at least somewhat) against me taking the trouble to go upstairs *given* Brendan’s testimony, even though Varun’s testimony taken by itself counts neither for nor against going upstairs.

[47] By saying that a mental state is “conscious” at time t, I mean that the content of the state is at least on the “edge” of K’s awareness, even if K doesn’t explicitly rehearse this content to herself in thought-speech. A further clarification: Williamson (2000) argues that knowledge itself is a mental state, and that in general there are many “factive” mental states, whose obtaining entails the truth of the proposition which is the state’s content. However,

outlook on the world as K has had—whether or not K has been mistaken or misled (in her beliefs, memories, perceptions, etc.) about certain facts. Here's an example, from the domain of physical action (rather than "epistemic" action), to support this second claim:

> Sandy is at work on Friday and is getting very hungry. She forgot to pack a lunch, and since it's chilly outside, she'd rather not leave the office for a meal. She walks into the communal kitchen, opens the refrigerator, and spots a wrapped sandwich on the shelf. Sandy notices that on the sandwich wrapper, her colleague Fred has written: "Fred's—please do not take". Nonetheless, Sandy picks up the sandwich, unwraps it, and eats it, feeling guilty but satisfied. As it happens, Fred had sent an email to the office on Wednesday, saying that he wouldn't be back at work until the following week, and that the sandwich he left in the fridge was up for grabs. However, Sandy neither read nor heard about this email.

I would certainly hesitate to tell someone else who read this story, "I would advise anyone in a similar position to do just as Sandy did"—as least without qualifying what I meant by "Sandy's position". I suspect my readers would hesitate as well. But our reluctance wouldn't make sense unless we expected our listeners to think that by Sandy's "position", we might be referring only to her *perspective* (as constituted by her mental states), and not to *all* the facts of the situation of which we, as observers, are aware. After all, if the only salient understanding of Sandy's position includes *all* the facts of which we're aware (including Fred's email), then we shouldn't hesitate at all to convey the prospective advisory message. Thus, this example strongly supports my second claim: that there are natural interpretations of K's "position" which are limited to K's mental states.

My third claim to support my upcoming hypothesis is that when we consider K's "position at t" in an advice-giving context, it might be natural to think about *any given subset* of

without meaning to pass judgment on Williamson's proposal, for my purposes I'm limiting the term "mental state" to *non-factive* mental states—those constituted fully by facts "within K's head".

K’s relevant mental states at t (both conscious and subconscious) each of which is within a given “level of recall effort”—including, at the limit, *all* relevant mental states which K has ever experienced, even if she can’t recall them at all (at least currently, and at least without prompting), no matter what her level of effort. By “recall effort”, I mean (roughly) the amount of cognitive attention which K would have to devote (without the help of external prompting) to bring a certain mental state m, formed at some time in the past, into consciousness. (If m is already conscious, then of course the recall effort is zero.) For instance, it takes me very little effort to recall my memory of what I ate for breakfast this morning; recalling my memory of what I ate for lunch last Wednesday takes quite a bit more effort. Likewise, it’s relatively easy for me to recall my belief that I indeed ate lunch last Wednesday (a belief which was subconscious until just a moment ago)—harder for me to recall my memories of what exactly I ate, when exactly I ate, whom I ate with, etc. Thus, for a given level of recall effort, we can imagine all the relevant mental states of K’s which are *within* that level: Recalling them would take either that much effort or less. We can then imagine the various recall effort levels as concentric circles, each encompassing more mental states than the previous one.

Now, I think that any one of these sets of mental states has a *prima facie* plausible claim to constitute K’s “position at t”. I’m certainly not claiming that we *consciously* or *explicitly* think about other people in terms of a series of discrete recall effort levels. Rather, my suggestion is that our implicit cognitive processes are disposed to form various representations of another person (vis-à-vis some relevant action), each one corresponding to what we implicitly perceive to be a given level of recall effort for that person at the time in question. At least, it would make sense for our cognitive systems to have evolved to naturally categorize someone else at a given time in multiple ways, each corresponding to a given recall effort level: We could then be at least implicitly sensitive, to varying degrees (as the situation seemed to require), to what that person might subsequently recall. (For instance, we could focus in on a given representation or set of representations based on how much cognitive effort that person seemed to be expending or willing to expend.) I'm assuming here that the recall effort levels which we naturally think in terms of (at least implicitly) are fairly coarse-grained: It's plausible that our implicit cognitive processes don't make *very fine* distinctions here, say between a memory which would take a bit of effort to recall and a memory which would evidently take *just a teensy bit* more effort to recall. Rather, both memories are (other things being equal) grouped together.

But enough abstraction—examples are probably more helpful here. Let's consider an alternative version of Sandy and the sandwich, where Sandy has forgotten (at least momentarily) a relevant experience:

> Sandy is at work on Friday and is getting very hungry. She forgot to pack a lunch, and since it's chilly outside, she'd rather not leave the office for a meal. She walks into the communal kitchen, opens the refrigerator, and spots a wrapped sandwich on the shelf. Sandy notices that on the sandwich wrapper, her colleague Fred has written: "From Fred—feel free to take". Sandy happily picks up the sandwich, unwraps it, and eats it. As it happens, Fred had sent an email to the office on Wednesday, asking his colleagues to disregard the note he left on the sandwich: He would be coming in on Friday afternoon, and he'd like to eat the sandwich then. Sandy read this email on Wednesday, but it's been a busy week, and she's forgotten all about it by Friday.

Once again, I think most of us would hesitate to say, without qualification, "I would advise anyone in a similar position to do just as Sandy did". Yet if we expect our audience to understand Sandy's "position" solely in terms of her mental states which were conscious *at the moment she ate the sandwich*, then it seems we wouldn't hesitate to give the advice. Thus, this case provides evidence that in any given case, there are potentially salient interpretations of K’s “position at t” which encompass more than just the facts about K’s *currently conscious* mental states.

We might now wonder whether *every* natural interpretation of K’s “position” includes *all* of her relevant mental states, past and present: Perhaps in the original Sandy example (where she hadn’t read or heard about Fred’s email), all the relevant mental states just happened to be conscious (or nearly so) at the moment of action. However, another Sandy variation tells against this suspicion:

> Sandy is at work on Friday and is getting very hungry. She forgot to pack a lunch, and since it’s chilly outside, she’d rather not leave the office for a meal. She walks into the communal kitchen, opens the refrigerator, and spots a wrapped sandwich on the shelf. Sandy notices that on the sandwich wrapper, her colleague Fred has written: “Fred’s—please do not take”. However, Sandy is in a bad mood and thinks to herself, “I don’t care whether or not Fred wanted to keep this sandwich for himself—I’m hungry!” Sandy picks up the sandwich, unwraps it, and eats it, feeling satisfied. As it happens, Fred had sent an email to the office on Wednesday, saying that he wouldn’t be back at work until the following week, and that the sandwich he left in the fridge was up for grabs. Sandy read this email on Wednesday, but it’s been a busy week, and she’s forgotten all about it by Friday.

Our likely hesitancy to say “I would advise anyone in a similar position to do just as Sandy did” is best explained, I think, by there being a natural interpretation, in any given case, of K’s “position” which is restricted to facts about K’s mental states which are conscious *at the moment of her action*, and thus require *no* effort to recall. If our understanding of K’s “position” always included *all* relevant mental states, past or present, we’d invariably understand Sandy’s earlier reading of Fred’s permission-giving email to be part of her “position”—and so not hesitate to convey the prospective advisory message.

But neither of the Sandy variations we just considered gave evidence that there's ever a natural interpretation of K's "position" which lies *between* all of K's currently conscious mental states, and all of the relevant mental states K has ever experienced. Here's an example to support the idea that *any given* level of recall effort yields a natural interpretation:

> Sahib has been chosen to serve on the jury for a criminal trial. On the first day of the trial, Sahib notices that the defendant looks quite similar to his own daughter, and Sahib begins to feel sorry for her. Throughout the trial, the defense presents a very weak case, relying just on a few character witnesses, and presenting no forensic or direct evidence. By contrast, the prosecution's case is extremely compelling: Forensic evidence which seems to make it almost certain that the defendant committed the crime. When the judge polls the jury, Sahib feels pretty sure the defendant is guilty. But he votes for acquittal, out of his sympathy with the defendant. "Someone who looks as sweet as her doesn't deserve to go to jail, even if she did commit the crime," he thinks to himself. As it happens, Sahib actually saw the defendant at the time of the crime: She was his server that day at a restaurant, miles away from where the crime was committed. Sahib's memory of that event was virtually lost—but had Sahib remembered, he would have been certain that the defendant was innocent.

I would again hesitate to say, *without any qualification*, "I would advise anyone in a similar position to do just as Sahib did when he voted for acquittal". But if I could expect my listeners to understand Sahib's "position" in terms of only his conscious mental states at the time he voted for acquittal, or in terms of *all* the relevant mental states he had ever experienced (or both), then I surely wouldn't hesitate. The best explanation, I think, is that in any given case, there are natural interpretations of K's position corresponding to *any given level* of recall effort. If that's right, then there's natural interpretation of Sahib's position which encompasses his memories of the defense's weak evidence, the prosecution's strong evidence, and his feelings of sympathy with the defendant, but *not* his (virtually inaccessible) exculpatory memory. And I certainly wouldn't want my listeners to think I meant, "I would advise anyone who remembers a weak case from the defense and a very strong case from the prosecution, and who also feels sympathetic with the defendant due to her appearance, to vote for acquittal".

To motivate my fourth and final claim (really, a clarification of the second and third), consider the following variation on Sahib's case:

> Sahib has been chosen to serve on the jury for a criminal trial. On the first day of the trial, Sahib notices that the defendant looks quite similar to his own daughter, and Sahib begins to feel sorry for her. Throughout the trial, the defense presents very compelling forensic evidence which makes it all but certain that the defendant could not have committed the crime. (And in fact she didn't.) Sahib himself finds this evidence completely non-probative: He thinks that circumstantial evidence without direct eyewitness testimony can never speak either for or against a defendant's innocence. When the judge polls the jury, Sahib is personally undecided about whether the defendant is guilty. But he votes for acquittal, out of his sympathy with the defendant. "Someone who looks as sweet as her doesn't deserve to go to jail, even if she did commit the crime," he thinks to himself.

Same story as before: I would hesitate to say "I would advise anyone in a similar position to do just as Sahib did", at least without carefully qualifying what I meant. And indeed, there does seem to be a "conscious mental state" interpretation of Sahib's position which yields a message I wouldn't want to convey: "I would advise anyone who feels sympathetic with a criminal defendant because of how she looks, and who remembers not finding any of the defense's evidence compelling at all, to vote for acquittal".

The second mental state of Sahib's referenced in the latter advice—his not "finding" any of the defense's evidence compelling—is what I'll call a "reason-representation" of another mental state. When I say that K has a reason-representation of mental state m, I mean that K *appreciates* the content of m as having reason-giving force for or against some action, where this action might be a "direct" physical or epistemic action (that is, making a certain movement or forming a certain belief), or it might be to regard some other mental state as being a reason to take some direct action—or to regard some mental state as being a reason to regard some other mental as being a reason to take some direct action, etc. The "appreciation" here need not be in the form of an explicit belief ("Ah yes, that tells in favor of action A"); perhaps more often, it's a non-linguistic affective representation, conscious or otherwise. (After all, how often do we have an explicit belief about reasons when we (intuitively) act for reasons throughout our daily lives?) The representation should represent the content c of mental state m as counting in favor of (or against) action A, to such-and-such extent. That is, it should represent c with a particular valence and strength: either encouraging or cautionary, to such-and-such (perhaps somewhat vague) extent.[48] The existence of affective (as opposed to conceptual) reason-representations—albeit in the context of reasons for physical action, rather than reasons for belief—has been hypothesized and researched over the past several decades by psychologists and cognitive scientists.[49] (Can a non-linguistic affective state represent as much as I've supposed? Arguably, similar representations are formed even in non-linguistic animals—for instance, when a foraging animal learns that a particular plant is poisonous.)

Returning now to Sahib: The story makes clear that, in the terms I've proposed, Sahib has a memory of his reason-representation of the defense's evidence as *not* providing reason to vote for acquittal. And plausibly, that memory is a relevant fact for us (as observers) about Sahib's mental states: At least by itself, it counts as a reason *against* Sahib voting for acquittal. Thus, as per my second and third claims, there is indeed a natural interpretation of Sahib's position which includes his memory of this reason-representation and thus yields a message

[48] Note that this characterization of "representation" is clearly compatible with non-human animals and pre-linguistic humans representing various facts as reasons for or against believing certain things.
[49] See, for instance, Schwarz and Clore (2003). (I thank Paul Boswell for pointing me to this literature.)

that we wouldn't want to convey: "I would advise anyone who feels sympathetic with a criminal defendant because of how she looks, and who remembers not finding any of the defense's evidence compelling at all, to vote for acquittal". My fourth claim, therefore, is that among the relevant mental states of a given agent at a given time are the relevant reason-representations (and memories of them) which she has at that time, consciously or otherwise.

Here, for easy reference, are the four claims I've made about the interpretations of K's "position" which are candidates for becoming salient when we at least implicitly consider conveying the message "I would advise anyone in a similar position to do the same sort thing K did":

> 1. In general, there are natural interpretations of K's "position" which don't include *every* fact about K's situation of which we're aware. Most or all interpretations are limited to facts which seem reason-relevant to the action of K's that we're considering.
> 2. There are natural interpretations which are restricted to K's mental states (that is, excluding any facts about matters "beyond K's head").
> 3. There is a natural interpretation corresponding to all mental states of K's within any given level of recall effort—including an interpretation which encompasses *all* relevant mental states K has ever experienced (even if she can't recall them).
> 4. All natural interpretations restricted to K's mental states include all relevant reason-representations (or memories of them) which K has (within the boundaries of the level of recall effort under consideration).

These four claims support my first hypothesis about the natural interpretations of "K's position"—and thereby of "positions similar to K's"—in the context of the *epistemic* prospective advisory message discussed in Chapter 2: "I would advise anyone in a position similar to K's to do the same sort of thing she did in believing that *p*".

> (KP_M) Given a believer K and her belief that *p* at time t, given a speaker S aware of various facts about K's situation, and given any level of recall effort r, there is a natural

> interpretation for S of K's "position" which includes all reason-relevant facts (of which S is aware) about the set of K's mental states at t which appear to S to be within level r.[50]

I'll say that any interpretation of K's position falling under KP_M is an "interpretation of K's position of type KP_M". More specifically, when the level of recall effort under consideration is *zero* (that is, when we're considering only those relevant mental states which are conscious—or nearly so—at time t), I'll say that the interpretation is of type KP_M-t. When the level of recall effort under consideration is "at the limit" (that is, when we're considering *all* relevant mental states which K has ever experienced, even those she can't recall), I'll say the interpretation is of type KP_M-h ("h" for "historical").

3.2 Internal Barriers to Knowledge

In this section, I'll apply SAGE along with KP_M to a number of representative cases from the recent literature in epistemology. What I aim to demonstrate, by induction, is the following: Consider any case where a significant number of philosophers judge that a certain believer K lacks knowledge, and where this judgment appears to be independent of any facts "beyond K's head". Then there's at least one natural interpretation of K's "position" of type KP_M such that nearly any of us would recognize that *we* wouldn't believe that *p* if we were in *that* "position". (This recognition would be mediated by our trust heuristic: We mentally "simulate" being in the position in question, and then implicitly perceive what other mental states we would

[50] "KP" stands for "K's position". I've used the subscript "M" to indicate that the natural interpretations falling under this hypothesis are restricted to the facts internal to K's "M"ind. Note that K's "belief at time t" may be either conscious or subconscious: K need not feel *consciously* certain that *p* at time t in order to believe that *p* at time t. Having a subconscious belief that *p* plausibly means being *disposed* to (say) feel certain that *p* upon consciously considering the matter, and to not take the possibility that ~*p* into account when deciding what to do.

consequently have—and in particular, whether we would be certain that *p*.)[51] And if this is correct, then SAGE and KP$_M$ offer an explanation for each of the philosophical community's verdicts on each of the cases under consideration: The type-KP$_M$ interpretation in question becomes at least implicitly salient to philosophers when they consider whether or not to say "K knows that *p*", and thereby (according to SAGE) implicitly consider whether they want to convey message, "I would advise anyone in a similar position to do just as K did". Implicitly recognizing that they don't want to convey this message—since *they* wouldn't be certain that *p* if they were in K's position (as understood), let alone a similar but different position—philosophers are inclined to say "K doesn't know that *p*". More specifically, they're inclined to say "K doesn't know that *p*" because they implicitly recognize a strong aversion ("That would be wrong!") to say "K knows that *p*", and because they at least implicitly assume that there's a fact of the matter about whether or not K knows, at least once they believe they've been apprised of all the relevant facts.[52]

Why wouldn't philosophers feel equal hesitation to say "K doesn't know that *p*"? After all, in theory "K doesn't know that *p*" might seem to convey "I advise K not to be certain that *p* (given all of my evidence about K's situation)". And in many of the cases we'll consider, it's stipulated that K's belief is true. However, if it's clear to a speaker and her (real or imagined) listeners that K's belief is true, then the speaker likely would have little concern about the

[51] See Seligman et al. (2013) on evidence for our abilities and automatic tendencies to simulate being in non-actual situations, and to predict what would occur (e.g., what mental states we would have) in those possibilities.

[52] Note that this explanation for philosophers' knowledge denials doesn't appeal to the fact that the prospective advisory message contains the idea of positions "similar" to K's. Rather, it relies only on the fact that the prospective advisory message implies "I would advise anyone in the *same* position as K to do what K did and believe that *p*". In Chapter 4, where I consider cases where philosophers' verdicts rely on facts *beyond* K's mental states, my explanations for the verdicts will rely crucially on the "similarity" notion in the prospective advisory message.

possible ambiguity here: It would be clear from the context that the speaker does not mean to convey "I advise K not to be certain that *p*". Hence, philosophers likely wouldn't feel hesitation to say "K doesn't know that *p*"—as they in fact do in the cases we'll be considering.[53]

3.2.1 Mental-State Defeaters

Let's first consider cases involving "mental-state defeaters," which draw a nearly universal denial of knowledge from philosophers. Taking after Pollock (1986), epistemologists often distinguish between "rebutting" and "undercutting" mental-state defeaters. Sudduth (2008) supplies standard definitions: A rebutting mental-state defeater for K's belief that *p* is any mental state of K's which provides for K "a reason for holding the negation of *p* or for holding some proposition, *q*, incompatible with *p*." An undercutting mental-state defeater is any mental state of K's which provides for K "a reason for no longer believing *p*, [but] not for believing the negation of *p*. More specifically, it is a reason for supposing that one's ground for believing *p* is not sufficiently indicative of the truth of the belief."

Here's an example Sudduth gives of a rebutting defeater:

> Mary sees in the distance what appears to be a sheep in the field and forms the belief that there is a sheep in the field. The owner of the field then comes by and tells her that there are no sheep in the field. (ibid.)

Let's assume here that Mary is aware that the person who speaks to her is the owner of the field; that she has no memories indicating that this person is untrustworthy about such matters as whether there are sheep in the field; that nothing which she perceives about the owner indicates that he or she is joking or trying to be deceptive; and that (for some reason) she

[53] I don't mean to imply here that there actually *isn't* a fact of the matter about whether K knows that *p* in the cases under consideration. My point is simply that if a speaker does implicitly assume that there's a fact of the matter about whether or not *q*, and she feels strongly averse to saying "*q*" but not to saying "~*q*", then she's very likely to say "~*q*"—at least if she's explicitly asked about the matter.

represents the person's testimony as neither a reason for nor against believing that there's a sheep in the field.[54] Let's further assume that after hearing the owner's remark, Mary continues to believe that there's a sheep in the field. Sudduth implies that such a belief, whether true or false, would not constitute knowledge—a verdict which, based on my reading of the literature, I expect every philosopher to agree with.[55] SAGE offers an explanation for this unanimous denial of knowledge, since the following is a natural implication of "I would advise anyone in a similar position to do as Mary did" according to KP_M:

> (MAR) I would advise anyone with a somewhat hazy visual perception as of a sheep in the distance (which they represent as counting in favor believing there's a sheep there), a memory of the person who oversees the area in which they're looking asserting that there's no sheep there, and no mental states indicating that this person is being disingenuous, to continue to believe that there's a sheep where they're looking.[56]

Moreover, surely almost none of us—and certainly no philosopher—would want to seem to be conveying this message to our listeners: It would imply that we ourselves have imprudent belief-forming habits (or at least imprudent advice-giving habits). After all, if we believed ourselves to be in a position described by MAR, we certainly wouldn't be certain that there's an S in the area where we're looking. Now, is MAR in fact based on an interpretation of Mary's "position" of type KP_M? From the description of Mary's case, it appears that there are only three facts about her mental states which are reason-relevant, by either Mary's lights or ours, for believing as she did: her visual perception of "what appears to be a sheep", her recent

[54] Or, we could suppose that Mary represents the person's testimony as a reason against believing that there's a sheep in the field. The philosophical verdict on Mary's belief, and the explanation of the verdict offered by SAGE, will be the same.

[55] In fact, Sudduth says that the belief is "unjustified". But since nearly every philosopher takes justification to be required for knowledge, I'll assume here and throughout that any philosopher who judges a belief to be unjustified also judges that it fails to constitute knowledge (unless, of course, they've indicated otherwise).

[56] After this example, I'll omit mention of reason-representations unless they're specifically relevant to the example. That is, the inclusion of reason-representations within K's "position" will be implicit.

memory of the owner denying that there's a sheep in the field, and the lack of any indication that the owner is lying or joking. Thus, MAR is indeed based on an interpretation of type KP_M of Mary's position (more specifically, of type KP_M-t and/or KP_M-h), so SAGE and KP_M explain the philosophical verdict that Mary's belief isn't knowledge.[57]

Let's next consider an example of an undercutting mental-state defeater, again from Sudduth (2008):

> Bridget enters a factory and sees an assembly line on which there are a number of widgets that appear red. Being appeared to red-widgetly, Bridget believes that there are red widgets on the assembly line. The shop superintendent then informs Bridget that the widgets are being irradiated by an intricate set of red lights, which allow the detection of hairline cracks otherwise invisible to the naked eye.

Let's suppose that Bridget is aware that the person who speaks to her is the factory superintendent; that she has no memories or perceptions indicating that this person is untrustworthy; that she represents the person's testimony as neither a reason for nor against believing that the widgets are red; and that she continues to believe that the widgets are red. I believe nearly any philosopher would follow Sudduth in saying that Bridget's belief would not qualify as knowledge, whether or not we suppose that the widgets are in fact red. SAGE explains this verdict, since the following message is based on an interpretation of Bridget's position of type KP_M and is clearly undesirable to convey: "I would advise anyone to whom it appears that the objects they're looking at are red, and who has a memory of someone in charge of the objects saying—without any indication of dishonesty—that the objects are being illuminated by red lights, to believe that the objects are red".

[57] Both here and throughout Chapters 3 and 4, I'll be using many examples proposed by other authors. However, the *explanations* I offer of the intuitions elicited by these examples are my own: Except where noted, no other philosopher has (to my knowledge) offered anything like the SAGE-based explanations I propose. (In Chapter 1, I reviewed some of the previous attempts philosophers have made to explain these sorts of intuitions.)

Presumably, the fact that Bridget's mental-state defeater (that is, her memory of the superintendent's message) is conscious at the time she maintains her belief is irrelevant to the philosophical verdict on Bridget. Suppose that the shop superintendent told Bridget several hours earlier that the widgets at the factory are irradiated by red lights: When Bridget finally gets to the assembly line, she's not thinking about what the superintendent said (although she would remember it if prompted), and she proceeds to believe that the widgets are red. By all indications, every philosopher would again say that Bridget's belief fails to constitute knowledge. SAGE again explains the verdict, since the following message is based on an interpretation of type KP_M-h of Bridget's position and is clearly something we wouldn't want to convey: "I would advise anyone who earlier heard someone in charge of a set of objects tell them that the objects are illuminated by red lights, and to whom it now appears that the objects in question are red, to believe that the objects are red". Thus, although the interpretation of Bridget's position of type KP_M-t here yields an unproblematic prospective advisory message, SAGE and KP_M still explain the philosophical verdict.

Indeed, we can push the matter of the "timing" of Bridget's relevant mental states even further. Consider another variation on the original case:

> Bridget enters a factory and sees an assembly line on which there are a number of widgets that appear red. Being appeared to red-widgetly, Bridget believes that there are red widgets on the assembly line. Several weeks ago Bridget's friend Chet, who works at the factory, told Bridget that the widgets really are red, and that whenever people visit, the superintendent likes to play an epistemic practical joke by saying that the assembly line is irradiated by red light (which it's actually not). And indeed, several hours ago, the shop superintendent told Bridget that the widgets are irradiated by an intricate set of red lights, which allow the detection of hairline cracks otherwise invisible to the naked eye. Currently, Bridget isn't thinking at all about what the superintendent said. And although at the time she spoke with the superintendent, she recalled her conversation with Chet and so didn't believe the superintendent, now she's forgotten all about what

> Chet told her: If she were to recall her conversation with the superintendent right now, she would believe that what he said was true.

Based on my reading of the literature, even with all the added layers here, hardly any philosopher would agree that Bridget's belief that the widgets are red constitutes knowledge. SAGE again explains the verdict, since Bridget can't *easily* recall her representation of the superintendent's testimony as not a reason to believe the widget aren't red—at least not as easily as her memory of the testimony itself. Thus, the following message is based on an interpretation of Bridget's position of type KP_M (with a level of recall effort *between* zero and the limit): "I would advise anyone who (1) earlier heard person S who's in charge of a set of objects tell them that the objects are illuminated by red lights, (2) now has a visual appearance of the objects as being red, and (3) doesn't have a representation of what S said as a reason for or against believing that the objects are actually red, to believe that the objects are red". Since conveying this message would indicate that we ourselves have imprudent belief-forming habits, SAGE again explains the philosophical verdict on Bridget—even though the prospective messages based on the interpretations of type KP_M-t *and* KP_M-h in this case are unproblematic.

Let's consider one final variation on Bridget's case, to confirm that SAGE and KP_M can indeed explain the philosophical verdict even in rather fanciful cases:

> Bridget enters a factory and sees an assembly line on which there are a number of widgets that appear red. Being appeared to red-widgetly, Bridget believes that there are red widgets on the assembly line. Several weeks ago Bridget's friend Chet, who works at the factory, told Bridget that the widgets really are red, and that whenever people visit, the superintendent likes to play an epistemic practical joke by saying that the assembly line is irradiated by red light (which it's actually not). And indeed, several hours ago, the shop superintendent told Bridget that the widgets are irradiated by an intricate set of red lights, which allow the detection of hairline cracks otherwise invisible to the naked eye. But by that time, Bridget had forgotten all about her conversation with Chet, and so believed what the superintendent told her. And currently, Bridget has forgotten about *both* conversations. (She gets very forgetful around red widgets.)

Again, I expect that nearly every philosopher will deny that Bridget knows the widgets are red (regardless of whether we suppose that they are in fact red). SAGE can explain the verdict because the following message is based on an interpretation of Bridget's position of type KP_M (albeit neither of KP_M-t nor of KP_M-h, the two "extremes") and is undesirable to convey: "I would advise anyone who (1) earlier heard person S who's in charge of a set of objects say that the objects are illuminated by red lights, (2) now has a visual appearance of the objects as being red, and (3) has a representation of what S said as a reason against believing that the object are red, to believe that the objects are red". In this case, it's clear that the interpretations of Bridget's position of type KP_M-t and KP_M-h yield an unproblematic prospective advisory message advice. Yet on the hypothesis that there's a natural interpretation of K's position corresponding to *any* level of recall effort, SAGE can still explain the verdict on Bridget's belief in this variation.

Rebutting and undercutting mental-state defeaters are implicated in the majority of cases in the literature where philosophers deny that a believer knows, and where that judgment seems to be independent of facts beyond the believer's head. I won't consider any further "straightforward" examples of rebutting or undercutting defeaters (although I'll soon consider some more "complicated" cases of defeat) since SAGE and KP_M offers analogous explanations for all such cases.

3.2.2 A Difference of Representation

In each of the Mary and Bridget cases, an undesirable prospective advisory message results by virtue of the perceptions and memories of the agents, and not specifically because of their reason-representations of those perceptions and memories. But there are other cases where it's undesirable to convey the prospective advisory message specifically *because* of K's reason-representations (or lack thereof) of some of her other relevant mental states. (These cases are structurally similar to the second variation of Sahib's case in § 3.1.) For instance, consider the following variation on Mary's case:

> Mary sees in the distance what appears to be a sheep in the field and forms the belief that there is a sheep in the field. The owner of the field then comes by and tells her that what she's looking at is indeed a sheep. The owner looks and sounds sincere. But Mary has an irrational prejudice against farmers (she used to watch a television show depicting a deceptive farmer, and she generalized this trait to all real farmers), and she thereby comes to suspect that the owner is lying to her. Yet she continues believing that there's a sheep in the field.

If my reading of the literature is correct, many (if not all) philosophers would reject that Mary's belief constitutes knowledge.[58] SAGE explains this verdict, since the following message is undesirable and is based on an interpretation of Mary's position of type KP_M (where one of Mary's mental-state facts is her representation of the owner's testimony): "I would advise anyone with a somewhat hazy visual perception as of a sheep in the distance, a memory of the person who oversees the area saying that what they see is indeed a sheep, and a representation of that memory as being a reason *not* to believe that what they see is a sheep, to continue to believe that there's a sheep where they're looking". Why is this message undesirable to convey? It's plausible that when people represent a piece of testimony that *p* as a reason *not* to believe that *p*, typically they have *good* reasons of this representation. For

[58] For instance, Horowitz (2013), among several other epistemologists, argues that if a believer takes herself to have strong reason not to believe that *p*, then even if from *our* perspective she's mistaken about her reasons (that is, she really *does* have overall good reason to believe that *p*), her belief that *p* is thereby not knowledge.

instance, they have memories and beliefs—even if not as easily accessible—indicating that their would-be informant is trying to deceive them.[59]

Another example of the "K vs. Us" phenomenon is the following case from Turri (2010):

> Imagine two jurors, Miss Knowit and Miss Not, deliberating about the case of Mr. Mansour. Both jurors have paid close attention throughout the trial. As a result, both have good reason to believe that Mansour is guilty. Each juror goes on to form the belief that Mansour is guilty, which he in fact is. Miss Knowit believes he's guilty because of the evidence presented during the trial. Miss Not believes he's guilty because he looks suspicious. (Turri 2010: 312)

Turri judges that "Miss Knowit knows that Mansour is guilty; Miss Not does not"—and I expect that nearly every philosopher would say the same. SAGE explains the verdict about Miss Not, because the following is based on an interpretation of type KP_M (where Miss Not's mental-state facts include her apparent representations of the evidence and of Mansour's appearance): "I would advise anyone with memories of a prosecutor presenting various pieces of evidence, no representation of these memories as reasons to believe in the defendant's guilt, and a visual perception as of the defendant looking suspicious, to believe that the defendant is guilty". Surely none of us would want to convey this message; and since it's based on an interpretation of type KP_M of Miss Not's position, SAGE offers an explanation of the philosophical verdict. The explanation hinges on Miss Not *not* representing certain of her mental states as reasons for her belief.

[59] On the other hand, it appears that at least some philosophers would insist that Mary's belief *does* constitute knowledge. For instance, Alston (2002) argues that a belief must be justified in order to defeat the positive epistemic status of another belief (whereas Mary's belief that the farmer is untrustworthy is unjustified). If these philosophers would indeed have an initial intuition (and not merely a considered, reflective opinion) that Mary's belief constitutes knowledge, then there are several possible SAGE-based explanations for this. For instance, it may be that some philosophers, upon cognitively processing Mary's case, would find the *irrationality* of Mary's reason-representation of the farmer's testimony so salient that, for them, there would be no salient interpretation of Mary's position which includes the reason-representation but not some indication of its irrational provenance. If so, then SAGE predicts that these philosophers will ascribe knowledge to Mary (since they won't feel much if any aversion to conveying, "I would advise anyone in a similar position to do just as Mary did").

What about Miss Knowit? For her, it appears the only interpretations of type KP_M yield unproblematic advisory messages (assuming, as the case suggests, that she represents her memories of the evidence as strong reasons to believe Mansour is guilty): “If you have memories of a prosecutor presenting various pieces of evidence, and representation of these memoires as providing strong reason to believe the defendant is guilty, believe the defendant is guilty”. Thus, SAGE explains why philosophers ascribe knowledge to Miss Knowit: They don’t perceive any problem, implicitly or otherwise, in conveying the message “I would advise anyone in a similar position to do as Miss Knowit did”, since all salient interpretations of Miss Knowit’s position yield an unproblematic advisory message. What if we suppose instead that the evidence presented by the prosecutor was very weak (say, a single piece of eyewitness testimony from someone who’s clearly an enemy of Mansour), yet Miss Knowit still represented it as a strong reason to believe in Mansour’s guilt? Then there would be an interpretation of type KP_M-h which includes her perceptions of this evidence (at the time it was presented), and thereby a problematic interpretation of “I would advise others to do as Miss Knowit did”: “I would advise anyone who heard a single eyewitness who’s clearly an enemy of the defendant testify that the defendant committed a crime, and who represents this testimony as a sufficient reason to believe the defendant is guilty, to believe the defendant is guilty”.

Before we move on: I’ve assumed, from Turri’s use of “because”, that Miss Not does not represent the prosecutor’s evidence as providing any reason to believe in Mansour’s guilt. We might wonder whether Turri instead meant (or at least didn’t mean to rule out) that Miss Not *does* represent the evidence as sufficient reason to believe that Mansour is guilty, but that her perception of him looking guilty was what *caused* her belief, and that the prosecutor's evidence did nothing to strengthen it. On this assumption, all interpretations of Miss Not's position of type KP_M yield an unproblematic prospective advisory message, since they presumably all include her memory of representing the prosecutor's evidence as sufficient reason for her belief. However, I suspect that most philosophers would no longer deny that Miss Not's belief qualifies as knowledge in this case. At least speaking for myself, so long as we suppose that Miss Not appreciates the full reason-giving force of the prosecutor's evidence, I feel little reluctance to say that she knows. Lehrer (1971) offers a similar case of a believer whose belief is initially caused by bad reasons (that is, bad from our perspective as observers) but who later acquires, and appreciates the force of, objectively good reasons for the same belief. Lehrer judges that this believer now has knowledge, even supposing that his belief isn't strengthened by the later evidence.[60]

Here's a follow-up example from Turri (2010), which again demonstrates the importance of K's reason-representations (or lack thereof) for SAGE's explanations of philosophical verdicts:

> Consider two of the other jurors, Miss Proper and Miss Improper, sitting in judgment of Mr. Mansour. Each paid close attention throughout the trial. As a result, each knows the following things:
>
> (P1) Mansour had a motive to kill the victim.
> (P2) Mansour had previously threatened to kill the victim.
> (P3) Multiple eyewitnesses place Mansour at the crime scene.
> (P4) Mansour's fingerprints were all over the murder weapon.
>
> … Miss Proper reasons like so: "(P1-P4) make it overwhelmingly likely that Mansour is guilty. (P1-P4) are true. Therefore, Mansour is guilty." Miss Improper, by contrast, reasons like this:

[60] What should we say if it turns out that not all philosophers agree with Lehrer and me about the cases in question? Perhaps there are interpretations of K's position which are limited to those mental states which *caused* her belief—interpretations which are more salient to some of us than to others (due to various differences in our psychology). I won't pursue this idea as an official hypothesis; but I do want to emphasize that SAGE is likely capable of explaining judgments different from the ones I've assumed above, with suitable additional hypotheses about natural interpretations (at least for some speakers) of K's positions.

> "The tea leaves say that (P1-P4) make it overwhelmingly likely that Mansour is guilty. (P1-P4) are true. Therefore, Mansour is guilty." (ibid., pp. 315-16)

Turri judges that Miss Proper has knowledge while Miss Improper does not, and I believe nearly any philosopher would agree. And indeed, there's an interpretation of type KP_M of Miss Improper's position which yields an undesirable advisory message (assuming, as the case suggests, that she doesn't represent the *content* of P1-P4 as providing reason for believing in Mansour's guilt): "I would advise anyone who believe that a particular set of tea leaves indicate that a certain set of facts make it overwhelmingly likely that a particular defendant is guilty, and who represents this latter belief as sufficient reason to believe that the defendant is guilty, to believe that the defendant is guilty". (Miss Improper's memories of the *exact content* of P1-P4 are surely not *as* easily recallable as her memory of the tea leaves indicating that P1-P4 make it overwhelmingly likely that the defendant is guilty.)[61]

3.2.3 Peer Disagreement and "Higher-Order" Evidence

The phenomenon of "peer disagreement" has provoked a good deal of debate among epistemologists in recent years, and it offers a good further test of whether SAGE can explain philosophers' judgments in more "complex" cases of mental-state defeat than those considered above. (We'll also again see the importance of distinguishing between KP_M-t and KP_M-h for SAGE's ability to explain certain philosophical verdicts.) Peer disagreement is disagreement over the truth of a proposition with someone whom you take to be just as good as you at evaluating evidence, and where each of you has evaluated the same body of evidence in arriving at your respective belief. (Arguably, most cases of peer disagreement involve a

[61] SAGE's explanation for the universal ascription of knowledge to Miss Proper is analogous to the explanation given above for the verdict on Miss Knowit.

candidate *rebutting* defeater, where you believe that *p* and then receive evidence which, taken by itself, supports the belief that ~*p*.) For instance, consider the following case from Lasonen-Aarnio (2014):

> My friend and I have often amused ourselves by solving little math problems in our heads, and comparing our answers. We have strikingly similar track records: we are both very reliable at doing mental math, and neither is more reliable than the other. We now engage in this pastime, and I come up with an answer to a problem, 457. I then learn that my friend came up with a different answer, 459. (p. 315)

Let's call the narrator of this scenario "Maria," and let's suppose that she retains her belief that the correct answer is 457 when she learns of her friend's answer. Let's also suppose that Maria is right, and that the mental process by which she arrived at her answer was impeccable, such that any of us would agree that Maria knew the answer prior to hearing from her friend. Even still, does Maria continue to know that 457 is the correct answer once she hears from her friend? Many epistemologists would say no on both counts. There's a popular view, sometimes called "conciliationism," that peer disagreement generally calls for suspension of judgment, at least pending further investigation. (Christensen 2007 and Elga 2007 are notable defenders of this position.)[62] On the other hand, some authors have argued that beliefs retained in the face of peer disagreement generally continue to count as knowledge so long as they counted as knowledge before the disagreement. (In particular, see Kelly 2005 and Lasonen-Aarnio 2010 and 2013.)

[62] To be more accurate, most of the literature on peer disagreement is framed in terms of "credences," or degrees of belief. Accordingly, most conciliationists argue that peer disagreement (where two peers arrive at significantly different credences in a given proposition) generally calls for a significant credal revision in the direction of the credence reported by one's peer. However, I think it's safe to assume that most or all of these authors would further agree that in cases of disagreement involving all-out belief, both peers generally ought to suspend judgment—and therefore that at least for the time being, both peers cease to have any knowledge which they did previously, even if they retain their belief.

But I find that I typically don't have strong "knowledge intuitions" about cases like Maria's until I assume more details about the agent's reasoning process than is typically offered by authors who write about peer disagreement. For instance, if I imagine that Maria's reasoning process was very involved and lengthy, then I'm inclined to think that she no longer *knows* that the answer is 457 at the moment she learns of her peer's disagreement (that is, pending further investigation on her part, or at least a recitation of her reasoning). But if I suppose that Maria's reasoning process was short and simple (say, if the problem was "What is 437 plus 20?"), then I'm quite willing to say that she continues to know. From what I can tell, epistemologists haven't discussed this "length of initial reasoning" consideration in the literature on peer disagreement. Thus, I can't say for sure that my intuitions about Maria's case are representative of philosophers' intuitions generally. But several philosophers have indicated their agreement with me in person, so I'll tentatively assume that my intuitions are widely shared.

And happily, SAGE easily explains my intuitions about each of the above two versions of Maria's case. First, suppose that Maria's reasoning process was complicated and lengthy—let's say that the problem was to find the square root of 208849, which Maria didn't have memorized—and that all she consciously remembers at time t (when she learns of her peer's disagreement) is that each step of her reasoning, which led to her answer of 457, felt sound to her at the time. Then according to hypothesis KP_M, the following is a natural implication of "I would advise anyone in a similar position to do as Maria did", where the interpretation of Maria's position is of type KP_M-t and so restricted to her mental states *conscious at t*: "I would advise anyone with a memory that each step of their mental reasoning towards the answer of a fairly difficult arithmetic problem felt sound, and who hears their peer—whom they believe has a similar track record on problems of this sort—report a similar but distinct answer, to retain their belief about what the correct answer is (and to not, say, engage in any double-checking)". I suspect most of us wouldn't want to convey this message, since it seems that beliefs retained in such positions will be false about 50% of the time. (But perhaps some of us—perhaps those who feel especially confident about their arithmetical abilities—*would* feel comfortable conveying this advisor message. And if so, SAGE could explain why some people might intuitively judge that Maria has knowledge in this case.) Note that there's another natural interpretation of type KP_M of Maria's position which yields an unproblematic advisory message—namely, the interpretation of type KP_M-h, which includes *all* of her memories of the exact steps of her reasoning process the past. But SAGE still explains the judgment that Maria doesn't know, on the hypothesis that there's always a potentially salient interpretation of K's position which is restricted to K's mental states *conscious at t*.

But now suppose that Maria's reasoning process was short and simple, and that the full process is still in her working memory when she learns of the disagreement (or that she can recalls it immediately thereafter in full detail). Say, for instance, that the given problem was to find the sum of 437 and 20. Then, it seems that any prospective advisory message based on an interpretation of type KP_M goes something like this: "I would advise anyone who remembers adding a particular three-digit number to a particular two-digit number, who feels that the answer is utterly, and who hears their peer—whom they believe has a similar track record on problems of this sort—report a similar but distinct answer, to retain their belief about what the correct sum is". I wouldn't feel much hesitation to convey this message, and I doubt many other philosophers would either. Thus, SAGE explains (what I take to be) our verdict that in *this* case, Maria retains her knowledge.

There are variations on the above two "archetype" cases (really hard vs. really easy reasoning process)—variations where philosophers' knowledge intuitions might sway to one side or the other. But I'm reasonably confident that SAGE will sway with them, including when intuitions are ambivalent. To take but one example, suppose that Maria's reasoning process was moderately long and complex, but she's checked it over numerous times in her head, and she has a memory of a near-perfect track record with problems of similar difficulty when she's checked over her answer a similar number of times. Then, I'm fairly inclined to say she'd continue to have knowledge upon learning of her peer's disagreement. But assuming that one of her conscious mental states is a memory of checking over the answer multiple times, it's plausible that all interpretations of her position of type KP_M yield an unproblematic advisory message—including those restricted to her mental states at t. I could multiply the examples here, but I think we can tentatively conclude that in the special case of mental-state defeat (or at least *candidate* defeat) involving peer disagreement, SAGE can explain each philosophical verdict. Moreover, I think SAGE helps us at least partially diagnose the continuing "battle of intuitions" among epistemologists over cases of peer disagreement: In the hypothetical vignettes offered in the literature, there's typically an omission of the precise length and complexity of the target agent's reasoning process—information which is very important (if SAGE is correct) to whether we'll be willing to ascribe knowledge. If these details are omitted, different philosophers may implicitly "fill them in" in different ways.

3.2.4 Cognitive Limitations and Expertise

Another interesting class of cases where philosophers deny knowledge involves agents who seem to conclude “too much” from their experiences, given their cognitive limitations. Such cases have been discussed at length by epistemologists. Here is a version of the famous “specked hen” case, apparently first offered by Chisholm (1942):

> Ken is visiting a friend’s farm for the first time, and comes across a cage full of hens. The first hen which Ken sees has three speckles, and Ken forms the belief, “That hen has three speckles.” The second hen which Ken sees has 48 speckles, and without counting each speckle individually, Ken immediately comes to believe, “That hen has 48 speckles.” Ken is very confident about both beliefs, although he has no memory of having any ability to accurately number a large quantity of items in his visual field.

So long as we suppose that Ken’s visual capacities are just like ours, then based on my reading of the literature, every philosopher will deny that Ken’s belief about the 48-speckled hen constitutes knowledge. But even if we suppose that Ken really does represent his visual impression as decisively supporting a belief that the hen has 48 speckles, SAGE can explain the philosophical verdict. For the following is a natural implication of “I would advise anyone in a similar position to believe as Ken did” based on an interpretation of type KP_M-t of Ken’s position: “I would advise anyone who has a single brief visual impression of a hen which they regard as decisively supporting a belief that the hen has exactly 48 speckles to believe that the hen has exactly 48 speckles”. This is surely a message we wouldn’t want to convey (at least not to others who share our visual capacities), so SAGE explains the philosophical verdict: The interpretation of Ken’s position of type KP_M-t becomes salient, and philosophers thereby register an aversion to implying, “I would advise anyone in Ken’s position to do as he did”.

Some cognitive limitations are due not to intrinsic features of our cognitive apparatus, but to lack of appropriate practice or training. Here is a case, based on one from Feldman (2003, p. 75), which highlights the expert/novice differential:

> Bert and Brent are bird-watching in a forest. Bert is an expert bird-watcher, with many years of training and experience, while Brent is his new apprentice, out for his first bird-watching trip. Suddenly, a pink-spotted flycatcher flies across the path in front of them. Bert, having seen many flycatchers in photos and in person, immediately comes to believe, “That’s a pink-spotted flycatcher.” Brent has never seen a flycatcher or an image of one. But he perceives a faint pink streak on the bird’s breast, and “pink-spotted flycatcher” is the only bird name he’s heard with the word “pink,” so he comes to form the same belief as Bert.

Based on the literature, it seems every philosopher would agree that Bert’s belief constitutes knowledge, while Brent’s belief does not. SAGE offers an explanation of both verdicts. Here, for instance, is a prospective advisory message based on an interpretation of type KP_M-h of Brent’s position: “I would advise anyone who has a visual impression of a bird with some pink on its feathers, a memory that ‘pink-spotted flycatcher’ is the name of a type of bird, and no memories of correctly identifying any other pink-spotted flycatchers (nor of seeing images or reading visual descriptions of them) to be certain that the bird they’ve just seen is a pink-spotted flycatcher.” This interpretation of Brent’s “position” includes a relevant fact about his mental-state history: He’s never experienced correctly identifying a pink-spotted flycatcher, nor has he seen images of one. Thus, even though Bert’s and Brent’s conscious mental states *at t* are identical, their mental-state histories have relevant differences. It’s those historical differences which SAGE points to as the explanation for philosophers’ different verdicts on Bert and Brent.

3.2.5 Risk and Lotteries

Neither Ken’s nor Brent’s case seemed to involve any explicit mental-state defeaters, in the sense defined in § 3.2.1. There’s another class of cases, involving so-called “aggregation of risk”, which also feature no mental-state defeaters and yet in which a good proportion of philosophers deny that the agent knows. Makinson (1965) offered the first such example, and I've based the following vignette on his:

> Effie, a mathematician, has written a draft of a new textbook with several hundred mathematical claims (counting all the intermediate claims included in the proofs), some of them original. Before sending it to a publisher, Effie read through the entire textbook once, looking for any errors or typos. She found none, and upon reading each claim, she felt confident that it was true. When she sends the draft to her publisher, she attaches a note saying, "I'm confident that all of the claims in this textbook are true".

Let's suppose that in fact all of the claims in Effie's new textbook are true, and that her evidence for each is so good that no philosopher would deny, of any of her beliefs in the individual claims, that it constitutes knowledge. Still, many philosophers have expressed the intuition that an "aggregate" belief like Effie's ("All of the claims in the book are true") typically fails to constitute knowledge, especially when the number of constituent claims is very large. For instance, Makinson himself judges that such aggregate beliefs fail to constitute knowledge, and Hawthorne and Lasonen-Aarnio (2009) find it implausible to say that someone can know the conjunction of all the high-chance propositions they know about the future (e.g., "My house will still be standing tomorrow", "My car will work tomorrow", "My alarm clock will be working tomorrow morning", etc.). Williamson (2009) and Smith (2010) offer some theoretical arguments in favor of granting that aggregate beliefs *do* constitute knowledge when all the constituent claims are known. However, both Williamson and Smith concede that this position runs counter to intuition, and each of them seems to be motivated by a desire to maintain the elegance and simplicity of a favored theory of justification or knowledge. Thus, I think it's reasonable to say that nearly all (if not all) philosophers would deny that aggregate beliefs like

Effie's constitute knowledge—especially when they approach the cases "intuitively", rather than reflectively.[63]

SAGE can explain the philosophical (near-)consensus. Consider the following prospective advisory message, based on an interpretation of type KP_M-t of Effie's position: "I would advise anyone who has a memory of checking carefully through each of several hundred claims in an advanced mathematical text only once, and not finding any errors, to believe that all of the claims are true." This is arguably not a message we'd want to convey: We ourselves wouldn't be certain that "all the claims are true" if our position were exactly as stated (especially given the stipulation that the claims involve advanced mathematical content). Plausibly, our trust heuristics evolved to prevent certainty in such aggregate beliefs: Those of our ancestors who acted as though there were no chance of falsity of the conjunction of many individually likely propositions (for instance, "We'll easily find enough food every day for the next twenty days", "No one in the community will fall ill in the next twenty days") were almost surely at a disadvantage.[64] But now suppose that Effie reads through her draft several more times, and receives feedback from several expert reviewers, none of whom finds a single error. Then, I suspect most or all philosophers would begin to agree that Effie knows that all of the claims in the book are true. (Remember our supposition that all of the claims are in fact true.) And

[63] Note that SAGE is compatible with us sometimes consciously "overriding" the verdict which our trust heuristic automatically yields when we simulate being in a given "position". SAGE predicts that *typically*, our knowledge judgments are mediated by our sense of whether we want to convey the message "I would advise anyone in a similar position to do as K did" and that *typically* this sense is mediated by our own trust heuristic (see § 2.5). I'll return to the possibility of "overriding" in Chapter 5.

[64] Meanwhile, SAGE explains philosophers' knowledge ascriptions for each of Effie's *individual* beliefs, since the following prospective advisory message seems unproblematic: "I would advise anyone with a memory of proofreading once carefully through a single claim in an advanced mathematical text, and finding no errors, to continue to believe that the claim is true."

indeed, it now appears that all interpretations of type KP_M of Effie's position yield an unproblematic advisory message, even the interpretation of type KP_M-t: "I would advise anyone with a memory of checking through each of the claims in a mathematical text multiple times and finding no errors, and memories of several experts telling them that they similarly found no errors, to believe that all the claims are true".

In Effie's case and others like it, it seems that from the agent's conscious perspective at the relevant time, there's a significant likelihood that the aggregate claim (e.g., "All the claims in the book are true") is false. However, there's another sort of case where the target claim seems to have *no* significant likelihood of being false, and yet where philosopher still deny that the claim can be known. I'm referring to the infamous "lottery" cases, like the following:

> Terry buys a ticket in a million-ticket lottery. She's aware of the size of the lottery, and that the winner will be chosen by a random-number generator. On the basis of the odds, Terry comes to believe that her ticket will lose. And in fact, her ticket is a loser.

Based on the recent literature, it appears that nearly all philosophers judge that Terry's belief does not constitute knowledge. (Hawthorne 2004 is a recent influential study which presupposes that lottery beliefs don't qualify as knowledge, and considers the implications of putative fact.) SAGE offers a clear explanation for the verdict, since the following is a prospective advisory message based on an interpretation of type KP_M of Terry's position: " I would advise anyone who recalls buying a lottery ticket in a large lottery, who's calculated that the odds of winning are extremely low, but who hasn't yet read or heard the winning number, to be certain that their ticket will lose". It seems that very few if any of us would be *certain* in a position like Terry's (as stated). Thus, if philosophers recognize that *they* wouldn't be certain that their ticket would lose in a position like Terry's, they almost certainly won't want to convey that they would advise *others* to be certain. SAGE thereby explains the robust philosophical judgment that pre-draw lottery beliefs don't constitute knowledge.[65]

3.2.6 Maintenance Problem?

I've now covered a representative sample of cases in the literature where a significant number of philosophers deny that the believer has knowledge, and where the verdict appears to be independent of facts other than those about the agent's mental states. I've demonstrated that SAGE, along with my hypothesis KP_M, offers an explanation for the philosophical verdict on each case. I think we have good inductive grounds, then, to suppose that SAGE and KP_M can explain the verdict on *all* such cases—both existing cases that I haven't covered here, and any cases which may be proposed in the future. Of course, I can't absolutely rule out that there will be (or is) some case for which SAGE and KP_M cannot explain the philosophical verdict. But for now, I suggest we at least tentatively accept that my hypotheses suffice to explain every verdict.[66]

Of course, a prerequisite for hypothesis H to explain a given observation O is for H not to predict ~O. I spent some time above demonstrating that SAGE and KP_M do not predict that philosopher's *won't* ascribe to knowledge to such agents as Miss Knowit and Miss Proper (§ 3.2.2), short-reasoning-process Maria (§ 3.2.3), and Bert (§ 3.2.4)—all agents to whom philosophers clearly do or would ascribe knowledge. Thus, I think we also have inductive grounds for supposing that SAGE and KP_M do not predict a significant rate of knowledge denial

[65] To be sure, the fact that we don't feel certain that our lottery tickets will lose—despite the long odds—has puzzled philosophers, since we often *are* certain of propositions (e.g., "My alarm clock will be working tomorrow morning") which seem to have a much higher likelihood of being false. (See, for instance, Hawthorne 2004.) In Chapter 5, I'll show how THYME can explain this puzzling state of affairs.

[66] Recall from Chapter 2 that by "explain" and "offer to explain", I mean simply that there's a physically possible casual pathway between the hypotheses in question (or rather, the putative facts therein) and the observations. Only in Chapter 6 will I argue that we should accept that the explanations offered by SAGE and KP_M are in fact *accurate*.

in cases where philosophers unanimously ascribe knowledge. However, I want to close by reviewing one class of such cases which my reader might have worried about: cases where a believer *maintains* or *retrieves* a belief she formed earlier, but without also maintaining or retrieving the mental states which she initially represented as reasons for her belief. For instance, consider a variation on the case of Mary (§ 3.2.1):

> Mary was told several minutes ago by the owner of a large field that there are several sheep in the field. Now, looking out over a portion of the field, Mary doesn't see any sheep-looking objects, but she tells her son, "There are some sheep in this field—let's keep looking to see if we can find them". At the time she says she, she doesn't consciously recall the owner's earlier testimony.

The interpretation of Mary's position of type KP_M-t in this case includes only her *lack* of visual perceptions as of sheep. Thus, it might seem as though there's a natural implication of the prospective advisory message which is undesirable: "I would advise anyone who has visual impressions of only one portion of a large field, and no current visual impressions as of any sheep, to believe that there are sheep somewhere in the field." However, I doubt that this interpretation of the prospective advisory message becomes salient to us when we consider ascribing knowledge to Mary: The fact that Mary's belief is *maintained* or *retrieved*, rather than *formed*, at the time in question is plausibly a component of *any* natural interpretation of "I would advise anyone in a similar position to *do as Mary did*". [67] Thus, I suggest that the only *natural* KP_M-t implications of the prospective advisory message will be along the following lines: "I would advise anyone with visual impressions of only one portion of a large field, and no

[67] I haven't offered any official hypothesis about the natural interpretations of the *second* half of "I would advise anyone in a similar position to *do as K did*". This is because I don't think there's much that needs to be said, other than to highlight the forming/maintaining distinction. (I'll cover one other notable issue in Chapter 4.) Otherwise, it seems to me that in each of the cases considered here and in the next chapter, it's quite intuitive what "doing the same sort of thing K did" amounts to, given a position which is plausible "similar" to K's position.

current visual impressions as of any sheep, to maintain their earlier belief that there are sheep somewhere in the field". This message doesn't seem problematic at all (given the "only one portion" qualifier): When we retrieve or maintain an earlier belief that *p*, and our current conscious mental states aren't clearly inconsistent with '*p*', we typically *do* continue to feel certain that *p*. Thus, SAGE does *not* predict that speakers who consider ascribing knowledge to someone maintaining or retrieving a "well-formed" belief will typically hesitate to ascribe knowledge—even if an interpretation of type KP_M-t of K's position becomes salient.[68]

[68] By similar reasoning, SAGE does not predict that philosophers won't ascribe knowledge that *p* to someone who earlier formed a belief that *p* consciously (on the basis of good reasons) but whose belief is now subconscious.

CHAPTER IV: Trouble from Without

4.1 Taking on the World

In Chapter 3, we saw how SAGE can explain the philosophical verdict on cases where a significant number of philosophers deny that the believer knows, and where this verdict seems to be independent of any facts other than those about the believer's mental states (past and present). At least some of these verdicts are, absent an explanation, puzzling—particularly when, as for instance with "Miss Not" (§ 3.2.2), the believer has *objectively* good evidence to believe as she does. (We saw that the crucial variable in these cases was the believer's own reason-representations of the objectively good evidence.) But ever since Edmund Gettier published his seminal paper, there's been a proliferation of cases in the literature where many or all philosophers deny knowledge, and where this verdict depends entirely on facts *outside of* the believer's awareness. Moreover, these cases don't involve false belief: Gettier's original two cases, and the similar ones that have followed, all involve justified true belief.[69] The existence of cases of justified true belief where many of us deny knowledge has remained a puzzle in epistemology. (After all, what more could knowledge ask for?)

[69] Recall the minimal sense of "justification" that I'm using here: K's belief that *p* is justified just in case most or all of us would also have been certain that *p* had we been certain of all the same propositions relevant to whether or not *p* as K was certain (at least implicitly) at the time of her belief.

In this chapter, I'll argue that SAGE can explain the philosophical verdict on each of these "Gettier-type" cases: cases where a significant number of philosophers deny knowledge, and where the verdict appears to depend entirely on facts of which the believer isn't aware. At the end of the chapter, I'll show how SAGE can further explain the appeal of several of the diverse "conditions for knowledge" which philosophers have proposed since Gettier's paper. SAGE can also explain each condition's relative degree of success in aligning with (although arguably not *explaining*) philosophers' verdicts. (Perhaps most relevantly, I'll show how SAGE overlaps with—but still differs importantly from—the "safety" condition which many epistemologists have endorsed in recent years.)

To begin, here's a reminder of SAGE and KP$_M$, my first hypothesis about the natural interpretations of K's "position":

> (SAGE) Speakers today at least implicitly expect their listeners to interpret "K knows that *p* at t" to convey the prospective advisory message "I would advise anyone in a position similar to K's at t to do the same sort of thing K did in believing that *p*".
>
> (KP$_M$) Given a believer K and her belief that *p* at time t, given a speaker S aware of various facts about K's situation, and given any level of recall effort r, there is a natural interpretation for S of K's "position" which includes all reason-relevant facts (of which S is aware) about the set of K's mental states at t which appear to S to be within level r.

To see that we still have work to do despite our exertions in Chapter 3, let's consider the following case, which I've based on a now-famous case proposed by Bertrand Russell (1948):

> Russell descends the stairs of his house at 7:34am—not having checked the time when he woke up—and sees the face of his old reliable grandfather clock reading 7:34. Russell comes to believe, correctly, that the time is 7:34am. However, he doesn't realize that the grandfather clock stopped working last night. In fact, the clock stopped just at 7:34pm.

To my awareness, every philosopher since Russell himself has denied that believers in "stopped clock" cases of this sort know the time. But it seems that every interpretation of Russell's position of type KP_M yields an unproblematic interpretation of the prospective advisory message posited by SAGE. For instance, the interpretation of type KP_M-t yields the following implication: "I would advise anyone who has a visual appearance of a clock reading 7:34, visual appearances consistent with early morning sunlight, and a representation of their visual appearance of the clock as a decisive reason to believe that it's exactly 7:34am, to believe that it's 7:34am".

If KP_M were the full story about the natural interpretations of K's "position", then SAGE clearly would be at a loss to explain philosophers' denial that Russell knows the time. But in fact, I don't believe KP_M is the full story. Indeed, it seems plausible in the abstract that when we hear reference to someone else's "position", we're at least initially open (that is, pending qualification by the speaker) to interpretations which involve facts of which that person wasn't aware. That is, the "position" in question might be the agent's position *in the world*, and not limited to her *perspective* on the world. To confirm this suspicion about the natural interpretations of "position", consider the following vignette:

> Travis is visiting Washington, D.C. from out of town, and he's just disembarked a Metro train at Smithsonian Station. Travis intends to visit the National Museum of African Art, but he's not sure how to get there. On his way out of the station, he sees a woman wearing a "Tourist Ambassador" shirt and holding maps. Travis asks her how to reach the Museum, and she replies, "Head east on Independence Avenue after exiting the station". Travis thanks her and proceeds to head east on Independence. What he doesn't realize is that this woman is trying to play practical jokes on unsuspecting tourists by giving them bad directions. And indeed, she randomly chose a street and a direction when she told Travis, "Head east on Independence". But in fact, unbeknownst to this woman, the Museum of African Art *is* east on Independence from the Metro station.

I would certainly hesitate to say, "I would advise anyone in a similar position to do just what Travis did in following the directions he was given"—at least without qualifying what I meant.

But my reluctance would be needless if I expected my listeners to interpret Travis's "position" exclusively in terms of his mental states. After all, I wouldn't hesitate to convey the following: "I would advise anyone who hears someone who is by all appearances a helpful authority say to take directions D to location L, and who wants to get to L, to take directions D". Travis's case is evidence, then, that there are natural interpretations of K's "position" which outstrip K's perspective. In fact, it seems plausible that in any given case, there's a natural interpretation of K's position which includes *all* reason-relevant facts about K's situation, whether or not K was aware of each of them. That is, the interpretation includes all facts which, by our lights as *observers*, provided reason for or against K doing what she did, either by themselves or conditional on one or more other facts.[70]

Why might such an all-inclusive interpretation of K's position always be natural—that is, a candidate for salience for speakers who are considering referring to K's position either directly or indirectly? This might have to do with the fitness-related advantages for our ancestors of easily forming cognitive representations of some third-party agent which included not only those relevant facts of which the agent was aware, but in addition all relevant facts of which the observer was aware as well. By having such an all-inclusive representation "at the ready", our ancestors would have been more easily able to (say) explain and/or predict the consequences of the agent's actions, and also bring the agent "up to date" if she subsequently

[70] That is, a fact might be "relevant" even if it doesn't seem to be a reason *all by itself* for or against the action, but rather is a reason (say) against the action conditional on some other fact which *is* a direct reason for the action. For instance: Suppose that Brendan tells me that there's currently free food on the top floor of our apartment building. But just before I venture upstairs, Varun tells me that Brendan has recently been playing a whole bunch of practical jokes. Varun's testimony seems to count (at least somewhat) against me taking the trouble to go upstairs *given* Brendan's testimony, even though Varun's testimony taken by itself counts neither for nor against going upstairs.

came within hearing distance. At any rate, if the Travis example is any indication, it does appear that we spontaneously consider all-inclusive interpretations of K's "position".

But now Travis's example raises a question: Once we take into account *all* the reason-relevant facts about Travis's situation, it clearly *is* a good idea for him to follow the directions he was given. Thus, the hesitation I (and I suspect my readers also) feel to say "I would advise anyone in a similar position to do just as Travis did" must have something to do with the word "similar". What I now want to suggest is that there's a natural interpretation of "similar position" on which "Tracy", described below, is in a position similar to Travis's:

> Tracy is visiting Washington, D.C. from out of town, and she's just disembarked a Metro train at Smithsonian Station. Tracy intends to visit the National Museum of African Art, but she's not sure how to get there. On her way out of the station, she sees a woman wearing a "Tourist Ambassador" shirt and holding maps. Tracy asks her how to reach the Museum, and she replies, "Head west on Independence Avenue after exiting the station". Tracy thanks her and proceeds to head west on Independence. What she doesn't realize is that this woman is trying to play practical jokes on unsuspecting tourists by giving them bad directions. And indeed, the woman randomly chose a street and a direction when she told Tracy, "Head west on Independence". In fact, the Museum of African Art is *east* on Independence from the Metro station.

I do think it's intuitively quite natural to say that Travis and Tracy are in similar positions, despite one getting good directions and the other being misled. Why might this be? Notice that to "get" from Travis's case to Tracy's case, all we need to do is make one small change in the relevant facts: Have the putative "Tourist Ambassador" randomly choose to say "west" instead of "east".

Travis's case, its apparent similarity to Tracy's case, and the observation about "getting" from Travis to Tracy with one small change of relevant fact motivate the following two hypotheses (which are specifically about *believers*, but of course can be generalized to agents taking physical actions as well):

> (KP_W) Given a believer K and her belief that *p* at time t, and given a speaker S aware of more facts about K's situation than K is at t, there is a natural interpretation for S of K's "position" which includes *all* reason-relevant facts of which S is aware—past, present, and future.[71]
>
> (SIM) Given an interpretation of type KP_W of K's "position" and some other agent L, it is relatively more likely, for any given speaker, that there's a natural interpretation of "a position similar to K's" on which L's position is similar to K's if:
>
> > (1) L's reason-relevant mental states (that is, the mental states relevant to L doing the same sort of thing K did in believing that *p*) differ only in details (if at all) from K's. For each of K's relevant mental states, L should at least have an analogous one.
> > (2) The differences between K's position and L's position involve only facts which are not involved in *causing* any of K's or L's relevant mental states—or, if there are differences between the facts which caused K's mental states and L's mental states, the differences are relatively "far back" in the causal chains.
> > (3) There are few and/or minor (as opposed to many and/or major) factual differences between K's total situation and L's total situation.

The idea behind KP_W should be clear already, in light of our observations about Travis's case. However, SIM of course bears some explanation.[72] First, I haven't attempted to give a definitive, exhaustive set of conditions for when there's a natural interpretation on which the positions of two agents are "similar" in the context of the message "I would advise anyone in a similar position to do just as K did in believing that *p*". Such a project would consume far more time and patience than I or my reader has—and besides, it's likely that we all have our own cognitive idiosyncrasies regarding our (implicit or explicit) similarity judgments. Thus, I think that general guidelines about what makes one agent's position *more likely* to strike any one of

[71] I've used the subscript "W" to indicate that the natural interpretations predicted by this hypothesis include all reason-relevant facts about K's "W"orld of which the speaker is aware.

[72] Readers might notice some similarities between SIM and Lewis's (1979) four rules for determining whether any two "possible worlds" (that is, fully-specified states of affairs) are similar to each other. However, while Lewis is concerned with similarity between worlds *in toto*, I'm concerned here with similarity between the positions of *agents* within a given world or worlds—and even more specifically, similarity between agents in the special context of forming a certain type of belief. Thus, while Lewis's rules clearly apply to my project—such as "It is of … importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails"—outlining similarity between the positions of particular agents requires finer-grained guidelines.

us as “similar” to another’s is a prudent middle course. As for the particular three guidelines: I think that (1) is straightforward, since the conscious perspective (past and present) of a given agent is arguably the most “central” aspect of her position on any natural understanding of “position”. Our intuitive notions of “similarity” between two agents taking or considering analogous actions A and A′ brook little if any significant difference between what they’re aware of. As for (2), I think it’s plausible in the abstract that the facts with which someone is *causally connected* are more “central” to their position than facts which (at least so far) have not physically influenced that person at all. Moreover, those facts which had a *direct* causal influence on the person seem more central than those whose causal influence was indirect (that is, mediated by other causes).[73]

As for (3), let me say something about “minor” vs. “major” differences. I do believe that our judgments here are vulnerable to context and framing. For instance, if we take a “wide angle” view of things (say, from the seat of an airplane), then the fact that Aggie is growing soybeans on her farm and that Rick is growing corn on his doesn’t seem to make for a big difference between their positions vis-à-vis being farmers. But if we’re on the ground, and we’re working for a soybean wholesaler, the difference is a clearly a bigger deal. A related point is that there might be “threshold” effects in our similarity judgments: If Aggie has already planted one acre of corn along with her twenty acres of soybeans, then it wouldn’t seem much

[73] I’ll soon offer examples to support these abstract claims. But I’ll note here that there might be a good evolutionary reason for our similarity judgments to be more sensitive to facts which causally influenced a given person than those which didn’t: There are plausibly advantages (for our human ancestors and for us) to be readily able to *explain* why a given person did what they did, or how they got to be the way they are. For this purpose, it would be advantageous to focus more closely on reason-relevant facts which causally influenced a given person than on facts (even reason-relevant ones) which had no causal influence. Moreover, given that our explanations can’t go back indefinitely in time, it would be prudent to focus more closely on *direct* causes than on causes further back in the chain.

different if subsequently she switches one of her soybeans acres to corn. But if she hasn't ever planted corn before, then adding an acre of corn might seem to bring her into a much different position (especially if we're one of those corn wholesalers out there). Thus, getting above a certain "threshold" (say, zero acres of corn to one acre) might entail an intuitively bigger difference than further increases above the threshold. A final note about "minor" vs. "major" differences: Note that minor differences can, if repeated or sustained over time, aggregate to constitute an intuitively major or significant different. For instance, if Aggie and Rick each started their farm only just this month, and neither is dead-set on continuing to plant only the one crop they've planted this season, then they might not seem so different from each other. But if they've both been at it for 30 years, and have only planted their one crop—soybeans in Aggie's case, corn in Rick's—then I think we're much more prone to regard them as very different farmers. After all, Aggie and Rick have been doing a *lot* of things differently over the years, even if each of these things (planting corn vs. planting soybeans in a given year) aren't so different from each other taken individually.

Now, with KP_W and SIM in hand, we can offer a full explanation for our hesitation to say "I would advise anyone in a similar situation to do just as Travis did": A case in which someone receives *bad* directions from a woman trying to give bad directions to tourists is, as per SIM-3, "similar" to Travis's case on a natural interpretation. Thus, when we consider saying "I would advise anyone in a similar position to do as Travis did", it's plausible that we at least implicitly recognize that there are cases—like Tracy's above—which involve only a few minor detail-changes from Travis's case, and in which we *wouldn't* advise the agent to "follow the directions". Although Travis's position and these alternative agents' positions do differ slightly in their reason-relevant mental states—that is, hearing the direction-giver say "west" rather than "east"—these mental states are caused by the *same* event: the direction-giver randomly choosing a direction to say. (The difference between the woman saying "east" and "west" requires only a small difference in her whims at the moment, but not in the fact that she's randomly choosing what to say.) SIM-2 therefore leads us to expect that we'll find Travis's position similar to Tracy's. How might it be that we, as I've put it, "implicitly recognize" the existence of cases where the agent's position is similar to Travis's, and where we wouldn't advise that agent to "follow the directions you receive"? There is strong empirical evidence that our subconscious cognitive systems have evolved to heuristically simulate situations that deviate in certain ways from situations that we've already encountered or that we've heard about—at least if we're trying to make plans for the future, or are otherwise primed to think about "similar but different" situations.[74] It seems plausible that the same "similarity simulation" heuristic activates when we consider (consciously or otherwise) conveying a message about how we'd respond to others in a position "similar" to the one under consideration.

To further motivate KP_W and SIM before we dive into the literature on epistemic (rather than physical) actions, consider the following case:

> Sean's doctor recently prescribed him an antibiotic for an infected wound. Sean picks up the medication at his local pharmacy and begins taking it as directed. Unbeknownst to Sean, one of his daily vitamin supplements counteracts the antibiotic and prevents it from combatting the infection. (Sean's doctor didn't realize that Sean was taking this particular supplement; otherwise, she would have told Sean to temporarily stop taking it.) Luckily, however, Sean's immune system turns out to be strong enough to battle the

[74] Seligman et al. 2013 review a large body of empirical evidence for this subconscious ability and proclivity to think "prospectively", both in humans and non-human mammals.

> infection on its own. Had it not, Sean might have developed a life-threatening case of sepsis.

I would certainly hesitate to say, without qualification, “I would advise anyone a similar position to do just as Sean did in taking his supplement along with the antibiotic”. Why the hesitation, if Sean turned out fine in the end? KP_W and SIM shed some light on this: Although Sean’s immune system does turn out to be strong enough to battle the infection on its own, it does intuitively seem that Seana, whose case is exactly analogous except that her immune system *isn’t* quite strong enough to do the job unaided, is in a position similar to Sean’s. And indeed, the difference between Sean and Seana involves only changes to facts which were not involved in causing their reason-relevant mental states (cf. SIM-2). (Neither has any evidence that their vitamin supplement will counteract their antibiotic, nor that their immune system might or might not handle the situation on its own.) Thus, KP_W and SIM (or rather their generalizations to physical action) offer to explain our hesitation here: We at least implicitly recognize that there are cases along the lines of Seana’s which are intuitively similar to Sean’s and in which we *wouldn’t* advise the agent to continue taking their supplement along with their prescribed antibiotic.

Finally, to further illustrate conditions (3) and (4) of SIM, consider the following case:

> Denny’s Toshiba laptop has crashed, and he brings it to a nearby computer repair shop. The manager tells Denny to return in a few hours, by which time his laptop should be fixed. After Denny leaves, his laptop is assigned to Ada, the technician who always repairs Toshibas. The shop’s other technician, Cody, specializes in Dell computers. (Denny and Ada were randomly assigned their specialties when they were hired three years ago.) Unbeknownst to anyone else, Cody is an identity thief: He applied for the repair job in order to steal personal information from the computers that he repairs, and then sell the information online to fraudsters. In fact, Cody has stolen information from all of the Dell computers that have been brought into the shop for repairs. But since Ada takes care of Denny’s laptop, there’s no chance of Cody stealing any of Denny’s information.

Despite the happy ending, I'd still feel *some* discomfort about saying, without qualification, "I would advise anyone in a similar position to do just as Denny did in turning his laptop over to the shop's employees". I suspect my reader will feel likewise—but perhaps not *as much* hesitation as in Travis's case. If so, then this is good evidence for SIM-3:

> (3) There are few and/or minor (as opposed to many and/or major) factual differences between K's total situation and L's total situation.

After all, an alternative case in which (say) Cody repairs Toshibas and Ada repairs Dells seems to involve either a minor change in the past which aggregated into a relatively major change (that is, Cody being assigned to Toshibas and Ada to Dells three years ago) or a major-ish recent change (for instance, we could suppose that Cody and Ada *sometimes* repairs laptops other than the ones they specialize in). (I say "major-ish" because there seems to be a threshold effect at work here: Going from "*only* Dells" to "*almost* always Dells, but sometimes Toshibas" might seem like a rather big change for Cody.) Thus, KP_W and SIM lead us to expect *some* hesitation (at least among some speakers) to say "I would advise anyone in a similar position to do just as Denny did", but probably not as much as in Travis's case or Sean's case.

Enough now with motivating examples: Let's get epistemic.

4.2 Gettier Explained

In this section, I'll begin using KP_W and SIM to show how SAGE can explain the verdicts of philosophers in various "Gettier-type" cases from the literature—that is, cases of justified true belief where a significant number of philosophers deny knowledge, and where their verdict clearly hinges on facts of which the believer is unaware. First, I'll focus on cases where there appears to be unanimity or near-unanimity among philosophers. (In §§ 4.3 and 4.4, I'll look at cases that have produced split verdicts among philosophers: a significant proportion ascribing knowledge and a significant proportion denying knowledge.) Before we begin, I'll repeat KP_W and SIM for reference:

> (KP_W) Given a believer K and her belief that *p* at time t, and given a speaker S aware of more facts about K's situation than K is at t, there is a natural interpretation for S of K's "position" which includes all reason-relevant facts of which S is aware—past, present, and future.
>
> (SIM) Given an interpretation of type KP_W of K's "position" and some other agent L, it is relatively more likely, for any given speaker, that there's a natural interpretation of "a position similar to K's" on which L's position is similar to K's if:
>
> > (1) L's reason-relevant mental states (that is, the mental states relevant to L doing the same sort of thing K did in believing that *p*) differ only in details (if at all) from K's. For each of K's relevant mental states, L should at least have an analogous one.
> > (2) The differences between K's position and L's position involve only facts which are not involved in *causing* any of K's or L's relevant mental states—or, if there are differences between the facts which caused K's mental states and L's mental states, the differences are relatively "far back" in the causal chains.
> > (3) There are few and/or minor (as opposed to many and/or major) factual differences between K's total situation and L's total situation.
>
> Let's first take care of the "stopped clock" case:
>
> Russell descends the stairs of his house at 7:34am—not having checked the time when he woke up—and sees the face of his old reliable grandfather clock reading 7:34. Russell comes to believe, correctly, that the time is 7:34am. However, he doesn't realize that the grandfather clock stopped working last night. In fact, the clock stopped just at 7:34pm.

As I've already noted, it appears that philosophers would unanimously deny that Russell knows the time, despite his belief being true and justified. SAGE can explain the verdict, because the following set of facts about "Alt-Russell" arguably constitutes a position similar to Russell's: (1′) descending the stairs at 7:37am; (2′) seeing the face of a clock that's always been exactly accurate in Alt-Russell's memory reading 7:34; (3′) representing his visual appearance of the clock as a decisive reason to believe that it's 7:34am. Alt-Russell's position involves only a minor change in one of the facts—namely, the time at which Russell descends the stairs—which caused Russell to see the clock at the time he did. (An alternative case in which the clock stopped at (say) 7:37pm, and Alt-Russell descends the stairs at 7:34am just like Russell, is also plausibly similar.) Thus, KP_W and SIM lead us to expect that most anyone who at least implicitly considers conveying the message "I would advise anyone in a similar position to do just as Russell did" will at least implicitly recognize that there are cases (along the lines of Alt-Russell's) in which the agent's position is similar, but where we would *not* advise the agent to do as Russell did ("Believe that the time is exactly what the clock says it is"). (As I suggested in § 4.1, this implicit recognition might occur through a heuristic simulation of situations which deviate in certain details from our conception of Russell's situation—a mechanism which likely evolved for the sake of planning for future scenarios similar to ones we've already encountered.) According to SAGE, then, philosophers deny that Russell knows the time because they implicitly register an aversion to conveying the message "I would advise anyone in a similar position to do as Russell did in believing that it's 7:34am". As philosophers implicitly recognize that they don't want to convey this message—since doing so would seem to suggest that they encourage the formation of false beliefs—they are inclined to say "Russell doesn't *know* that the time is 7:34am". More specifically, they're inclined to deny knowledge because they implicitly recognize a strong aversion ("That would be wrong!") to saying "Russell knows the time", and because they at least implicitly assume that there's a fact of the matter about whether or not Russell knows, at least once they believe they've been apprised of all the relevant facts.[75]

Let's now re-approach Gettier's (1963) two cases, which set off the whole firestorm in epistemology. (Bertrand Russell's 1948 case wasn't widely appreciated as a counterexample to the "justified true belief" account of knowledge until after Gettier's paper.) Here's the first case:

> Smith and Jones have each applied for a certain job. Smith has strong evidence that Jones will get the job and that Jones has ten coins in his pocket. (Suppose that the president of the company assured Smith that Jones would be selected, and that Smith recently counted the coins in Jones's pocket.) Smith infers the following proposition from the latter evidence: "The person who will get the job has ten coins in his pocket." But unknown to Smith, he and not Jones will get the job. Also, unknown to Smith, he himself has ten coins in his pocket. (ibid., p. 122)

Philosophers overwhelmingly agree that Smith does not know that the man who will get the job has ten coins in his pocket. Let's see if SAGE offers an explanation for this. According to KP_{W}, the following is a natural interpretation of Smith's position: (1) a memory of the company president saying that Jones will get the job; (2) a memory of counting ten coins in Jones's pocket; (3) the fact that he and not Jones will ultimately get the job; (4) Smith having ten coins in his own pocket.[76] (Note that facts (1) and (2) are reason-relevant mental states, while facts

[75] I don't mean to imply here that there actually *isn't* a fact of the matter about whether or not Russell knows. My point is simply that if a speaker does implicitly assume that there's a fact of the matter about whether or not *q*, and she feels strongly averse to saying "*q*" but not to saying "~*q*", then she's very likely to say "~*q*"—at least if she's explicitly asked about the matter. (And it is plausible that philosophers don't register an aversion to saying "Russell doesn't know": Even though this theoretically could be interpreted to imply "I advise Russell not to belief that it's 7:34am (given all my evidence)", the stipulation that Russell's belief is true arguably removes this interpretation from the conversational context. Moreover, the stipulation—implicit or explicit—that Russell's belief is justified cancels the interpretation "I wouldn't advise Russell to believe that it's 7:34am if my only evidence were his evidence" (which philosophers presumably also don't want to convey).

[76] Gettier leaves it unclear how exactly the counting in (2) transpired, but let's just go along with the example. For instance, suppose that Jones and Smith got bored and decided to empty their pockets and examine each other's pocket change (although Smith paid no attention to the contents of his own pocket, nor did Jones convey any information about them).

(3) and (4) are reason-relevant but not mental states.) As SIM suggests, the following set of facts about Alt-Smith constitutes a position similar to Smith's, since it involves only a minor change (one less coin in Smith's pocket) to a fact which didn't play a causal role in forming Smith's relevant mental states: (1′) a memory of the company president saying that Jones will get the job; (2′) a memory of counting ten coins in Jones's pocket; (3′) the fact that Alt-Smith and not Jones will ultimately get the job; (4′) Alt-Smith having nine coins in his pocket. Thus, SAGE explains the philosophical verdict that Smith's belief does not constitute knowledge: Philosophers implicitly recognize that there are cases, along the lines of Alt-Smith's, which are intuitively similar to Smith's but in which they *wouldn't* advise the agent to "do as Smith did"—that is, "Believe that the man who will get the job has however many coins you counted in the other candidate's pocket". Thus, philosophers are averse to implying "I would advise anyone in a similar position to do just as Smith did" and so deny that Smith has knowledge.

Let's move on to Gettier's second case:

> Smith has strong evidence that Jones owns a Ford: Jones has at all times in the past within Smith's memory owned a car, and always a Ford. Smith has another friend, Brown, of whose whereabouts he is totally ignorant. Smith thinks of the city Barcelona quite at random, and constructs the following proposition: "Either Jones owns a Ford, or Brown is in Barcelona." Smith realizes the entailment of this proposition from his belief that Jones owns a Ford, and he proceeds to accept it on the basis of this entailment. But in fact, Jones does *not* own a Ford: He is merely renting one right now. And as it happens, Brown is in fact in Barcelona. (ibid., pp. 122-23)

Philosophers again overwhelmingly deny that Smith knows that either Jones owns a Ford or Brown is in Barcelona. Now according to KP_W, the following is natural interpretation of Smith's position: (1) memories which he represents as providing decisive reason to believe that Jones owns a Ford; (2) no mental states which he represents as reasons to believe that Brown is in Barcelona; (3) the fact that Jones does not currently own a Ford; (4) the fact that Brown is currently in Barcelona. Since the fact that Brown is in Barcelona did not have any causal influence (direct or otherwise) in producing Smith's reason-relevant mental states, SIM predicts that we'll likely perceive the following set of facts about Alt-Smith as constituting a position similar to Smith's: (1′) [as above]; (2′) [as above]; (3′) [as above]; (4′) Brown is currently in Mataro (just up the coast from Barcelona). And of course we wouldn't advise Alt-Smith to "do just as Smith did"—"Believe that either Jones owns a Ford or Brown is in Barcelona". Moreover, even if we could somehow fill in the details of Gettier's original case such that the change from (4) to (4′) seems like a *major* change vis-à-vis Smith's position, plausibly we still wouldn't want to advise *Smith himself* to "do just as you did again"—since he might naturally interpret this to mean, "Choose a location L at random, and believe that either Jones owns a Ford or Brown is in L". Thus, SAGE explains the philosophical verdict on Smith: Philosophers register an aversion to convey the message "I would advise anyone in a similar position to do just as Smith did"—whether because they recognize that there are agents in different but plausibly similar positions to whom they wouldn't give the advice, or because they recognize that they wouldn't give the advice to Smith himself (or both).

After Gettier's two cases caused a stir in epistemology, subsequent authors have offered a wide array of cases which I've been calling "Gettier-type": cases in which K's belief is true and justified, yet in which a significant number of philosopher (including at least the author of the case him- or herself) deny that K knows. Since I certainly don't have space to consider each of these cases individually, I'll devote the rest of this section to a small but representative sample of Gettier-type cases where philosophers (by all appearances) unanimously deny that K knows, and show that SAGE offers an explanation for the verdict in each case. On the basis of the sample, I'll propose that we have strong inductive grounds to believe that SAGE can explain the philosophical community's verdict on any other Gettier-type case (existing or future) where the verdict is unanimous in denying knowledge. (I'll consider split-verdict cases in the next section.)

To begin, here's an early Gettier-type case from Chisholm (1966):

> While gazing over a field, Rod sees what looks to be a sheep and comes to believe there is a sheep in the field. And he's right: just beyond a hill in the middle of the field, there is a sheep. It's out of view, though, and Rod has no idea it is there. What he sees is a dog, convincingly dressed up as a sheep. (cf. Chisholm 1966, p. 23; reprinted in Ballantyne 2011)

Philosophers deny that Rod knows that there's a sheep in the field. By KP_W and SIM, the following set of facts about Alt-Rod plausibly constitutes a position similar to Rod's: (1′) a visual appearance as of a sheep in a field; (2′) the fact that the object causing this appearance is a dog dressed up as a sheep; (3′) the fact that there are no sheep in the field. (3′) marks a difference from Rod's position, but the change does not implicate any facts which caused Rod's reason-relevant mental state. (At least, no fact is implicated which *directly* caused Rod's visual perception. Perhaps there's a dog dressed up as a sheep in the field *because* there are sheep in the field—say, to help round them up. But then the fact of there being sheep in the field still isn't a *direct* cause of Rod's visual perception.) Thus, SAGE suggests that philosophers deny that Rod knows because they at least implicitly recognize that there are cases (along the lines of Alt-Rod's) which are intuitively similar to Rod's but in which they would not advise the agent to "do just as Rod did".

There are also cases in the literature which illustrate the importance of SIM-1—in particular, the idea that someone whose reason-relevant mental states are *similar but not identical* to K's might still seem to be in a "similar position". For instance, consider the following case from Collier (1973), which we examined in § 1.3 as a counterexample to Goldman's (1967) causal theory of knowledge:

> Suppose that unbeknown to Smith I administer an hallucinogenic drug to him. Since he doesn't realize that he has been drugged, he believes that his hallucinations are real. But one of the hallucinations is that I gave him the drug that I, in fact, gave him, and in particular, he believes that his hallucination is real. (p. 350)

Let's call the protagonist of this story "Kenneth". Collier, along with probably every other philosopher, judges that Smith doesn't know that Kenneth gave him a hallucinogenic drug, even though his belief is true and justified. But note that the following set of facts about Alt-Smith plausibly constitutes a position similar to Smith's: (1′) a visual perception as of being given a hallucinogenic drug by Gwyneth (one of Alt-Smith's friends); (2′) no memories or perceptions indicating that the visual perception in (1′) is a hallucination; (3′) the fact that the visual perception in (1′) is a hallucination, caused by a drug given to Alt-Smith by Kenneth (another of Alt-Smith's friends); (4′) the fact that Gwyneth did not give Alt-Smith a hallucinogenic drug. Although Alt-Smith's relevant visual perception differs slightly from Smith's, it almost surely lies within the bounds of "similar" on a natural interpretation. Moreover, it presumably wouldn't require much of a change in the properties of the drug Kenneth gave Smith, and/or in Smith's neurophysiology, to have him hallucinate that one of his friends other than Kenneth gave him the drug. Finally, we would not advise Alt-Smith to "do just as Smith did"—that is, to trust the contents of his visual perceptions relating to being given a particular substance by a particular person. Thus, SAGE explains the philosophical verdict on the hallucinating Smith: Philosophers don't want to seem to be conveying that they would advise hallucinating agents like Alt-Smith to "do just as Smith did".

Moving on now, there are Gettier-type cases in the literature where the crucial element appears to be the *timing* of a particular event beyond the believer's awareness. The event is question makes the believer's belief go from "true" to "false". Philosophers typically judge that so long as the adverse event is quite a ways in the future, the believer still has knowledge. But as the event impends, the verdict switches to being that the believer does *not* know. Here's an example from Feit and Cullison (2011):

> In the office, Smith wears a smock to protect his clothes. A few moments before 5pm, Smith leaves the office after draping his smock over the back of his chair, as he has done for years. After locking his door and walking out of the office, he believes at 5pm that his smock is then draped over his chair. The smock is just where Smith believes it to be at this time. However, a few seconds before 6pm, thieves break into the office and steal Smith's smock. (p. 294)

Feit and Cullison judge that Smith knows at 5pm that his smock is draped over his chair, but *not* at 5:59pm, even though his smock is still there at the later time. (The authors' judgments here appear to be representative of all philosophers.) Notice now that intuitively, 5pm-Smith and Alt-5pm-Smith (where the latter leaves just before 5pm and whose smock is stolen seconds after he leaves) are in very different positions. This is as SIM predicts: Moving up the time of the burglary by an hour seems to be a rather major change, even though the facts involved don't causally influence Smith's mental states. But now consider 5:59pm-Smith. It seems that there's not much difference between him and Alt-5:59pm-Smith, whose smock is stolen at 5:58 rather than 6. Thus, SIM and SAGE offer an explanation for the verdicts here: Philosophers don't register any cases similar to 5pm-Smith's where they wouldn't advise doing as Smith did, but they *do* register cases similar to 5:59pm-Smith's where they wouldn't advise doing as Smith did (that is, continuing to believe that one's smock is draped over one's office chair).

I suggested in § 4.1 that our intuitive notions of similarity can be subject to framing effects, and I think Feit and Cullison's case offers a good way to test that suggestion. Consider the following variation (my variation, not theirs):

> In her office, Sabina wears a smock to protect her clothes. On Friday June 24, Sabina leaves the office for her month-long summer vacation after draping her smock over the back of her chair, as she has done for years. Sabina is the only one with key access to her office, and she thereby believes throughout her vacation that her smock is where she left it. But on Friday July 15, thieves break into Sabina's office and steal her smock.

Now consider Sabina's belief on Thursday July 14: Does she know that her smock is still on her chair? I feel quite inclined to say no—and I suspect many philosophers will agree with me. But if we're not willing to grant that Sabina knows her smock is on her chair a full day ahead of its removal, why do we grant that Smith knows his smock is on his chair merely *an hour* before its removal? SIM offers an explanation. In Smith's case, the wording leads us to adopt a relatively small "unit of significant time"—on the order of a minute or two—since all the action takes place within an hour. In Sabina's case, by contrast, the wording suggests a relatively large unit of significant time—on the order of a day or two—since the events take place over the period of a month. Thus, if I'm right about the effects of framing on our intuitive similarity judgments, moving up the burglary by an hour for Smith involves a lot of significant time units, whereas moving up the burglary by a day for Sabina does *not*. Once we recognize this, SAGE clearly offers an explanation for our different verdicts about Smith and Sabina.[77]

Moving on, there are several cases in the literature which have come to be known as "guardian angel" cases: Some believer K forms her beliefs in a way which would normally yield

[77] Researchers have repeatedly found that the way situations or decisions are described can have big impacts on people's reactions. For an overview, see Levin et al. (1998).

false beliefs, but there is some hidden agent or device (unbeknownst to K) ensuring that K's beliefs turn out true after all. We saw one such case in Chapter 1—from Pritchard (2010)—so let's return to it as a representative of the guardian-angel category:

> Imagine that our agent—let's call him 'Temp'—forms his beliefs about the temperature in his room by consulting a thermometer on the wall. Unbeknownst to Temp, however, the thermometer is broken and is fluctuating randomly within a given range. Nonetheless, Temp never forms a false belief about the temperature by consulting this thermometer since there is a person hidden in the room, next to the thermometer, whose job it is to ensure that whenever Temp consults the thermometer the temperature in the room corresponds to the reading on the thermometer. (Pritchard 2010, p. 49)

Pritchard judges, as do apparently most other philosophers, that Temp's beliefs about the temperature fail to constitute knowledge, despite being justified and true. Notice now that Temp is being *deceived* in this case: The person hidden in the room is allowing Temp to assume that the thermometer is functioning normally, even though it's not. It plausibly wouldn't be so different, then, if the hidden person were further deceiving Temp into having false beliefs about the *temperature*. There's arguably a threshold effect here: Given that Temp is already being deceived in one way, it doesn't seem to make much of a difference if we alter the case so that he's being deceived in another, related way as well. In particular, consider the following position of Alt-Temp: (1′) a visual perception of a thermometer steadily reading 69°F; (2′) no memories indicating that this thermometer is inaccurate; (3′) the fact that the thermometer is broken; (4′) the fact that a hidden person in the room is secretly adjusting the thermometer, whenever Temp looks at it, to give a slightly inaccurate reading; (5′) the fact that the room is currently 70°F. Moving from Temp's to Alt-Temp's position requires only a small change in the direct cause of Temp's relevant mental states (in particular, the thermometer reading 69°.) If this analysis is correct, then SAGE explains the philosophical verdict: Philosophers register an aversion to conveying the message, “I would advise anyone in a similar position to do just as Temp is doing”, since there are plausibly similar cases where this advice would be encouragement to form false beliefs.

Finally, I’d like to illustrate the importance of the first condition of SIM, which I repeat here:

> (1) L’s reason-relevant mental states (that is, the mental states relevant to L doing the same sort of thing K did in believing that *p*) differ only in details (if at all) from K’s. For each of K’s relevant mental states, L should at least have an analogous one.

The following case from Goldman (1976) might at first suggest that SAGE and SIM make false predictions:

> Oscar is standing in an open field containing Dack the dachshund. Oscar sees Dack and (noninferentially) forms a belief that “The object over there is a dog”. Now suppose that “The object over there is a wolf” is a relevant alternative (because wolves are frequenters of this field). Further suppose that Oscar has a tendency to mistake wolves for dogs (he confuses them with malamutes, or German shepherds). Then if the object Oscar saw were Wiley the wolf, rather than Dack the dachshund, Oscar would (still) believe “The object over there is a dog”. (p. 779)

Goldman takes it as obvious, as does apparently every other philosopher, that Oscar knows that the object he sees is a dog. By KP$_W$, something like the following is a natural interpretation of Oscar’s position: (1) visual perceptions as of a moving thing with four legs, a short tail, droopy ears, and a hotdog-shaped body; (2) representations of these visual perceptions as reasons to believe that he’s looking at a dachshund, and therefore a dog; (3) the fact that the object causing these visual perceptions is a dog. SAGE might seem to predict that philosophers would *deny* knowledge in this case, since Alt-Oscar might seem to be in a similar position: (1′) visual perceptions as of a moving thing with four legs, a long tail, a shaggy coat, and a large body; (2′) representations of these visual perceptions as reasons to believe that he’s looking at a malamute, and therefore a dog; (3′) the fact that the object causing these visual perceptions is a wolf. (Since there are already wolves in Oscar's area, (3′) plausibly isn't a major departure from (3).) Yet notice that Oscar's reason-representations of his visual perceptions differ from Alt-Oscar's reason-representations of *his* visual perceptions: Oscar represents his perceptions as reason to believe that he's looking at a *dachshund*, while Alt-Oscar represents his as reason to believe he's looking at a *malamute*. Moreover, the visual perceptions themselves are clearly very different. (At least given the frame of reference suggested by the case: Oscar evidently would notice significant differences between his visual perceptions caused by a dachshund and those caused by a wolf, even if he would identify both animals as dogs). Thus, in light of SIM-1, SAGE arguably does *not* falsely predict that philosophers will deny that Oscar knows that he's looking at a dog.[78]

4.3 Cause for Concern?

I've now reviewed what I take to be a representative sample of Gettier-type cases in the literature where philosophers unanimously deny that the believer has knowledge. Since SAGE offers an explanation for each verdict in the sample, I think we have good inductive grounds for supposing that SAGE can explain the verdict in any other Gettier-type case that draws a unanimous denial of knowledge from philosophers. And in light of my analysis of Goldman's Oscar case, we also have good reason to expect that SAGE doesn't make any false predictions,

[78] But note that SAGE can still explain what would almost surely be the philosophical verdict that Oscar would *not* know that he's looking at a dog if his visual perceptions were caused by a malamute. For then there's plausibly a similarly-positioned Alt-Oscar whose similar visual perceptions are caused by a wolf, and who moreover represents these visual perceptions *exactly* as Oscar represents his visual perceptions of the malamute: as reasons to believe that the animal is a malamute, and therefore a dog.

since philosophers' ascription of knowledge to Oscar at first glance *seems* to disconfirm SAGE. Oscar's case seems to be the best candidate for SAGE making a false prediction—but as we saw, SAGE in fact doesn't go wrong there.

I'll now consider a few cases in the literature which have drawn mixed reactions from philosophers: A significant faction asserts that K knows, while another significant faction asserts that K clearly does not know. Before I begin, here for reference are restatements of KP$_W$ and SIM:

> (KP$_W$) Given a believer K and her belief that *p* at time t, and given a speaker S aware of more facts about K's situation than K is at t, there is a natural interpretation for S of K's "position" which includes all reason-relevant facts of which S is aware—past, present, and future.
>
> (SIM) Given an interpretation of type KP$_W$ of K's "position" and some other agent L, it is relatively more likely, for any given speaker, that there's a natural interpretation of "a position similar to K's" on which L's position is similar to K's if:
>
> > (1) L's reason-relevant mental states (that is, the mental states relevant to L doing the same sort of thing K did in believing that *p*) differ only in details (if at all) from K's. For each of K's relevant mental states, L should at least have an analogous one.
> > (2) The differences between K's position and L's position involve only facts which are not involved in *causing* any of K's or L's relevant mental states—or, if there are differences between the facts which caused K's mental states and L's mental states, the differences are relatively "far back" in the causal chains.
> > (3) There are few and/or minor (as opposed to many and/or major) factual differences between K's total situation and L's total situation.

Perhaps the most infamous split-verdict case is the "Fake Barn" case from Goldman (1976):

> Henry is driving in the countryside with his son. … [U]nknown to Henry, the district he has just entered is full of papier-mâché facsimiles of barns. These facsimiles look from the road exactly like barns, but are really just façades, without back walls or interiors, quite incapable of being used as barns. They are so cleverly constructed that travelers invariably mistake them for barns. Having just entered the district, Henry has not encountered any facsimiles; the object he sees [right now] is a genuine barn. (pp. 772-73)

Goldman supposes that most anyone will be inclined to deny that Henry knows that the object he sees is a real barn (p. 773). In fact, many philosophers since Goldman have disputed this intuition. (See, for instance, Feit and Cullison 2011, Lycan 2006, and Gendler and Hawthorne 2005). I personally share Goldman's intuition; but in my experience I've found that only about half of my colleagues do so. The other half insist that Henry's belief *does* constitute knowledge, or at least have mixed feelings about the matter. SAGE and SIM offer an explanation for this state of affairs. By $\mathrm{KP_W}$, the following is a natural interpretation of Henry's position: (1) a visual perception as of a barn; (2) the fact that there are several facsimile barns in the area which look from the front just like real barns; (3) the fact that Henry's visual perception is caused by a real barn. If there is to be any position "similar" to Henry's in which we would *not* advise the agent to do as Henry did, it would have to be along the following lines: (1′) a visual perception as of a barn; (2′) the fact that there are several facsimile barns in the area which look from the front just like real barns; (3′) the fact that Alt-Henry's visual perception is caused by one of the facsimiles.

Does (3′) depart too much from (3) to make Alt-Henry's position qualify as similar to Henry's? SIM suggests that there's room for debate: On the one hand, there *are* fake barns in Henry's area, so it might not seem to make much of a difference if we replaced the real barn he's looking at with a fake one. (Recall my discussion of threshold effects in § 4.1.) On the other hand, this replacement would entail a change to the *direct* cause of Henry's relevant mental state: His visual perception as of a barn would be caused by a fake barn, rather than a real one—and these are very distinct objects. Thus, SAGE can explain the philosophical debate over Henry as follows: Some philosophers have an implicit notion of similarity which *does* allow for

Alt-Henry's position to count as similar to Henry's, while others have a notion of similarity which does *not* assimilate their positions. Thus, when considering whether or not to ascribe knowledge to Henry, some philosophers register an aversion to conveying "I would advise anyone in a similar position to do as Henry did" (and thereby deny that Henry knows) while others do not (and thereby assert that Henry knows).

Moreover, framing effects might be important here: Goldman doesn't specify in his case *how close* the nearest fake barns are to Henry, nor *how soon* Henry will be passing them. If we suppose that the nearest fake barns are miles away, and/or that Henry is traveling very slowly (say, he's stopped for a picnic on the side of the road, right across from the real barn he's looking at), then I find myself more inclined to say that Henry knows. But if we suppose that the nearest fake barns are very close (say, within a few hundred yards), and that Henry will see them very soon (if he hasn't already), I once again feel confident saying that he doesn't know the real barn is a real barn. [79] Notice now that these two alternative setups suggest different units of significant distance: In the first, moving one of the fake barns to the position of the real one seems to involve a big change, whereas in the second it might not. If I'm right about framing effects, then our intuitive notion of similarity might indeed register a "fake barn" alternative as constituting a similar position in the "fast and close" alternative, but not so in the "slow and far" alternative. SAGE then offers an explanation for my (and I suspect at least some others') variable intuitions about Henry, depending on how we fill in the details.

Another controversial case comes from Gendler and Hawthorne (2005):

[79] Gendler and Hawthorne (2005) imply that they also find their intuitions about Goldman's changing based on how quickly Henry is moving through the landscape.

> Robert enters a room and asks someone the time. She replies truthfully and correctly, and she is extremely reliable. But Robert's informant happens to be surrounded by a roomful of compulsive liars [which Robert doesn't realize]. (p. 346)

Gendler and Hawthorne report from their informal survey of selected philosophers that most ascribe knowledge to Robert in this case. However, a few philosophers in their survey denied that Robert comes to know the time. SAGE offers an explanation for the split verdict, relying on the same idea as in the Fake Barn case: An alternative in which Alt-Robert asks one of the liars in the room what the time is, and is told the wrong time, involves a change compared to Robert's case in a fairly direct *cause* of the relevant mental states. In particular: Robert hears that the time is (say) 3:13pm because someone looked at her reliable watch and reported what she saw with the intention of conveying the truth, whereas Alt-Robert hears that the time is (say) 3:14pm because someone looked at her watch and reported a different time than what she saw with the intention of misleading Alt-Robert. Thus, by SIM-2, it's relatively less likely for any given speaker to perceive a case like Alt-Robert's as similar to Robert's. (That is, less likely than if there were no significant differences in the causal formation of their relevant mental states.)[80] On the other hand, there *are* liars in Robert's room. Thus, despite the change in causal formation of mental states, to some speakers Alt-Robert's position might not seem *too* different from Robert's. (By contrast, if there were no liars in the room or anywhere nearby, Alt-Robert's position almost certainly wouldn't seem similar to Robert's for *any* speaker.) SAGE thereby offers an explanation for the split philosophical verdict, by analogous reasoning as in the Fake Barn case.

[80] By comparison, consider Russell and Alt-Russell in § 4.2. Compare also Temp and Alt-Temp (§ 4.2), where the change to the causal formation of the mental state (the hidden person being deceptive only about the thermometer, as opposed to being deceptive about both the thermometer *and* the temperature) doesn't seem as large as the change between a thoroughly honest informant and a thoroughly deceptive one.

There are several other Gettier-type cases in the literature which have provoked mixed reactions from philosophers (particularly those featured in Gendler and Hawthorne 2005, who themselves don't take an explicit stand on the cases they propose). All such cases that I'm aware of involve, just like Henry's and Robert's cases, a believer whose reason-relevant mental states are caused by the true proposition which they believe, but whose environment offers opportunities to form similar or identical mental states which would *not* be caused by the proposition which the believer would thereby believe. Thus, my analysis of Henry's and Robert's cases generalizes to the other controversial cases: Some philosophers' intuitive similarity notions are relatively permissive about changes to the direct causes of a believer's relevant mental-states, while the similarity notions of others are relatively less so.

4.4 Evidential Horizons

There's one class of Gettier-type cases I haven't yet discussed: cases where K's belief is caused by the true proposition '*p*' which she believes, but where there is nearby "misleading" evidence which K has not seen, and which (if she did see it) would remove her justification for believing that *p*. These cases were first reported by Harman (1973), one of whose examples is the following:

> Donald has gone off to Italy. He told you ahead of time that he was going; and you saw him off to the airport. He said he was to stay for the entire summer. That was in June. It is now July. … [F]or reasons of his own Donald wants you to believe that he is not in Italy but in California. He writes several letters saying that he has gone to San Francisco and has decided to stay there for the summer. He wants you to think that these letters were written by him in San Francisco, so he sends them to someone he knows there and has that person mail them to you with a San Francisco postmark, one at a time. You have been out of town for a couple of days and have not read any of the letters. You are now standing before the pile of mail that arrived while you were away. Two of the phony letters are in the pile. (p. 143)

Let's name the "you" of this scenario "Gil". Harman insists that Gil, standing before his pile of mail, does not know that Donald is still in Italy, despite his belief being true and justified. Many philosophers have agreed with Harman's intuition, although a sizeable contingent disagrees. (See, for instance, Lycan 1977 and Leplin 2009.)

Can SAGE explain this split verdict? If KP_M and KP_W are the full story about the natural interpretations of K's "position" in any given case, then it seems like SAGE is at a loss. For the only natural interpretations of Gil's position of type KP_M fall along the following lines: memories of Donald saying he would be in Italy all summer, memories of seeing Donald off to Italy, no memories indicating that Donald is not still in Italy. Thus, SAGE does not as yet predict any aversion to conveying the message "I would advise anyone in a similar position to do just as Gil did in continuing to believe that Donald was in Italy". Further, the following is the interpretation of Gil's position of type KP_W: (1) memories of Donald saying he would be in Italy all summer; (2) memories of seeing him off to Italy; (3) no memories indicating that Donald is not in fact in Italy; (4) the fact that there are letters written in Donald's handwriting indicating that he's now in San Francisco; (5) the fact that Donald is in fact still in Italy. Plausibly, the position of Alt-Gil, whose friend Alt-Donald is in fact no longer in Italy, is in a very different position from Gil: Donald apparently never intended to leave Italy during the summer. Thus, as far as KP_W is concerned, speakers should have no hesitation to convey the message "I would advise anyone in a similar position to do just as Gil did".

But I believe there is in fact a natural interpretation of K's position which falls *between* her mental-state perspective (on the one hand) and *all* reason-relevant facts about her situation (on the other). I'll call this the "accessible evidence" interpretation of K's position at t with regard to proposition '*p*': It constitutes all of K's reason-relevant mental states up through t, along with any reason-relevant facts which could become the content of one of K's sensory perceptions (vision, hearing, touch, smell, taste) without much if any physical effort on K's part beginning at t. I'll say that the latter set of reason-relevant facts, along with K's relevant mental states at t, constitutes K's "accessible evidence":

> (KP_A) Given a believer K and her belief that *p* at time t, and given a speaker S aware of more facts about K's situation than K is at t, there is a natural interpretation for S of K's "position" which includes all of K's reason-relevant mental states up through t, along with all reason-relevant facts of which S is aware and which are accessible via sensory perception for K at t.[81]

Why have a natural interpretation of K's position which is restricted to her accessible evidence? Here's one idea: It might have been advantageous for our ancestors to be disposed to form representations of another agent which included only those relevant facts which it was at least possible for that agent to become aware of in the near-future. Our ancestors could thereby be "prepared" for what that agent might be apprised of in the near-future, and so be prepared for how she might act, or what she might say.

Here's an example, from the domain of action, to support my hypothesis KP_A:

> Desi and her friends are attending a party tonight at Nate's house, and Desi has agreed to be the designated driver (and so to not consume any alcohol). At the drinks table, Desi is surprised to find that there are apparently no alcoholic beverages: The four pitchers are labeled "Apple Juice", "Orange Juice", "Grape Juice", and "Seltzer". And the contents of these pitchers indeed correspond to their labels. However, Nate has decided to try a sociological experiment: He's placed a sign at one end of the table saying, "All of the juice pitchers are spiked with vodka". He wants to see whether his friends will act drunk because they *expect* to get drunk (after reading the sign), despite not actually imbibing any alcohol. However, when Desi approaches the drinks table, someone else is standing in front of the sign, blocking it from Desi's view. She pours herself a glass of orange juice and finishes it off in a few gulps.

[81] I've used the subscript "A" to indicate that the natural interpretations predicted by this hypothesis include all of K's "A"ccessible evidence of which the speaker is aware.

At least speaking for myself, I feel at least *some* hesitation to say, without any qualification, "I would advise anyone a similar position to do just as Desi did in pouring herself a drink from the table". But my hesitation would be needless if I could expect my listeners to understand Desi's "position" either exclusively in terms of her perspective, or exclusively in terms of *all* relevant facts. I think the best explanation is that there's an interpretation of Desi's "position" which is salient for at least some of us (including me) and which encompasses all and only her *accessible* evidence—in particular, the fact of Nate's sign indicating that the beverages are alcoholic, but *not* the fact that the drinks don't in fact contain alcohol.

If I'm correct that accessible-evidence interpretations of K's "position" are natural for at least some speakers, then SAGE can explain the philosophical verdict on Gil after all: At least some philosophers implicitly register the following as a natural interpretation of Gil's "position": (1) memories of Donald saying he would be in Italy all summer; (2) memories of seeing Donald off to Italy; (3) no memories indicating that Donald is not in fact in Italy; (4) the fact that there are letters written in Donald's handwriting indicating that he's now in San Francisco. Surely *none* of us would advise someone in this position to believe that Donald is still in Italy. SAGE posits that the philosophers for whom [(1)-(4)] is a salient interpretation of Gil's position register an aversion to saying "Gil knows that Donald is still in Italy", since they implicitly recognize knowledge ascriptions to convey the (unqualified) message, "I would advise anyone in a similar position to do just as K did". These philosophers consequently deny that Gil's belief constitutes knowledge. Now what about the sizable contingent of philosophers who insist that Gil *does* know? It may simply be that interpretations of type KP_A are not salient for *everyone*, or at least not equally so: Some of us (for reasons of personal experience, or of psychology) naturally gravitate only towards interpretations of type KP_M and KP_W, with no "middle ground" between them. And for what it's worth, I do agree that my inclination to deny that Gil knows is less strong than in (say) Gettier's original two cases.

Harman (1973) offers two more cases along the same lines: K's belief is caused by the fact that *p*, but there's nearby misleading evidence which would (if "accessed") negate K's justification for believing that *p*. SAGE's explanations for the split philosophical verdicts on these cases (and others like them that have been proposed since) are analogous to the one I've offered for the verdict on Gil and the Letters.

4.5 Post-Gettier Explained

We've now reviewed what I believe to be a fully representative sample of the Gettier-type cases in the literature: cases where K's belief is both true and justified, but where a significant proportion (if not all) philosophers deny that K knows. I've shown that in each representative case, SAGE offers an explanation for the philosophical verdict, in light of my additional hypotheses KP_W, SIM, and KP_A. The theme in each case (other than in the accessible-evidence cases) was this: There is a way to change the details of the case such that the position of the alternative agent Alt-K is intuitively similar to that of K, but in which advising Alt-K to "do just as K did" would be an encouragement for Alt-K to form a false belief (which Alt-K would not be aware of, but which we as observers would be). I posited that in such cases, philosophers considering whether to ascribe knowledge to K implicitly recognize that there are alternative cases in which an agent's position (understood as comprising all reason-relevant facts of the situation) is similar to K's position, but in which they wouldn't advise the agent to do just as K did. If this is so, SAGE explains why philosophers deny knowledge: We all implicitly understand "K knows that *p*" to convey the prospective advisory message "I would advise anyone in a similar position to do just as K did"—a message which philosophers wouldn't want to convey in the cases in question, for the reasons just reviewed.

Since my survey of the literature hasn't been exhaustive, there may well be some Gettier-type cases where KP_W, SIM, and KP_A still leave it mysterious why at least some philosophers deny knowledge. The following variation which I've devised of Harman's Letters case might be one such case:

> Gil is now standing before the pile of mail that arrived while he was away. Two of Donald's phony letters are in the pile. In addition, several minutes ago Donald sent an email to Gil's Hotmail account with an explanation about the letters: "If you've read my letters already: I was just kidding! I'm still here in Italy". Attached to the email is video footage of Donald at the Venice Biennale, with a banner behind him reading "La Biennale di Venezia, Luglio [July] 2016". However, Gil almost never checks his Hotmail account, and in fact he won't check it until September.

I'm inclined again to judge that Gil doesn't know that Donald is still in Italy. Yet by the lights of KP_W and KP_A, there should be no problem in conveying "I would advise anyone in a similar position to do just as Gil did in continuing to believe that Donald is in Italy". In particular, Gil's accessible evidence includes both the letters in which Donald claims to be in San Francisco *and* Donald's follow-up email. (It wouldn't take Gil much effort to check his Hotmail account, even though he won't.) Yet it's possible that there's a further natural interpretation of K's "position" which is restricted to the accessible evidence which K is actually *fairly likely* to access—not just the evidence which it would be *possible* for K to access in the near-future. (The corresponding representation of K perhaps would have been useful for our ancestors in making a less-cautious assessment of what K is likely to be apprised of in the near-future.)

However, for the sake of space, I won't propose this as an additional formal hypothesis about the candidate salient interpretations of K's "position". What I'd like to do instead is to offer SAGE as a *framework* for explaining philosophers' verdicts on cases in the literature. My hypotheses KP_M, KP_W, and KP_A may well not exhaust the natural interpretations of K's position. And SIM surely doesn't cover all of the factors that can affect our intuitive judgments of similarity between different agents' positions. (At the least, we could surely say something more specific about what counts in favor of a change in facts being "major" vs. "minor".) But I do think I've done enough here to make it clear how to use SAGE as an explanatory tool for any further case where K's belief is true yet a significant proportion of philosophers judge K not to know. We may well need more hypotheses than I've offered here, but I think we have good inductive grounds for expecting that any case where KP_M, KP_W, KP_A, and SIM don't suffice can be explained by some other plausible hypothesis about natural interpretations and/or our implicit similarity judgments.

I'd like to close the chapter by considering whether SAGE can explain the appeal of each of the conditions for knowledge which has been proposed in the wake of Gettier's seminal paper—conditions meant to help "fill the gap" between true belief and knowledge. The conditions on offer are many and diverse, and they've had varying levels of success in aligning with the actual verdicts of philosophers on cases in the literature. If SAGE can explain why each of these conditions appealed to at least some philosophers—and also systematically explain why each condition has failed to gain consensus as fully "solving" the problem of what knowledge is—then so much the better: If we do accept THYME and SAGE as accurate accounts of the origins and functioning of "know" (as I'll argue in Chapter 6 that we should), we won't need to look elsewhere for explanations for these puzzling developments in recent epistemology.

Since I can't give individual attention to even a small portion of the many conditions that have been proposed, I'll analyze just those four which I reviewed in Chapter 1: the causal condition, the safety condition, the virtue condition, and Lewis's "Rule of Resemblance" condition. By showing that SAGE can explain the appeal and degree of success of each of these conditions, I hope to make it plausible that SAGE can do likewise for all other conditions that epistemologists have proposed. Let's begin, then, with Goldman's (1967) casual condition:

> S knows that *p* if and only if the fact *p* is causally connected in an "appropriate" way with S's believing *p*. "Appropriate" knowledge-producing causal processes include the following: (1) perception [that *p*]; (2) memory [that *p*]; (3) a causal chain [which includes the fact that *p* and] which is correctly reconstructed by inferences, each of which is warranted …; (4) combinations of (1), (2) and (3). (pp. 369-70)[82]

The rough idea is that K's belief that *p* constitutes knowledge just in case the belief was *caused* by the fact that '*p*' (with an exception for beliefs about the future: the belief should be caused by the same facts which cause '*p*' to be true). SAGE explains the appeal of requiring '*p*' to be in the causal chain leading to K's belief: If *p* is not in the causal chain, then by SIM there are likely to be alternative agents whose positions are similar to K's, but in which it's *not* the case that *p*. (Altering the fact that *p* won't constitute a major change since '*p*' did not help cause—even indirectly—K's relevant mental states.) This phenomenon is illustrated by the case of Rod in § 4.2: His relevant mental state is caused by a dog dressed as a sheep, and *not* by the fact which he believes ("There's a sheep in the field").

[82] Goldman offers this as an account of *empirical* knowledge—that is, knowledge of contingent propositions such as "Edmund Gettier published a short but influential paper in 1963", as opposed to "necessary truths" such as Fermat's last theorem.

But SAGE also explains why the causal condition failed to fully satisfy philosophers: Even if ‘*p*’ does belong to the causal chain leading to K’s relevant mental states, it may be an *indirect* cause, such that changing it might still leave the alternative position intuitively “similar” to K’s. For instance, in Pritchard’s Temp case (§ 4.2), Temp’s belief that the temperature is (say) 69°F *is* caused by this fact—but not directly. Rather, the temperature being 69° causes the person hidden in the room to manipulate the thermometer to read 69° just when Temp glances at it. Since (as I suggested above) Temp is being deceived here, it’s plausible to think that Alt-Temp, who’s being deceived not only about whether the thermometer is working normally but *also* but the temperature, is in a “similar” position. Thus, here’s at least one explanation SAGE offers for why the causal condition isn’t sufficient to capture the philosophical verdict on all cases: If the fact that ‘*p*’ causes K’s relevant mental state, but isn’t a *direct* cause, there may well be intuitively similar cases where we wouldn’t advise the agent to “do just as K did”.

Let’s now move on to the “safety” condition, formulated in the following way by Pritchard (2005): For all agents and all propositions ϕ, an agent’s belief that ϕ is “safe” if and only if:

> … in nearly all (if not all) nearby possible worlds in which she forms the belief about ϕ in the same way as she forms her belief in the actual world, that agent only believes that ϕ when ϕ is true. (p. 163)

(Pritchard defines “possible worlds” as fully-specified sets of circumstances, where “’distant’ possible worlds are very unlike the actual world, whilst ‘nearby’ possible worlds are very alike the actual world” (p. 128).) Many authors, among them Williamson (2000) and Sosa (1999), have argued for a similar condition as at least *necessary* for knowledge (if not sufficient). And indeed, SAGE quite clearly explains the appeal of safety: If there’s a similar alternative set of circumstances in which K forms her belief "in the same way" (which I take to mean "with roughly the same relevant mental states") but in which her belief is false, then philosophers clearly won't want to convey the message "I would advise anyone in a similar position to do just as K did"—which is what, according to SAGE, philosophers implicitly understand knowledge ascriptions to convey.

But it seems most safety-defenders concede that safe belief isn't *sufficient* for knowledge—and SAGE gives a clear explanation for why the safe belief condition doesn't capture the philosophical verdict on *all* cases. For even if there's no alternative set of circumstances in which Alt-K forms a false belief and which is similar to K's actual "possible world" *in toto*, there may still be an alternative set of circumstances which is similar *when its details are weighted as per SIM*. That is, even if changing from K's *world* to Alt-K's *world* entails relatively major changes, if those changes aren't changes to the causal chains leads to K's relevant mental states (or at least not to the most *direct* causes), then Alt-K's *position* may still well be intuitively similar to K's *position*. Again, Pritchard's Temp case is a good illustration here (and Pritchard himself uses the case as a counterexample to the sufficiency of safety): Changing the facts so that the person hidden in the room is deceiving Temp about the temperature, and not just about the status of the thermometer, might seem like a relatively big change (perhaps requiring a large change in the goals and past experiences of the hidden person), such that these two "possible worlds" are quite "distant" from each other. In general, Pritchard suggests, when Temp comes to believe that the temperature is 69°, in all the *nearest* possible worlds where he forms the same belief in the same way (that is, by consulting the thermometer), the hidden person is in the room and ensuring that Temp's belief is accurate. But since the similarity relation for agents' *positions* is weighted towards the direct causes of their relevant mental states (as per SIM), changing the nature and motives of the hidden person in the room doesn't intuitively yield that big of a difference between the positions of Temp and Alt-Temp (the latter being deceived about the temperature, and not just about how the thermometer works).

In light of the apparent insufficiency of safety, Pritchard (2010) offers a further "virtue" condition, which is similar to conditions offered by self-labeled virtue epistemologists like Sosa (2007) and Greco (2007):

> Knowledge [requires that one's belief] arises out of the reliable cognitive traits that make up one's cognitive character, such that one's cognitive success is to a significant degree creditable to one's cognitive character. (p. 54)

SAGE sheds light on the appeal of the virtue condition: If the truth of K's belief is *not* to a significant degree creditable to her cognitive character, then it may well be due to the whims of something or someone else. But if so, there will likely be an intuitively similar position for Alt-K in which the whims of that something or someone else swayed in the direction of *falsity*, rather than truth. For instance, Pritchard gives the example of "an agent with poor mathematical skills who is trying to work out a series of mathematical problems, but who is unbeknownst to him being helped by a wizard who ensures that all his beliefs formed on this basis are true" (p. 61). Would we readily say "I advise anyone in a similar position as the careless mathematician to do just as he does in believing that each of his mathematical beliefs is true"? Probably not. And this is plausibly because we at least implicitly think of positions where a careless mathematician is being *hindered* by some unseen and unknown agent as "similar". (After all, the careless mathematician is being deceived about the causes of his mathematical success: It doesn't seem to introduce too much of a difference to have him be deceived about at least some mathematical propositions as well.)

Yet as some philosophers have argued, the virtue condition seems to require *too much* for knowledge: Suppose that Temp is in a room with an accurate electronic thermometer functioning normally. Like most of us, he's not aware of exactly how electronic thermometers work. Is the truth of his belief that it's 69°F "to a significant degree creditable to his cognitive character"? It's hard to see how it is.[83] Pritchard indicates (in the course of his reply to a similar concern) that he would insist that the truth of Temp's belief here is at least *somewhat* creditable to Temp, since Temp forms the belief by looking at a *thermometer*, and not by (say) reading tea leaves (cf. Pritchard 2010, p. 41). So perhaps what Pritchard's virtue condition amounts to is this: K *represents* mental states which are in fact good reasons for her belief *as* reasons for her belief—and, moreover, the fact that these mental states accurately reflect reality isn't due to the whims of something or someone else.

But if this is what the condition amounts to, then SAGE can perfectly well explain its appeal and success: If, for instance, K *doesn't* represent her relevant mental states as reasons for her belief, then by KP_M there's a natural interpretation of K's position such that we wouldn't want to convey the unqualified message "I would advise anyone in K's position to do just as K did" (an implication of the full prospective message, "I would advise anyone in a similar position to do just as K did"). For instance, if Temp doesn't represent his visual perception of the thermometer as a reason for his belief, we would seem to be conveying, "I would advise

[83] A similar point against virtue conditions on knowledge is made by Lackey (2009), who points to cases where we putatively gain knowledge through testimony, such that the truth of our belief is due fully to the cognitive character of *someone else*.

anyone who has a visual perception as of a thermometer reading 69°, but who doesn't represent this as providing any reason to believe that it's 69°, to believe that it's 69°". And we presumably wouldn't want to convey this: When people don't represent a mental state which otherwise seems to provide evidence that *p* as a reason to believe that *p*, typically they have *good reasons* (even if those reasons aren't currently conscious) to be suspicious of the mental state in question.

Finally, let's revisit Lewis's (1996) contextualist account of knowledge:

> K knows that *p* iff K's evidence eliminates every possibility in which ~*p*—Psst!—except for those possibilities that we are properly ignoring. (p. 554)

K's evidence (that is, her perceptual experiences and memories) "eliminates" possibility W just in case if W were in fact the actual situation, K's evidence would be different than it actually is (p. 553). Lewis posits, among other things, that possibilities which "saliently resemble" the actual situation are never properly ignored. SAGE gives a straightforward explanation for the appeal of this "Rule of Resemblance": If there is a possibility which "saliently resembles" K's actual situation and in which ~*p*, but which K's evidence does not eliminate, then this ~*p* alternative is one in which Alt-K's relevant mental states are the same as K's actual mental states. Thus, by SIM Alt-K's position is very likely to strike us as similar to K's position. And according to SAGE, since we wouldn't advise Alt-K to do the same as K did (since that would be encouraging Alt-K to form a false belief), we deny that K knows.[84]

Lewis's Rule of Resemblance arguably takes care of Pritchard's Temp case: The alternative possibility in which the hidden person is deceiving Alt-Temp about both the

[84] In Chapter 5, I'll show that THYME can explain the appeal of one of Lewis's other rules—the "Rule of Attention"—in the course of discussing skepticism.

thermometer *and* the temperature does seem to saliently resemble the actual situation (in which Temp is being deceived only about the workings of the thermometer). So even though Alt-Temp's world isn't "nearby" to Temp's world, the Rule of Resemblance yields the verdict that Temp doesn't know that it's 69°, due to this uneliminated but not-properly-ignored possibility in which it's not 69°. However, because Lewis's account of knowledge entails that only *uneliminated* possibilities can stand in the way of knowledge—that is, only possibilities where Alt-K has the exact same perceptual experiences and memories as K—the account doesn't suffice to rule that unjustified beliefs in necessary truths don't constitute knowledge.[85] For instance, suppose that Erma was a mathematician in the early 1900s who came to believe firmly that Fermat's Last Theorem is true, despite no one having offered a successful proof of it. (In fact, the theorem wasn't proven until 1994.) Erma felt that it just "must" be true. Now I doubt that many, if any, philosophers would grant Erma knowledge here. But Lewis's account yields the verdict that Erma *does* know: The theorem is true in *all* possibilities, so there is *a fortiori* no uneliminated possibility in which it's false.

SAGE can diagnose why Lewis's account doesn't align with the philosophical verdict here: Even if Alt-K's relevant mental states are slightly different in details from K's, Alt-K's position may still be intuitively similar to K's (as per SIM). Hence, if we consider conveying the message "I would advise anyone in a position similar to Erma's to do just as she did", we'll likely recognize that there are cases where the agent's position is similar to Erma's, but in which the mathematical proposition which they feel "must be true" is actually false. For instance, perhaps

[85] To be fair, Lewis seems to admit as much (p. 552), but doesn't take this to count against his theory overall. And he may be right. But my point here is that we'd like an explanation for why his account doesn't accord with the philosophical verdict on *every* case.

Alt-Erma feels that Matt's Last Theorem "must" be true, despite being aware of no proof for it. Unfortunately for Alt-Erma, Matt's Last Theorem, although *prima facie* compelling, *isn't* true. Erma and Alt-Erma are intuitively in similar positions, since their relevant mental states are fully analogous: a feeling that a particular, *prima facie* compelling mathematical theorem must be true, and no memories of anyone offering a successful proof of the theorem. Thus, SAGE and SIM posit that there are some cases where philosophers deny that K knows because there are alternative situations where Alt-K has *slightly different* mental states, but in which we wouldn't advise Alt-K to do as K did. Lewis's account, as stated, cannot account for the verdict on any such case.

CHAPTER V: Beyond the Armchair

5.1 Linguistic Sensitivity

Throughout Chapters 3 and 4, I showed that my hypothesis SAGE can explain the verdict of philosophers on a wide range of puzzling cases offered in the epistemology literature. I've suggested that based on the representative sample of cases we've considered, we have strong inductive grounds to believe that SAGE can explain the philosophical verdict on *any* case in the literature, present or future. But now what about the verdicts of speakers who aren't professional philosophers? Do they agree that K knows in all the cases where philosophers ascribe knowledge, and deny that K knows in all the cases where philosophers deny knowledge? If not, why not? In this chapter, I'll consider whether my hypotheses THYME and SAGE can help explain several results obtained recently by self-labeled experimental philosophers, who have asked non-philosophers about their judgments on the sorts of hypothetical cases discussed in the epistemology literature. I'll focus first on results which indicate significant group-level differences between philosophers and non-philosophers. But I'll also look at whether THYME can explain several puzzling features of knowledge ascription practices which I haven't yet discussed and which are, by all accounts, shared by philosophers and non-philosophers.

To begin: Several philosophers have recently conducted experiments to test whether non-philosophers share the verdicts of philosophers on several Gettier-type cases, of the sort discussed in Chapter 4: hypothetical cases where K's belief is true and justified, yet where a significant proportion of philosophers deny that K knows. Many of the results have been surprising. For instance, Weinberg et al. (2001) report a 40% knowledge ascription rate among 112 non-philosophers in a "classic" Gettier-type case. Turri et al. (2015) asked large groups of non-philosophers participants for their verdicts on cases where philosophers would unanimously deny that the believer knows. In three of these cases, the knowledge *ascription* rates among the participants were 38%, 39%, and 55%, respectively. Similarly, Starmans and Friedman (2012) found knowledge ascription rates of about 70% among groups of around 100 non-philosophers in cases like the following:

> Katie is in her locked apartment writing a letter. She puts the letter and her blue Bic pen down on her coffee table. Then she goes to the bathroom to take a shower. As Katie's shower begins, two burglars silently break into the apartment. One burglar takes Katie's blue Bic pen from the table. But the other burglar absentmindedly leaves his own identical blue Bic pen on the coffee table. Then the burglars leave. Katie is still in the shower, and did not hear anything. (ibid., p. 276)

Based on my reading of the literature, no philosopher would agree that Katie continues to know that there's a blue Bic pen on the table, despite her belief being true and justified.

Can the hypotheses I've proposed so far help to account for the apparently significant group-level differences between philosophers and non-philosophers when judging Gettier-type cases?[86] I believe so. First, recall my central hypothesis THYME and its corollary SAGE:

> (THYME) Our words "certainty" and "belief" (in at least some uses) are descendants of words that our early ancestors used to describe the subjective "no doubt" state arising from an evolved subconscious trust heuristic. Our word "know" is the descendant of a word our early ancestors used to urge others to be either certain or uncertain that *p*, and subsequently also to convey that third parties were already certain that *p*. Typically,

[86] To be clear, none of the researchers just cited proposes an *explanation* for the group-level differences they find between philosophers and non-philosophers. Rather, they only suggest that their results provide some evidence against the assumption that any believer in a Gettier-type case (no matter what the details or structure of the case) lacks knowledge.

> our ancestors' desire to urge others in this way was due to their own evolved trust heuristics having produced (or not produced) trust that *p*.
>
> (SAGE) Speakers today at least implicitly expect their listeners to interpret "K knows that *p* at t" to convey the prospective advisory message "I would advise anyone in a position similar to K's at t to do the same sort of thing K did in believing that *p*".

THYME and SAGE leave open the possibility that some speakers today are *more sensitive* to the prospective advisory message which we all implicitly understand knowledge ascriptions to convey. That is, some speakers are (1) more tuned in (albeit subconsciously) to whether there might be positions "similar" to K's in which they wouldn't want to advise Alt-K to do as K did, and/or (2) more concerned (albeit subconsciously) about conveying a message to their listeners which would seem to reflect poorly on their judgment. Moreover, this sensitivity might vary by context, depending on such things as the speaker's current level of risk aversion, what they believe about their audience, and whether they've thought systematically about knowledge ascriptions in the past.

What I'd like to propose is that at least in contexts where a hypothetical and somewhat fanciful case of belief is under consideration, philosophers (as compared to non-philosophers) tend to be both more *alert* (at least subconsciously) to how their audience might interpret their words, and more *risk-averse*: they are primed (at least subconsciously) to take measures to avoid being interpreted in ways they don't want to be. Let's say that as someone becomes either more alert or more risk-averse (or both) regarding possible listener interpretations of their speech, they become more "linguistically sensitive". Here, then, is an official statement of the auxiliary hypothesis I've just proposed:

> (SENS) Philosophers tend to be more linguistically sensitive than non-philosophers, at least in situations where the question of whether or not to ascribe knowledge to a given believer is explicitly under consideration.

Consider the following observations in support of SENS: Modern analytic philosophers are, almost as a rule, alert to possible misinterpretations of their spoken or written words in professional settings, and motivated to avoid or correct such interpretations. Articles in contemporary analytic philosophy journals tend to contain rigorous definitions of important terms used, along with parenthetical remarks and footnotes to ensure against misunderstandings. Carefully managing the interpretation of one's language—often at the fine-grained level of individual words—is part of the professional responsibility of philosophers. Non-philosophers, however, may not be as alert and/or averse to possible unwanted interpretations of their knowledge attributions—at least not in the contexts in which they first explicitly consider whether to ascribe knowledge to a believer in a hypothetical vignette (say, in taking a survey given to them by an experimental philosopher) [87]. This is not to say that non-philosophers are less linguistically sensitive than philosophers *in all circumstances*—just that each of us has our own habits and patterns of vigilance, depending on on our experiences, personalities, and domains of expertise.

It's plausible, then, that philosophers are on average more vigilant about unwanted interpretations than non-philosophers when explicitly asked about their knowledge intuitions—since for philosophers, this occurs almost exclusively in professional contexts (for instance, in a philosophy classroom, or in the course of writing a paper or article). And if SENS and SAGE are correct, we can begin to see a causal story leading to the observed result that non-philosophers

[87] Even if taking a survey activates linguistic risk-aversion (say, because one doesn't want one's responses to be misconstrued by the experimenter), it may not necessarily suffice to activate alertness to possible unwanted interpretations of a word which is used so causally and frequently as "knows".

ascribe knowledge in Gettier-type cases at higher rates than philosophers. The story is this: Philosophers are more likely to at least subconsciously register an aversion to conveying the message "I would advise anyone in a similar position to do just as K did" in Gettier-type cases—because their implicit judgments of similarity are more capacious (so as to anticipate the possibly capacious judgments of their listeners), and/or because they are more implicitly concerned about their listeners not giving them the "benefit of the doubt" when it comes to interpreting what they mean to convey. Hence, philosophers are more likely to deny that a hypothetical believer in a Gettier-type case has knowledge.

Now, if this SENS-based explanation is correct, then we should see *higher* rates of knowledge denial among non-philosophers if efforts are taken to make them more linguistically sensitive than usual (that is, to make them "more like philosophers"). Is this prediction of SENS and SAGE borne out by the data? In fact, the results of Turri (2013) arguably confirm SENS: Turri surmised that "philosophers notice … features of the [Gettier-type] cases that untutored laypeople overlook" (p. 2).[88] He predicted that by partitioning Gettier-type cases into three physically separated blocks of text, he could lead non-philosophers participants to notice the relevant features which similar participants apparently overlooked in previous studies (such as Starmans and Friedman 2012). Turri took these relevant features, in at least some Gettier-type cases, to be the following "bad luck/good luck" structure:

> Start with a belief that is well enough justified to satisfy the justification condition on knowledge. All seems well. Then introduce an element of bad luck that would normally prevent the justified belief from being true. All seems ill. Then introduce a conspicuously distinct element of good luck that makes the belief true anyway. (p. 2)

[88] Although he doesn't think that the relevant feature is whether there are positions intuitively similar to K's in which we wouldn't advise Alt-K to do as K did. Below, I'll discuss Turri's proposal of what the relevant features are.

Accordingly, Turri presented some of his non-philosopher participants with "bad luck/good luck"-type cases with each of the three stages (justification, bad luck, good luck) on a separate screen. (Participants took the survey on a computer.) Others received the same cases with no text partitioning. For instance, here are the three stages Turri used for one experiment (with two alternative third stages, which I'll discuss below):

> *Stage 1*: Katie is in the living room of her locked apartment writing a letter with a blue Bic pen. She puts the letter and the blue Bic pen down on her coffee table. Then she goes into the bathroom to take a shower. It takes her fifteen minutes to finish.
>
> *Stage 2*: Just after Katie started her shower, two burglars, a master and his apprentice, broke into her apartment. As they made their way around the apartment, the master burglar stole Katie's blue Bic pen from the coffee table. After five minutes, the burglars left, well before Katie finished her shower. Katie did not hear anything.
>
> *Stage 3—Burglar*: Right before the burglars left Katie's apartment, the apprentice burglar started feeling a bit dizzy, so he sat down on the couch for a moment to recover. When the apprentice burglar sat down, he absentmindedly set his own blue Bic pen on the coffee table, and forgot it there. This was five minutes before Katie finished her shower.
>
> *Stage 3—Husband*: Right after the burglars left, Katie's husband came home. Tired from a long journey, he put his wallet, keys and his own blue Bic pen down on the coffee table in the living room. Then he immediately lay down on the living room couch and fell asleep. This was five minutes before Katie finished her shower. Katie hasn't yet noticed that her husband is even home. (pp. 8-9)

For the Control condition, Turri presented 28 non-philosopher participants with Stage 1, Stage 2, and Stage 3—Burglar, with no textual separation. (That is, all three stages were joined together in one paragraph.) 57% responded that Katie "really knows", as she's getting out of the shower, that there is a pen on the coffee table; the other 43% chose the other option, "only thinks". Starmans and Friedman (2012) found that 69% of about 25 subjects responded "really knows" for the same vignette, presented all in one paragraph. (Turri purposely used the text from Starmans and Friedman in order to compare his results with previous ones.)

But when Turri presented 23 new participants with Stage 1, Stage 2, and Stage 3—Burglar, *each appearing on a different screen*, only 44% responded “really knows”—a statistically significant difference from Control (p = 0.04). And among a further 23 new participants who read Stage 1, Stage 2, and Stage 3—Husband, each appearing on a different screen, only 24% responded “really knows”. These results are encouragingly consistent with SAGE and SENS: By separating the three phases of Katie’s story into three different passages, Turri forced his participants to digest the case more slowly. The additional time may have allowed some of these participants, who wouldn’t otherwise have done so, to subconsciously register an aversion to conveying the message “I would advise anyone in a similar position to do as Katie did in continuing to believe there’s a pen on the table”. (In particular, the extra time may have prevented activation of the participants’ “justified true belief” *heuristic* for knowledge ascriptions.) After all, if SIM is on the right track, then at least when we’re tuned in to the prospective advisory message of knowledge ascriptions, we’ll likely register that the following constitutes a position similar to Katie’s: (1′) a memory of placing a pen on a table several minutes ago; (2′) no sensory perceptions indicating that the pen might not still be on the table; (3′) the fact that a burglar has stolen Alt-Katie’s pen; (4′) the fact that another burglar has absentmindedly left a *pencil* on the same table. Plausibly, the enforced slowdown in reading the vignette induced greater linguistic alertness among Turri’s participants, giving them more time to implicitly register that there are alternative situations—along the lines of (1′)-(4′)

above—where Alt-Katie's position is similar to Katie's, but in which they wouldn't advise Alt-Katie to do just as Katie did. [89]

Thus, Turri's results arguably provide some confirmation of SENS: Its predictions are so far borne out, so it continues to be a viable explanation for the typically lower knowledge denial rates among non-philosophers than among philosophers in Gettier-type cases. Of course, future empirical work could offer further significant tests of SENS's predictions.

5.2 The Consequences of Certainty

In this section, I'll consider some further recent empirical findings that indicate both similarities and differences among the knowledge ascription patterns of philosophers and non-philosophers. I'll propose that THYME can help explain the relevant patterns we see both among both groups.

First, here's an observation about philosophers: Many philosophers have noted their greater hesitation—if not outright denial—to ascribe knowledge that *p* to a given believer K when the *non-epistemic stakes* of believing that *p* are particularly high for K, in light of her situation and the actions available to her. By the "non-epistemic stakes of believing that *p* in situation S", I mean the goodness or badness of the physical consequences of taking actions which are readily available in S and whose success relies on the truth of '*p*', if in fact it turns out that ~*p*. For instance, DeRose (1992) says he would not ascribe knowledge to someone who believes that a bank will be open next Saturday—based on seeing the bank open on Saturday

[89] Why the even lower rate of knowledge ascription in the version where Katie's husband is the one who puts a pen on the table? It could be that the fact of the second burglar having an identical blue Bic pen struck some participants as "meant to be" (that is, somehow causally ensured, and not just a coincidence), such that it was hard to implicitly think of alternative situations which lack this feature as "similar" to Katie's.

two weeks prior, and assuming that the bank won't change its hours—and who will face *severe financial consequences* if she's wrong. (DeRose reports that he has this intuition even when assuming that the bank *will* be open next Saturday.) By contrast, DeRose doesn't find the same knowledge ascription problematic if the person in question doesn't face any serious consequences if the bank changes its hours (ibid., pp. 913-14). Other philosophers have expressed similar intuitions, among them Hawthorne (2004), Stanley (2005), Fantl and McGrath (2009), and Weatherson (2011). Moreover, based on his extensive review of the literature and his anecdotal impressions, Pinillos (2012) is aware of only one author (Schaffer 2006) who seems not to share the stakes-sensitive hesitation reported by others.

I'll consider evidence for similar stakes-sensitive hesitation among non-philosophers in a moment. For now, notice that THYME offers a straightforward explanation for the hesitation widely reported among philosophers: Suppose, as THYME proposes, that our word "know" evolved through memetic evolution such that we tend to ascribe knowledge that *p* to K only if we judge that we ourselves would be certain that *p* if we had only K's evidence—where this judgment is typically mediated by our evolved trust heuristic.[90] If this is so, then it's clear why our intuitive willingness to ascribe knowledge depends, at least sometimes, on the *non-epistemic* stakes of trusting that *p* for a given believer in a given situation. After all, trusting that *p* has implications for our future actions, since trusting that *p* (as I've defined it) means not making any contingency plans for the event that ∼*p*, and not taking the possibility that ∼*p* into

[90] What if K isn't aware of the high stakes of her situation? Many philosophers still report strong hesitation to ascribe knowledge in such cases. But THYME and SAGE still offer an explanation, based on the idea that we understand knowledge ascriptions to convey, "I would advise anyone in a similar position to do just as K did", where "position" may be ambiguous between K's evidence alone and further relevant facts of which K wasn't aware.

account (consciously or otherwise) when deciding what to do. And our future actions can have implications for our survival, reproductive abilities, and capacity to protect and nurture our biological relatives. Thus, if our willingness to ascribe knowledge to some third party K is indeed typically mediated by our own evolved trust heuristic, we can see why philosophers hesitate to ascribe knowledge to K in cases where the consequences of K trusting that *p* would likely be particularly bad for K if in fact it were the case that ~*p*. The causal link here is that our trust heuristic evolved to promote *survival and reproduction*, not merely true beliefs.

Let's condense this idea into another auxiliary hypothesis:

> (STAKE) Due to the pressures of natural selection, our trust heuristics evolved to generally not produce all-out trust that *p* when the negative physical consequences of being wrong about '*p*' are particularly high.

THYME and STAKE offer a clear explanation of the stakes-sensitive hesitation we see among philosophers. However, if THYME and STAKE are true, we won't expect this hesitation to be unique to philosophers: *All* of us descend from ancestors facing similar evolutionary pressures. The trust heuristics of non-philosophers should bear the same imprints of natural selection.

So do we see increased hesitation to ascribe knowledge among non-philosophers when the non-epistemic stakes are high? By "increased hesitation", I mean that either (1) a higher percentage of the population is disposed to deny knowledge in high-stakes cases than in similar cases where the non-epistemic stakes aren't so high, and/or (2) individuals tend to feel more reluctance to ascribe knowledge in high-stakes cases than in otherwise similar low-stakes cases—even if they do ultimately ascribe knowledge in both cases.[91] Feltz and Zarpentine

[91] As we'll see below, some researchers have attempted to measure hesitation of type (2) by asking participants to rate their *level* of agreement with a statement of the form "K knows that *p*" or "K doesn't know that *p*", using a Likert-type scale (e.g. strongly disagree, disagree, neither agree nor disagree, agree, strongly agree).

(2010) presented several cases with varying non-epistemic stakes to non-philosopher participants. Each participant received a “high-stakes” or a “low-stakes” version of a particular vignette. For instance, one of the case pairs was as follows:

> *Low Stakes*: Bill, Jim, and Sarah are hiking and they come to a ravine. There is a bridge five feet over the ravine. Bill sees Sarah and Jim cross the bridge, and Bill says to Jim, “I know the bridge is stable enough to hold my weight”.
>
> *High Stakes*: Bill, Jim, and Sarah are hiking and they come to a ravine. There is a bridge one hundred feet over the ravine. Bill sees Sarah and Jim cross the bridge, and Bill says to Jim, “I know the bridge is stable enough to hold my weight”. (pp. 689-90)

Non-philosopher participants were assigned to either the high-stakes or low-stakes version of a given vignette, and asked to rate their level of agreement on a 1-7 scale (1 = strongly agree, 7 = strongly disagree) with the knowledge ascription made in the vignette (e.g., Bill’s statement “I know the bridge is stable enough to hold my weight”). Across all case-pairs tested, Feltz and Zarpentine did *not* find a significant difference between the mean level of agreement with the knowledge ascription in the low-stakes case and the high-stakes case. Beebe and Buckwalter (2010) and May et al. (2010) report similar null results with similar high-stakes/low-stakes case pairs presented to non-philosophers.

This does appear to be a strike against THYME and STAKE. However, Pinillos (2011) points to a potential confounder in all three of the studies finding null results: In each of the cases tested, one of the protagonists *explicitly ascribes knowledge* to the believer in question (either themselves or someone else). Yet there is both theoretical and empirical support for the phenomenon of “accommodation”, by which we tend to interpret the assertions of others, if possible, so as to make them true. (Buckwalter 2011 finds very convincing empirical evidence for accommodation of knowledge ascriptions and knowledge denials.) Thus, given that in each tested case-pair a protagonist states “I/you/she know(s) that *p*”, it may be that the null findings are an artifact of accommodation, rather than reflecting whether participants would ascribe knowledge (and/or how hesitant they would be in doing so) without the protagonist’s precedent. In light of this methodological concern, Pinillos (2012) used a different experimental technique: He gave non-philosopher participants the open-ended question of *how much* evidence someone would need to gather before knowing that a given proposition was true. Pinillos varied the stakes for the evidence-collector across different versions of a given case. For instance, in one version of a case, a student who was about to proofread his college term paper would not face any major consequences from missing a typo; in a second version, his professor would not give an A to any paper with even one typo. Participants in each case were asked, “How many times do you think [the student] has to proofread his paper before he knows that there are no typos?” Pinillos found a strong and statistically significant difference in the amount of evidence which participants in the low-stakes vs. the high-stakes version of the case said the student would need to gather before having knowledge: a median of 2 proofreads in the low-stakes version, compared to a median of 5 in the high-stakes version. Pinillos proceeded to obtain similar results when he gave a new set of participants both the high-stakes and low-stakes version of a given case *at the same time*, and asked for their judgment on how much evidence the protagonists in each version would need to collect.[92]

Given the apparent tension between the earlier null results and those of Pinillos—as well as lingering methodological concerns about each of the studies mentioned here—we need

[92] To be clear, Pinillos does not propose an evolutionary explanation (of the sort THYME offers) for why knowledge intuitions might vary with non-epistemic stakes. He simply argues that his findings provide some evidence for thinking that the property “knowledge” is sensitive to non-epistemic stakes.

more empirical work to determine whether non-philosophers do in fact tend to hesitate more to ascribe knowledge when the non-epistemic stakes are high. The results of that further work will be important tests for the plausibility of THYME, given its predictions about the sensitivity of knowledge ascriptions to non-epistemic stakes. To be sure though, if THYME is correct, we should still expect to see some individual variation in knowledge ascription tendencies, due to genetics, epigenetics, and personal experience. Thus, THYME won't be disconfirmed if not *everyone* is disposed to hesitate more in ascribing knowledge in *every* high-stakes case. In this vein, recall the fact that at least one philosopher, Schaffer (2006), reports that he does *not* have the intuition that knowledge possession changes across otherwise similar cases with different non-epistemic stakes—at least once he concentrates on the case-pairs closely. THYME may be able to explain Schaffer's intuitions, on the plausible assumption that we can sometimes consciously "override" the verdicts of our trust heuristics (that is, the verdict yielded when we simulate being in someone else's position) when we're in certain theoretical mindsets—such as when we consciously resolve, as epistemologists sometimes do, to ignore all dimensions of value other than the quantity and proportion of our beliefs that are true. In general, THYME leaves open the possibility that when we're *consciously* deliberating about whether to ascribe knowledge to someone else, our verdict may be influenced not only by our trust-heuristic-mediated sense of whether we would be certain that *p* if we were in that person's position (or a similar one), but also by such factors as our current linguistic context, mood, and levels of empathy and risk-aversion. However, I'll leave for future work the task of carefully explaining such "outlier" intuitions as Schaffer's.[93]

As we await further empirical work on non-epistemic stakes, let's turn to an existing finding about the knowledge ascription tendencies of non-philosophers—a finding which clearly conflicts with most or all philosophers' conception of knowledge but for which THYME offers a plausible explanation. This finding, now called the "epistemic side-effect effect" (ESEE), involves cases where K has some evidence that an action she might take will have a certain side-effect which isn't her primary aim. When K takes the action, the side-effect does in fact occur. Researchers have repeatedly found that non-philosophers agree much more strongly that the person *knew* the side-effect would occur if the side-effect was morally bad than if it was morally neutral or good. (In each pair of cases presented to participants, the evidence for the future occurrence of the side-effect—good or bad—is the same.) For instance, Beebe and Buckwalter (2010) presented 749 non-philosophers with one version of the following case (half received the "help" version and half received the "harm" version):

> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits, and it will also *help/harm* the environment." The chairman of the board answered, "I don't care at all about *helping/harming* the environment. I just want to make as much profit as I can. Let's start the new program." They started the new program. Sure enough, the environment was *helped/harmed*. Did the chairman know that the new program would *help/harm* the environment? (pp. 475-76)

Participants who received the 'harm' version were significantly more likely to agree with the statement that the chairman knew the side-effect would occur than were participants who

[93] None of this is to say that Schaffer is necessarily *wrong* when he reports the equivalence of his knowledge intuitions across high-stakes and low-stakes cases. Recall that THYME is a hypothesis about what *causes* us to ascribe and deny knowledge as we do, not a hypothesis about the actual truth conditions for knowledge ascriptions. (I'll return to this point in Chapter 6.)

received the ‘help’ version (92% vs. 70%, p < .001). These results were replicated by Beebe and Jensen (2012) and Turri (2014).

Philosophers have been puzzled by these findings. After all, the evidence for the side-effect’s occurrence—namely, the vice-president’s report—seems to have the same probative value (i.e., probabilistic relevance) in each version of the case. Moreover, it’s puzzling why only 70% of participants judged that the chairman had knowledge in the Help case: Why not 100%, given the very clear evidence from the vice-president? As to this second question, it seems plausible that at least some participants inferred from the chairman’s statement “I don’t care at all about helping the environment” that she never *formed the belief* that the environment would be helped. Moreover, some participants may not have regarded the vice-president’s report as decisive evidence. (After all, the vice-president doesn’t provide any facts to back up his assertion.)

Now, as to the much higher rate of knowledge ascription in the “harm” condition: Suppose that our willingness to ascribe knowledge is, as THYME proposes, mediated primarily by an evolved trust heuristic (at least if we don’t have prior theoretical commitments to a particular view about the concept *knowledge* which are consciously salient at the moment). Then we seem to have a plausible explanation for non-philosophers’ much greater willingness to ascribe knowledge of a possible side-effect when the side-effect is morally noxious: It’s plausible that our early ancestors who tended to *fully trust*, on the basis of decent evidence, that an action they were contemplating would have a very bad negative side-effect on other people had an advantage over those who didn’t become similarly certain unless they had *decisive* evidence. After all, the former generally would have planned for the worst—or would not have taken the action at all—and thereby would have faced less social opprobrium (or worse).[94] Let's frame this idea as an auxiliary hypothesis:

> (SIDE) Our trust heuristics evolved to produce all-out trust that a certain action we're considering will have a morally noxious effect on other people, even if our evidence for this side-effect is merely decent and not decisive.

THYME and SIDE now offer an explanation for the difference between the "help" and "harm" conditions: It's very likely, based on the results in the "help" condition, that many of the participants in the "harm" condition would have denied that the chairman knew the side-effect would occur had the side-effect been helpful: either because it's not clear that she *believed* that the side-effect would occur, or because it's not clear that her evidence was decisive. But presumably, many of those participants in the "harm" condition who inferred that the chairman didn't believe the environment would be harmed still wanted to convey the message that she *should* have believed that the environment would be harmed—and hence should have acted on that assumption. As we saw in § 2.3, we do sometimes use "know" in a hortatory sense, as in "Oh, you knew that I'd be mad if you didn't take your shoes off before walking into the house". Moreover, consider any participant who assumed the chairman believed (or at least was pretty sure) that the environment would be harmed, but who didn't regard the evidence as decisive. If SIDE is correct, then these participants likely still would have wanted to convey the message, "I would advise anyone in a similar position to do as the chairman did in being certain that the environment would be harmed (despite her evidence not being decisive)". And if THYME is correct, then saying "The chairman knew the environment would be harmed" is a way of

[94] I'm assuming here (as I think is plausible) that the disadvantages of generally being more cautious in situations where the welfare of other people was at grave risk did not outweigh the advantages.

conveying exactly that message. Thus, if THYME and SIDE are correct, we have an explanation for why most or all of those participants in the “harm” condition who otherwise would have denied that the chairman knew the side-effect would occur (had the side-effect been positive) instead ascribed knowledge.

One more recent empirical finding deserves mention here: Shin (2014) assigned 81 non-philosopher participants to one of two stories about “Sally”, a medical student who needs to decide whether to prescribe medicine A, B, or C to her patient Harry for a persistent cough. In the “less time” version, Sally needs to make a decision within two minutes. Within those two minutes, she suddenly remembers reading in a textbook that C is a very good treatment for the sort of cough Harry has, and she thereby comes to believe firmly that C is the best medication for Harry out of A, B, and C (and she’s right). In the “more time” version, Sally has four months to decide which medication to prescribe (since all three are on order and won’t arrive for a while), yet she doesn’t do any research in the interim. When the time comes for her to make her decision after four months, she suddenly remembers reading in a textbook that C is a very good treatment for the sort of cough Harry has, and she thereby comes to believe firmly that C is the best medication for Harry (and she’s right) (ibid., pp. 160-61).

Shin asked participants to rate on a seven-point scale how strongly they agreed with the statement that Sally knew that C was the best medication for Harry (0 = strongly disagree, 3 = neither agree nor disagree, 7 = strongly agree). Those who received the “less time” version were on average significantly more in agreement that Sally had knowledge than those who received the “more time” version (3.55 vs. 2.67, p = .014). As with the ESEE findings, these results are puzzling if we conceive of knowledge (as many philosophers do) as generally sensitive only to the probabilistic relevance of our evidence to the proposition in question. After all, Sally has the exact same evidence in both versions. Moreover, it’s somewhat puzzling that participants in the “less time” condition were overall in agreement that Sally knew C was the *best* medication: Her memory indicated that C was a *very good* medication for Harry’s condition, but she didn’t have evidence about the *comparative* effectiveness of A, B, and C. But Shin’s results make good sense in light of THYME’s proposal that our willingness to ascribe knowledge is typically mediated by our willingness to convey the message, “I would advise anyone in a similar position to do just as K did”. First, consider the “less time” condition: It may be that some participants simply didn’t think about the distinction between (1) C being a very good medication, and (2) C being the *best* medication. (This distinction wasn’t explicitly highlighted in the text Shin gave to his participants.) But also, some participants who noted the distinction might still have registered no hesitation to convey, “I would advise anyone in a similar situation to do just as Sally did in being certain that C was the *best* medication”. After all, she has decent evidence that C is a *very good* medication for Harry’s condition, so the stakes of being wrong about C being the *best* don’t seem all that high.

Now consider the “more time” condition: Based on the “less time” results, it’s very likely that many of the “more time” participants would have agreed less with the statement “Sally knew that C was the best mediation” had they been given the “less time” version. But it’s plausible that many of these “more time” participants would not want to convey the message, “I would advise anyone in a similar position to do just as Sally did in believing that C was the best medication”. For in the “more time” condition, “doing what Sally did” might well seem to mean “not collecting evidence about a pertinent decision until moments before the decision needs to be made, despite having months to collect more decisive evidence". And if that's how we implicitly conceive of "doing what Sally did" (or if that's at least *one* of the interpretations that becomes salient to us), then we almost surely won't want to convey the message "I would advise anyone in a similar position to do just as Sally did". Now if THYME is correct, then we at least implicitly recognize that ascribing knowledge to Sally conveys that very message. Thus, THYME explains why those participants in the "more time" condition who otherwise would have agreed —or agreed more strongly—that Sally had knowledge (that is, had the timeframe for her decision been shorter) instead denied that she knew.[95]

5.3 Skepticism and its Evanescence

As discussed in Chapter 2, philosophers and non-philosophers alike are generally quite willing to attribute knowledge to those who form beliefs through fallible processes like sensory perception, testimony, and inductive inference (projecting that a hitherto reliable pattern will continue in the future). And yet, generations of philosophers have remarked how quickly they can lose their willingness to ascribe knowledge to themselves or others. To take only some contemporary examples, Nagel (2011) compiles the following pairs of propositions from several cases discussed recently by epistemologists:

> (1a) Lucy knows that the current US President is Barack Obama.
> (1b) Lucy knows that Barack Obama has not had a fatal heart attack in the last five minutes.
>
> (2a) Looking at the striped animal in the zoo, Fred knows that it is a zebra.
> (2b) Looking at the striped animal in the zoo, Fred knows that it is not a cleverly disguised mule.
>
> (3a) Smith knows that his car is on Avenue A, where he parked it an hour ago.

[95] To be clear, Shin doesn't consider an explanation of the sort I've offered for his results. He simply takes his results as evidence to support the claim that the property "knowledge" is sensitive to factors other than the probabilistically-relevant features of the believer's situation.

> (3b) Smith knows that his car is not one of the hundreds of cars that are stolen every day across the country. (ibid., p. 5)

Many philosophers, Nagel included, have remarked on their hesitation to agree with statements of type (1b), (2b), and (3b)—and then consequently to feel less certain about the knowledge ascriptions in the corresponding (a) statements. Several empirical studies have found similar trends among non-philosophers: Once the mere possibility of a break in a reliable pattern has been explicitly mentioned, subjects are much less inclined to ascribe knowledge to someone who believes that the pattern will continue—even if there's clearly no significant likelihood of the pattern being disrupted. (See, for instance, Nagel 2012.)

What could explain this sort of "induced" skepticism among both philosophers and non-philosophers? If THYME is correct that our knowledge ascriptions are typically mediated by our own evolved trust heuristics, then I think we can readily explain our bouts of skepticism. For consider the times when our early ancestors had the occasion to *consciously* think about the possibility of a break in a hitherto reliable pattern (such as the pattern of the berries they typically ate not making them sick, or of their community never suddenly turning against them), or to think about the possibility that the information coming from their senses or an informant was misleading on a particular occasion. We can imagine that such conscious reflection was typically occasioned by *actual evidence* of a break in the pattern or of the unreliability of one's senses or one's informant. Or, if such ancestors didn't have any positive *evidence* for being skeptical, it was likely the case that the non-epistemic stakes were particularly high for them at the time.

Consider that given the much higher rate of salient sensory perceptions of nature that our ancestors had (due to their being outdoors most or all of the time), the greater number of decisions they had to make every day on which their health and survival hinged, and their less frequent use of language, our ancestors probably had much less time or occasion to consider low-probability possibilities of breaks in reliable patterns, or of misleading information from their senses or from informants. When the possibility of a pattern disruption or misleading information actually made it to their conscious awareness, it's very likely that it was prompted by a recognition that the non-epistemic stakes of being wrong were very high, and/or by evidence (including perhaps just a gut feeling) that a pattern disruption was significantly likely in that case. For instance, we can imagine that our ancestors didn't typically have the time or occasion to consider the possibility that a stream on which they'd always relied on for water might suddenly run dry—*unless* they (for instance) suddenly found themselves without any other source of water, or observed that there hadn't been any rain for weeks.

Thus, we can expect that our trust heuristics evolved to generally reduce trust in a previously trusted proposition whenever a specific possibility of that proposition being false—consistent with our current sensory perceptions—enters *conscious* consideration. However, it's also plausible that our trust heuristics evolved to *reinstate* trust in a given proposition if, after consciously considering an adverse possibility, we at least subconsciously recognize that the non-epistemic stakes aren't high after all, and that there actually doesn't seem to be a significant likelihood of the adverse possibility. After all, if our ancestors had continued not trusting that *p indefinitely* after considering adverse possibilities that weren't in fact likely or high-stakes, they would have been at considerable disadvantage—particularly as their leisure time and use of language increased (and hence as they had more time to consider "idle" possibilities). Let's organize these ideas into another auxiliary hypothesis:

> (SKEP) Our trust heuristics evolved to revoke trust that *p* if we come to consciously consider a specific possibility in which ~*p* that is consistent with our current sensory perceptions. However, our trust heuristics also evolved to reinstate trust that *p* if we subsequently determine (consciously or otherwise) that there actually isn't a significant likelihood that ~*p*, and that the non-epistemic stakes of trusting that *p* actually aren't high.

THYME and SKEP explain why we hesitate to say we *know* that reliable patterns will continue (or that they haven't stopped)—or that our senses or informants are reliable—when we're *consciously considering* specific possibilities in which they won't or aren't. For instance, we may well hesitate to say we *know* that our car is parked where we left it once someone raises the possibility of it being one of the hundreds of cars stolen every day. For our ancestors, losing trust in these sorts of situations was generally advantageous, because there was usually *good reason* to stop trusting that *p*, and to start taking into the account the possibility that ~*p*.

Furthermore, THYME and SKEP explain the return of trust that we typically experience once adverse possibilities are no longer in our mind (that is, assuming that there really wasn't a significant probability of the possibility, and that the non-epistemic stakes weren't actually that high), or perhaps once we've *consciously* assured ourselves that the adverse possibilities are very improbable and that the stakes aren't high. The explanation here is based on the plausible idea that those of our ancestors whose trust heuristics *didn't* reinstate trust in such situations were at a significant disadvantage, due to the accrual of unnecessary contingency plans. By the same token, THYME and SKEP explain the appeal of "contextualist" accounts of knowledge, such as that proposed by Lewis (1996), on which a knowledge ascription "K knows that *p*" can be true in a context where the speaker isn't thinking about adverse possibilities, but false in a context where adverse possibilities are on the table. Contextualist accounts are motivated by the observation that we're perfectly willing to ascribe knowledge when we aren't thinking about (say) the rare times when parked cars are stolen, yet refuse to ascribe knowledge when we *do* focus on the possibility. THYME and SKEP explain this observation, and thereby offer an explanation for why Lewis and others have opted for contextualism.[96]

Do we indeed experience a reinstatement of trust as SKEP proposes? I've certainly had my trust return subconsciously after considering adverse possibilities. For instance, when I was first presented in an epistemology class with the possibility that I might be a "brain in a vat"—hooked up to a supercomputer and being stimulated to have sensory perceptions entirely consistent with being embodied and moving through world—I genuinely felt I couldn't *know* I wasn't envatted.[97] And I've felt the same hesitation nearly every other time I've thought about the possibility since (including right now). But I clearly haven't gone about my life over the past seven years never trusting that my body is real, or that my sensory perceptions are genuine—let alone having contingency plans for suddenly being fed completely different experiences (say, due to a glitch in the putative supercomputer). I'm sure that every other epistemology student initially swayed by the case has had the same experience: Skepticism just doesn't stick around. This point was perhaps put most eloquently by David Hume, who after finding no rational basis for expecting the reliable patterns of nature to continue, concludes with the following:

> Most fortunately it happens, that since reason is incapable of dispelling these clouds, nature herself suffices to that purpose, and cures me of this philosophical melancholy and delirium, either by relaxing this bent of mind, or by some avocation, and lively impression of my senses, which obliterate all these chimeras. I dine, I play a game of back-gammon, I converse, and am merry with my friends; and when after three or four hour's amusement, I wou'd return to these

[96] This isn't to say that contextualists are *wrong* about the truth conditions for knowledge ascriptions. Rather, my point here is that THYME and SKEP can help explain why a *prima facie* strange hypothesis like contextualism can seem attractive to many philosophers. But for all I've said here, contextualism may indeed be the correct theory.

[97] See Putnam (1981) for one of the first philosophical discussions of the "brain in a vat" hypothesis.

> speculations, they appear so cold, and strain'd, and ridiculous, that I cannot find in my heart to enter into them any farther. (Hume 1739/1978, p. 269)

If THYME is correct, then Hume can thank natural selection, at least in part, for "dispelling these clouds".

5.4 Lotteries and Patterns

I observed in the last section that after consciously considering the possibility that some highly probable proposition is in fact false, we almost invariably return to being certain that the proposition is true, and thereby to ascribing knowledge of that proposition to ourselves and others. THYME and SKEP explain this dynamic. Yet there is a notable exception: The odds of us winning the jackpot when we buy a lottery ticket are vanishingly low. Presumably, at the time we buy a ticket, we're consciously thinking about the possibility that it might win. Thus, THYME and SKEP explain why, at that moment, we wouldn't claim to *know* that our ticket will lose. And indeed, it appears from the literature that nearly every philosopher agrees that anyone who buys a losing lottery ticket can't *know* that her ticket will lose—even if she becomes certain that it will—until the winning number is drawn. (See, for instance, DeRose 1996 and Hawthorne 2004.) As Turri and Friedman (2014) show, this intuition appears to be widely shared by non-philosophers as well.[98] But why, once we stop thinking about the matter, don't we become certain that our ticket will lose? Why do we maintain our contingency plans (say, to not lose track of our ticket, and to watch the lottery draw on TV) for the possibility of our ticket winning, given that the possibility is so incredibly unlikely? This has greatly puzzled epistemologists, particularly because many of the propositions we take ourselves to know—say, that our car is

[98] And also by lottery marketers: The slogan of the New York Lottery is "Hey, You Never Know".

still where we parked it an hour ago, or that our invited guests will indeed arrive for dinner next week—surely have a *lower* probability of truth (at least conditional on our relevant evidence) than that a given ticket in a large lottery will lose. (See Hawthorne 2004 for in-depth discussion of this point.)[99]

Amid all the plausible but still speculative claims I've made about our species' evolutionary past, here's an incontrovertible fact: Our early human ancestors did not play the Powerball lottery. In fact, it's almost certain that they never confronted any scenario where the number of conceivable and salient outcomes was much larger than a handful. (By "scenario", I simply mean a branching point—past, present, or future—where, given one's current mental states, various alternative outcomes seem possible.) "Will it rain tomorrow, or will it be sunny?" "Will there be a drought next year, or will there be moderate rainfall—or perhaps flooding?" "Did the gathering group find berries yesterday, or only leaves?" "Will we encounter wildebeest today, or maybe antelopes, or gazelles—or nothing at all to hunt?" Moreover, from a fitness perspective, we can understand why humans might have evolved to generally not *trust* that any particular outcome would or would not occur in scenarios which seemed to be unassociated with any reliable pattern or trend**.** Let's call these "patternless" scenarios. By calling a scenario "patternless" (for a given agent A), I mean that the *type* of scenario in question is not, to A's awareness, associated with any reliable outcome pattern. (As defined, a scenario itself can't have a pattern, since it's a one-time event.) So, for instance, the *exact* amount of next year's

[99] The following observation deepens the puzzle further: Many philosophers (and, I suspect, at least some non-philosophers as well) insist *upon reflection*, and despite an initial feeling to the contrary, that we *do* know that we're not brains in vats, and that we *do* know that our car is still parked where we left it (so long as it actually is, and so long as there's no significant threat of theft). (See, for instance, Putnam 1981: ch. 1 and Hawthorne 2004: ch. 4.) Yet I'm not aware of any philosopher who clearly advocates on reflection that we can know (and not merely be justified in believing) that a given lottery ticket will lose.

rainfall is a patternless scenario, since past years' *exact* rainfall amounts do not seem to follow any pattern.

Plausibly, most features of our ancestors' personal and social worlds followed reliable patterns. For instance, they didn't wake up with a different body each morning, nor did all of their friends suddenly become hostile for no apparent reason. Thus, probably most patternless scenarios faced by our ancestors involved either the workings of nature (within which I include non-human animals) or the behavior of humans outside their own community (and so who weren't necessarily going to be cooperative). If so, then being caught off-guard without a contingency plan for a salient possible outcome of an patternless scenario was generally more likely to result in personal harm (if not death) than being caught without a contingency plan for an outlier outcome of a "patterned" scenario. That is, it generally would have been easier to cope on-the-fly with unexpected events involving only oneself or members of one's community than with those involving forces of nature or hostile outgroup humans. It's plausible, then, that most of the patternless scenarios our ancestors faced were *high-stakes*: Not having a contingency plan for (say) rain tomorrow, or flooding next year, or no gazelles to hunt, would have put our ancestors at a significant disadvantage.

Meanwhile, the number of patternless scenarios which our ancestors were thinking about (consciously or otherwise) at any one time was probably fairly small. The vast majority of salient scenarios likely involved reliable patterns—precisely because most of the things humans think about, and which are relevant to our daily actions, involve either stable aspects of nature (such as the continued flow of rivers, or the daily rising and setting of the sun) or the people in our community (most notably, ourselves). Most of the scenarios our ancestors were considering (consciously or otherwise) at a given time probably involved questions like the following: “Will my hut still be standing when I return tonight?” “Will my regular hunting partner continue to go hunting with me?” “Will my community continue to accept me, as they always have up to now?” Thus, contingency planning for the occurrence *and* non-occurrence of each salient outcome of all salient patternless scenarios, at a given time, probably would not by itself have involved too much cognitive burden. Moreover, it’s likely that all or most salient outcomes in any given patternless scenario our ancestors faced had a significant probability of occurring. (Or, to be more precise, each type of outcome would have occurred in a significant proportion of scenarios of the type in question). This is partly because an outcome generally would have been salient only if it had already been observed at least once in the past. For instance, looking at all the times when our ancestors wondered whether there would be sun or rain the next day, each outcome surely occurred in significant proportions.

Put all of these observations together, and they point to the following auxiliary hypothesis:

> (PATT) Our trust heuristics evolved to never produce trust in the occurrence or non-occurrence of any particular outcome of a patternless scenario—that is, a scenario analogous to scenarios we (or others) have observed before, which to our awareness have not shown any detectable pattern or trend in their outcomes.

Why wouldn’t our trust heuristics have evolved to produce certainty that a particular outcome in a patternless scenarios would not occur once it appeared that that outcome was *very unlikely*? My suggestion here is that in virtually *no* patternless scenarios which our ancestors faced was any given salient possible outcome very unlikely. (Recall that by “unlikely” here, I mean that the *type* of outcome in question occurred never, or almost never, in the long run across all scenarios of the type in question.) Plausibly, possible outcomes became salient to our ancestors either because they had witnessed such outcomes in the past (or heard of others witnessing them), or because they'd become certain in some other way (say, by logical reasoning) that the outcome was both physically possible and, if it occurred, relevant to their goals and plans. And in the patternless scenarios which our ancestors faced—sun or rain? gazelles or antelope?—any outcome which they had witnessed before or had otherwise become certain was possible typically would have occurred in significant proportions in the long run. Thus, we can see why those of our ancestors whose trust heuristics included a heuristic to never trust that any particular outcome in a patternless scenario would or wouldn't occur—regardless of apparent likelihoods in any given case—would have had competitive advantages over those whose trust heuristics tried to carefully track likelihoods in patternless scenarios. The latter would have used up more cognitive energy to track likelihoods, and may well have been unpleasantly surprised—when there was a lot at stake—much more often.

Now what does all this have to do with lotteries? Lotteries are patternless scenarios *par excellence*: A different number is drawn for the jackpot every time (aside from a few possible repeats here or there). There is no reliable pattern to who wins. Thus, THYME and PATT offer an explanation for our refusal to say we know that a lottery ticket will lose: Never mind that the probability of winning a lottery may be astronomically low. (For instance, the odds of winning the jackpot in the Powerball lottery in October 2015 were around 1 in 300,000,000). Once we cognitively register the scenario as patternless, our trust heuristics forbid trust that any particular possible outcome will or won't occur—including the outcome that our ticket (or our friend's ticket, etc.) will lose. According to THYME and PATT, there's a good evolutionary reason for this: As discussed above, it's plausible that in all or nearly all patternless scenarios which our early ancestors faced, each salient outcome had a decent probability of occurring (at least in the long run), and that each outcome was consequential for action. Thus, in our evolutionary past it was advantageous to plan for *all* salient outcomes. THYME and PATT thus assimilate our lottery intuitions to our typical unwillingness to say we know exactly what the weather will be next week (even if we've looked at the current forecasts, and they all give a very low probability of rain).

It's true, of course, that nearly all *serial* lottery players come to observe the following pattern: "I've never won a jackpot". Why wouldn't the trust heuristic of a serial lottery player seize on this pattern to yield trust that the "I haven't won" pattern will continue? (Recall from § 2.4 that we all tend to trust that reliable patterns we've observed in the past will continue.) It's plausible that our ancestors' trust heuristics evolved not to allow "not-me" pattern detection to override awareness of "objective" long-term paternlessness, at least when they weren't aware of any causally-relevant difference in properties between them and other people. Let me explain: Suppose that one of our ancestors lived in a community in which, over the last several weeks, about a third of the members had suddenly fallen ill (all with similar symptoms). There was no detectable pattern in who began displaying symptoms: Young or old, man or woman, skinny or muscular—the symptoms did not appear to discriminate. Suppose that our ancestor in question had not, through all this time, fallen ill. If his trust heuristic then produced trust that the pattern of *him* not being the one who fell ill would continue, he would have made no contingency plans for falling ill, nor would he have hesitated to take actions whose success depended on him not falling ill in the near-future. (For instance, he might not have hesitated to embark on a multi-day hunting trip, thereby leaving behind his community's stores of food and water.) Or consider one of our ancestors who kept narrowly averting injury during hunts, but many of whose hunting partners had become injured over the past.

We can now see why trust in the continuation of “not-me” patterns (or, similarly, “not-her/him” patterns) would not have been advantageous. Thus, we can see why our trust heuristics would have evolved to never produce trust in scenarios which were “objectively” patternless (that is, scenarios which would appear to have no pattern from the perspective of an observer unaffected by the outcomes), even when there *was* a pattern of a particular outcome not occurring for oneself or some other particular person. We can thus reasonably tweak PATT as follows:

> (PATT) Our trust heuristics evolved to never produce trust in the occurrence or non-occurrence of any particular outcome of an *objectively* patternless scenario—that is, a scenario analogous to scenarios we (or others) have observed before, which to our awareness have not shown any detectable pattern or trend in their outcomes (setting aside “not-me” and “not-her/him” patterns).

Consequently, THYME and PATT (as amended) explain why even serial lottery players won’t claim to know that they won’t win the next jackpot, and why we won’t ascribe such knowledge to them either.[100] To wit: Our trust mechanisms have evolved to never produce trust that any particular outcome of a scenario we register as objectively patternless (at least in the long run) will or won’t occur, *regardless* of apparent likelihoods. It was advantageous for our ancestors’ trust heuristics to not respond to apparent likelihoods in objectively patternless scenarios, so long as each outcome was judged to be at least *possible*.

[100] At least, I assume that serial lottery players don’t make claims to such knowledge. If they did, presumably they wouldn’t continue to buy lottery tickets.

Now, THYME and PATT predict that we'll strongly tend to deny knowledge in *all* objectively patternless scenarios we face, not just lotteries—including those in which at least some outcomes are highly unlikely.[101] Is this prediction borne out? I can't exhaustively list all the patternless scenarios modern humans encounter, but here's an example: In scenarios where we visually take in a large number of objects, and can't immediately tell exactly how many there are, there is no pattern to exactly how many objects there end up being (that is, once we count them individually). For instance, when we count up how many quarters are in our piggy bank on various occasions, then unless we have highly systematic spending habits, there's no pattern to how many quarters there are. Hence, THYME and PATT predict that we won't claim to know, without counting, how many of a given type of object there are or aren't in a particular location (at least when we can't tell with a single glance). And indeed, I find myself unwilling to say I know that there aren't exactly 3,476,879,324 grains of sand under any one-block stretch of the Coney Island boardwalk. Likewise, I don't take myself to know that that there aren't currently exactly 286 leaves on the oak tree behind the house where I grew up. Of course, I think it's extremely unlikely that there *are* exactly 3,476,879,324 grains of sand under any one-block stretch of the Coney Island boardwalk, or that there *are* exactly 286 leaves on the oak tree behind the house where I grew up. But I can't bring myself to confidently say I know that there *aren't*. I suspect my reader feels the same inclination—and this, I suggest, is a decent confirmation of the prediction yielded by THYME and PATT. So far as I can tell, then,

[101] From here on, I'll use "patternless" to mean "objectively patternless".

THYME and PATT constitute a plausible explanation for our judgment that knowledge and lotteries don't mix.[102]

[102] I should note that at least some philosophers (myself included) feel a rather strong inclination to say that we *do* know that we won't win the lottery *many times in a row*—say, every drawing for the next thirty years (Hawthorne 2004, p. 20). I suspect that many non-philosophers would likewise claim to know such things. But notice that this is no longer a patternless scenario: We've *never* observed *anyone* winning the lottery even twice in a row, let alone for thirty years running.

CHAPTER VI: Social-Cognitive Epistemology

6.1 Feasibility Study

Over the past four chapters, we've seen that THYME, along with its corollary SAGE and several auxiliary hypotheses, offers *prima facie* plausible explanations for a wide range of potentially puzzling observations about our modern-day uses of the word "know":

- The remarkable prevalence of "know" compared to "believes" and "is certain";
- The universal tendency to deny knowledge when someone is not certain that *p*, or when their belief is false or unjustified;
- The widespread ascription of knowledge to beliefs formed by the fallible processes of perception, memory, and testimony;
- Philosophers' denial of knowledge in both Gettier-type cases (Chapter 4) and several types of cases where the believer seems to have objectively good evidence for her belief (Chapter 3);
- The lower tendency among non-philosophers to deny knowledge in Gettier-type cases;
- The apparent greater hesitation of both philosophers and non-philosophers to ascribe knowledge when the non-epistemic stakes are relatively high;
- The tendency of at least non-philosophers to more readily ascribe knowledge of a negative side-effect of an action than of a positive side-effect;

- The apparent influence of how much time someone has to make a decision on our willingness to ascribe knowledge of propositions bearing on that decision;
- Our tendency to deny knowledge of propositions which we typically take ourselves to know once specific possibilities of falsehood are raised—and our subsequent willingness to ascribe knowledge of those same propositions once the adverse possibilities are no longer salient;
- Our univocal unwillingness to say that we don't know our lottery tickets will lose.

However, we might worry that one or more of the explanations I've proposed ultimately isn't viable—let alone likely to be true—because THYME, SAGE, or one of the auxiliary hypotheses makes unrealistic presuppositions about our cognitive capacities. Here, as a reminder, are THYME and SAGE:

> (THYME) Our words "certainty" and "belief" (in at least some uses) are descendants of words that our early ancestors used to describe the subjective "no doubt" state arising from an evolved subconscious trust heuristic. Our word "know" is the descendant of a word our early ancestors used to urge others to be either certain or uncertain that *p*, and subsequently also to convey that third parties were already certain that *p*. Typically, our ancestors' desire to urge others in this way was due to their own evolved trust heuristics having produced (or not produced) trust that *p*.
>
> (SAGE) Speakers today at least implicitly expect their listeners to interpret "K knows that *p* at t" to convey the prospective advisory message "I would advise anyone in a position similar to K's at t to do the same sort of thing K did in believing that *p*".

Now, THYME supposes that we have a subconscious trust heuristic which takes our current mental states as input, presumably abstracts away from the details of these mental states to yield likelihood-relevant types, and then, if it registers the event that ∼*p* sufficiently unlikely (at least given our goals and plans), not only preempts any subjective feeling of doubt about whether *p*, but also prevents us from taking the possibility that ∼*p* into account when choosing actions, and inhibits the formation of contingency plans for the event that ~*p*. Perhaps even more demandingly, SAGE supposes that when judging whether to ascribe knowledge, we can subconsciously and automatically register whether or not there are plausible interpretations of "I would advise anyone in a similar position to do as K did" which yield a message we don't want to convey. Could our cognition really support something like a trust heuristic? And could we really be capable of subconscious and automatic sensitivity to all the finely-grained factors which, as we saw in Chapters 3 and 4, can affect whether there are intuitively "similar positions" in which we wouldn't want to advise someone to do as K did?

In fact, there's some evidence to think that the answer to both questions is "yes". Psychologists have repeatedly found that humans' affective systems can have systemic effects on thought, emotion, and action with minimal or no conscious-level processing (see, for instance, Schwarz and Clore 2003.) By way of example, the affective state of fear "shapes attention, cognition, motivation, action-readiness and behavior ('fight or flight') in a coordinated way that is relevant to contending with risk or threat of harm" (Railton 2014, p. 144). Moreover, fear can be activated prior to any conscious-level processing of incoming sensory data, and in fact can affect how the conscious mind goes on to interpret those sensory data—for instance, by limiting our conscious focus to just those aspects of our sensory experience which have activated fear (Berridge 2004). It's not so far-fetched, then, to hypothesize a trust heuristic as a component of our affective systems: a neural pathway or network which can "narrow in" on relevant features of our current mental states and, depending on the features it detects, produce systemic changes in our behavior, thought, and feeling (along the lines I've proposed).[103] We already have good evidence that the affective system is capable of just this sort of thing.[104]

Next, how about our putative ability to subconsciously register whether there are natural interpretations of K's "position" such that we wouldn't want to convey, without qualification, "I would advise anyone in a similar position to do just as K did"? As I've argued in Chapters 3 and 4, there are a remarkable number of factors that affect whether there are any such interpretations: the exact nature of K's reason-relevant mental states and how easily K could recall each of them, the existence of reason-relevant facts of which K was unaware, whether or not these facts were involved in causing K's mental states, whether certain changes to these facts would intuitively count as "major" or "minor" changes to K's position, and which of these facts (if any) is "accessible" to K. Could our subconscious mental systems really have near-instantaneous sensitivity to all of these factors at once? In fact, our subconscious cognition appears to be capable of plenty of tasks equally or more sophisticated than what SAGE proposes. Even among non-human mammals, recent neuroscientific research has found evidence of remarkable subconscious tracking of complex variables. Summarizing some of this research, Seligman et al. report:

> Foraging mammals have systems of neurons whose firing rates and sequences correlate with differences in the identity of stimuli, their intensity, the magnitude of specific positive versus negative hedonic rewards or food values, the relative value of a stimulus (e.g., deprivation versus satiation), the absolute value of a stimulus (e.g., physiological need), the probability or expectation of a given outcome, the occurrence of a better- or worse-than-expected predicted error, and the absolute risk and expected value of given actions. (p. 125)

[103] Plausibly, the same trust subsystem of the affective system which produces all-out trust (of the sort I've been discussing throughout) can also issue in *degrees* of trust, such that (for instance) we won't take the possibility that ∼*p* into account unless the stakes appear to be quite high.

[104] In his (2014) and (ms.), Peter Railton begins to flesh out the details of trust as a component of the affective system, citing some compelling psychological and linguistic evidence along the way. I can't do justice to his discussion here, but I point the interested reader to his work.

Among humans, the range of subconscious processing is predictably even more impressive:

> Unlike conscious processes, the unconscious processes underlying implicit prospection [that is, the representation of possible future situations] are capable of handling very large numbers of statistical relationships at once—think, for example, of the decisive feint and pass made at the last minute by a champion soccer player, setting up the winning goal. … [E]ven 8-month-old children appear capable of rapidly learning conditional probabilities implicitly, as they spontaneously parse the speech they hear. (ibid., p. 126)

As the authors note in this second passage, in certain domains our subconscious processes are capable of much *more* sophisticated and technical processing, and at much faster speeds, than our conscious minds. Once we reflect on the incredible complexities which infants subconsciously register about the ways language works, or about motor coordination—let alone the sensitivities of non-human mammals which yield such adaptive behaviors as near-optimal foraging—I don't think it's far-fetched to suppose that we can subconsciously register whether there are plausible interpretations of the message "I would advise anyone in a similar position to do just as K did" which we wouldn't endorse.

At this point, we might worry that the emerging evidence of humans' subconscious sophistication actually tells *against* THYME, which proposes that a subconscious trust *heuristic* developed because our ancestors' brains couldn't handle the complexity of storing contingency plans for the negation of *every* proposition on which their actions relied (see § 2.1). However, we need to keep in mind the difference between *sensitivity* and *storage*. A digital camera, for instance, might have incredible sensitivity to light, thereby producing highly defined images with sharp contrast. But if the camera has only a few megabytes of memory, then it can store only a few dozen of these sophisticated images. Greater storage capacity requires more *hardware*; increasing the sophistication of the software won't make much difference to how much information can be stored. Likewise, it's plausible that our brains have evolved with extremely sophisticated "software", but still face limits on "hardware" (due to physical limits on head size and neuronal density). There is no necessary tension, then, between the optimism of SAGE about our subconscious sensitivities, and THYME's proposed limits on our mental storage capacities.

Clearly, further research by cognitive scientists and neuroscientists would be needed to provide positive evidence for the existence of something like a subconscious trust heuristic, and for the particular subconscious capacities for registering aversion to conveying certain prospective advisory messages. However, the existing evidence not only doesn't rule out the existence of the cognitive capacities THYME and SAGE presuppose, but shows that similar capacities are already in place. There are good grounds for provisionally accepting, then, that the explanations offered by THYME, SAGE, and the other auxiliary hypotheses are physically feasible.

6.2 Accepting THYME

I've now demonstrated that THYME offers physically feasible explanations for a very broad range of existing data about our knowledge ascription practices. But should we actually accept THYME, and its auxiliary hypotheses, as providing *the truth* about how our knowledge ascription practices came to be what they are? Admittedly, the evolutionary story I offered in Chapter 2—about the derivation of our word "know" from a word our ancestors used originally to urge others (with the apparent force of objective reasons) to be certain or uncertain of particular propositions—is wholly speculative, since I have no direct evidence to support it. And yet: We don’t appear to have any *simpler* explanations for *any* of the features of our uses of “know” which I listed in § 6.1. In particular, as I reviewed in Chapter 1, it appears that no existing account of knowledge offered by philosophers to date yields a convincing explanation for *why* we tend to deny knowledge unless K believes that *p*, it is true that *p*, K is justified in believing that *p*, and K is not “Gettiered” (as some epistemologists like to put it). That is, no existing account of knowledge suggests how we humans might have developed a word which follows the contours of “justified, true, non-Gettiered belief”. THYME offers a full explanation for this state of affairs, based on the presumptive desires of our ancestors to authoritatively urge others to be certain (or not) of certain things in particular situations, and also to report efficiently on whether or not others in their community were certain of particular truths.

The only attempt I’m aware of—by a philosopher or anyone else—which I haven’t yet mentioned to explain why at least some of us deny knowledge in Gettier-type cases is Nagel’s (2012) “cognitive-strategy sanctioning” theory. Nagel motivates her account with her empirical finding that only 44% of non-philosopher participants in one study ascribed knowledge to “Wanda”, who glances at a clock at a train station showing 4:15pm, and thereby comes to believe that it’s 4:15pm. In the story, Wanda’s belief is true, but she doesn’t realize that the clock is broken and has been showing 4:15pm for the past two days (p. 174). (By contrast, 86% of participants who received a story in which Wanda forms her belief in the same way, but in which there is no suggestion of the clock being broken, ascribed knowledge to Wanda.) Nagel seeks to explain this result by discussing “cognitive strategies”: Given the situation around us, we implicitly adopt more or less demanding rules for forming judgments on certain matters. When nothing seems “amiss”, we tend to form beliefs about our environment more or less automatically based on our perceptions. But when we have some evidence that our typical sources of evidence are unreliable, we won't trust these sources unless we can gather additional evidence (p. 186).

In the Broken-Clock Wanda case, Nagel suggests that in order to judge whether or not Wanda's belief is knowledge, we begin by (at least implicitly) imagining ourselves in Wanda's situation. When we do so, we see ourselves adopting a fairly demanding cognitive strategy for determining what time it is, given the fact that the clock at the train station is broken. (For instance, we would look for other sources of the time before believing anything about the exact current hour and minute.) But when we recall that Wanda did *not* adopt such a demanding strategy (rather, she simply believed what the broken clock said), we "sanction Wanda for her failure to adopt either our cognitive strategy or the range of evidence we now find intuitively necessary [for certainty], given the strategy we have adopted" (p. 186). Nagel implies that such sanctioning is incompatible with ascribing knowledge: Only believers who don't deserve sanction for not adopting an appropriate cognitive strategy can have knowledge (ibid.). This hypothesis about why we tend to deny that Wanda has knowledge is intriguing, and I admire Nagel for considering how certain features of human psychology might help explain both philosophers' and non-philosophers' intuitions.

However, as Nagel herself points out, neither philosophers nor non-philosophers tend to judge that believers like Wanda in Gettier-type cases are *unjustified*. In fact, Nagel and her collaborators found that their non-philosopher participants on average rated believers in Gettier-type cases as just slightly less justified than believers in "normal" cases (average ratings of 1.91 vs. 1.55, respectively, on a 1-7 scale where 1 was "completely justified" and 7 was "completely unjustified") (pp. 185-86). Turri et al. (2015) found similar results: Participants overwhelmingly judged beliefs as "reasonable" in Gettier-type cases. But if knowledge denials are at least implicitly about "sanctioning" believers for not adopting appropriate cognitive strategies, given their environments, then we would at least *prima facie* expect *low* justification/reasonableness ratings in Gettier-type cases. On this point, Nagel suggests that in justification assessments, "we focus directly on the subject's point of view"; but "when we are asking about knowledge we look instead at the first between the subject's point of view and the subject's environment as seen from our own perspective" (ibid.) So perhaps Nagel can save her sanctioning account after all. But even still, her account doesn't immediately offer an explanation for why there are split verdicts among philosophers on some Gettier-type cases and not on others. Recall, for instance, Goldman's (1976) Fake Barn case:

> Henry is driving in the countryside with his son. … [U]nknown to Henry, the district he has just entered is full of papier-mâché facsimiles of barns. These facsimiles look from the road exactly like barns, but are really just façades, without back walls or interiors, quite incapable of being used as barns. They are so cleverly constructed that travelers invariably mistake them for barns. Having just entered the district, Henry has not encountered any facsimiles; the object he sees [right now] is a genuine barn. (pp. 772-73)

Given the presence of fake barns in the area, Nagel's account seems to predict that we'll sanction Henry for not adopting a more demanding cognitive strategy for determining whether or not what he sees is a barn. Yet a significant proportion of philosophers *do* ascribe knowledge to Henry. Moreover, Colaço et al. (2014) and Turri et al. (2015) found very high knowledge ascription ratings among non-philosophers in structurally similar cases.

As we've seen, THYME offers an explanation for these data, by way of SAGE and SIM: In the Fake Barn case, but not in stopped-clock cases like Nagel's Wanda case, the only alternatives in which the agent might seem to be in a similar position, and in which we wouldn't advise them to “do just as K did”, involve a significant change to the direct causes of K’s relevant mental states. Thus, speakers are on the whole less likely to register an aversion to convey “I would advise anyone in a similar position to do just as K did” when the agent in question is Henry rather than Wanda. Perhaps Nagel can find a way to develop her sanctioning account to explain these differences among Gettier-type cases. But it seems reasonable to expect that any such development will render her account no simpler than THYME. And if Nagel’s account were to compete with THYME, we would still need an explanation for *how* our word “know” came to be sensitive to our implicit desires to sanction (or endorse) others agents’ cognitive strategies.

Thus, in light of the current absence of competing explanations which are both equally satisfactory and simpler, I believe we’re warranted in provisionally accepting THYME and its auxiliary hypotheses as providing roughly the truth about the origins of our knowledge ascription practices. I say “provisionally” because I can’t guarantee that future empirical data won’t decisively contradict one or more of THYME’s predictions, or that no equally satisfactory and simpler hypothesis will be developed in the future. Further, I say “roughly the truth” because of the speculative nature of evolutionary hypotheses: All current evidence is consistent with our word “know” evolving along a path slightly different from the account I offered in Chapter 2. For instance, perhaps our ancestors’ word “gno” developed first for reporting on one’s own and others’ states of certainty, and only later acquired an implicit “advisory” message (“I advise K to be certain that *p*”), rather than (as I suggested) the advisory function arising first. But the upshot would still be the same: We all at least implicitly understand “K knows that *p*” to convey the prospective advisory message “I would advise anyone in a similar position to do just as K did in believing that *p*". And this would be due to the memetic evolution of a word our ancestors developed to convey messages relating to certainty—a state typically caused by an evolved trust heuristic.

Finally, I can't guarantee that there aren't other significant and puzzling features of our knowledge ascription practices which will come to light in the future, or of which I'm not currently aware—particularly as experimental philosophers continue to devise and conduct experiments. However, given the very broad range of features which THYME does explain, I think we have strong inductive grounds for supposing that, with suitable auxiliary hypotheses, THYME suffices to explain all aspects of our uses of "know". That is, I suggest we provisionally accept THYME as a *comprehensive* explanation of our knowledge ascription practices.

6.3 Philosophy and Explanation

If we do accept THYME, at least for now, as providing an accurate explanation for our uses of "know", does this have any bearing on the traditional concerns of epistemologists? I'm not certain that it does. First of all, as I mentioned in Chapter 1, very few epistemologists have expressed interest (at least in writing) about the question of how our word "knowledge" came to be what it is and to work as it does. Second, most epistemologists who study knowledge seem to take the following assumption as their starting point:

> (KNORM) Whether or not K's belief that *p* instantiates the property **knowledge** has normative implications for K.

KNORM typically goes unargued for, but it does seem to pervade most contemporary work in epistemology. For instance, Williamson (2000) argues that any given instantiation of **knowledge** is a mental state, just like believing and seeing and feeling and remembering—but one which requires a proper “fit” between mind and world. He goes on to argue that whether or not our belief that *p* constitutes knowledge has implications for whether we’re justified in believing other propositions (ch. 9) and what it’s appropriate for us to assert (ch. 11). Following Williamson’s lead, Hawthorne and Stanley (2008) argue that whether our belief that *p* constitutes knowledge determines whether we can properly count the proposition that *p* as a reason for any of our actions. Fantl and McGrath (2009: ch. 6) similarly argue for normative relations between knowledge, belief, and action. Several other researchers extend these ideas in Carter et al. (forthcoming).

To be sure, some epistemologists are skeptical about Williamson’s claim that whether our belief that *p* constitutes knowledge has direct normative implications for our other beliefs, let alone for our assertions or actions. However, many of these same epistemologists also identify as “virtue epistemologists”, holding that whether our belief that *p* constitutes knowledge can have *ethical* importance. For instance, Greco (2007) defends the idea that knowing is “a kind of success from ability”, such that we (roughly) *deserve credit* for the truth of our belief when and only when we know. Sosa (2007) develops a similar thesis: Knowledge is true belief that is true *because* it was formed by a reliable cognitive disposition. Thus, virtue epistemologists clearly endorse KNORM as well: Knowledge has or reflects ethical value for the knower.

Now let’s turn back to THYME and SAGE: Do these theories help illuminate the normative implications of whether K’s belief instantiates **knowledge**? We might think so if THYME and SAGE reveal that our knowledge ascription patterns clearly align with some normatively significant property. So let’s consider this possibility. If THYME and SAGE are correct, then the following is a true *ceteris paribus* statement about our knowledge ascription patterns:

> Modern-day speakers are willing to assert "K knows that *p*" just in case they don't register (implicitly or otherwise) an aversion to conveying the message, "I would advise anyone in a similar position to do just as K did in believing that *p*".

We might think this shows that we ascribe knowledge precisely when K is a sort of "role model" for belief. And it does seem to be value in being a role model. Yet as we saw in the Gettier-type cases in Chapter 4, many of us (at least philosophers) sometimes deny that K knows even when K's belief is justified. Is there any sense in which a justified yet "Gettiered" believer fails to be a role model in a way that's normatively significant for her? In particular, is there any sense in which a justified non-Gettiered believer is always "better off" than an equivalent justified Gettiered believer? (Would Russell have accrued more ethical value had the clock he saw reading 7:34 *not* been broken?) I don't currently see any convincing way of answering these questions affirmatively. And several other philosophers have likewise expressed doubt about whether being "non-Gettiered" is at all normatively significant for the believer (e.g. Kvanvig 2003 and Foley 1993: pp. 54-59).

Thus, I don't see much promise for THYME and SAGE helping to substantiate KNORM. Since the standard concern of contemporary epistemology has been to substantiate KNORM, I'm not optimistic about THYME and SAGE contributing to mainstream epistemology. It seems that most epistemologists are concerned with our actual linguistic habits regarding "know" only insofar as these habits provide some evidence for theories about the nature of the property **knowledge**. And, to be sure, many epistemologists doubt that this property corresponds exactly with our concept *knowledge*, let alone our verbal uses of "know". However, I should note here that at least some philosophers *do* regard our linguistic usage as highly relevant to the nature of the property **knowledge**. These philosophers take after Lewis (1983), who suggests that the reference of any word in our language is a function of both our usage patterns and considerations of "naturalness". That is, to determine which property word w picks out, we should first look at the usage patterns of w among competent speakers of the language, in order to identify a set of candidate properties which w might refer to. Within this set, we should use considerations of naturalness—in particular, simplicity and theoretical significance—to determine what the actual referent of w is (p. 372).

If we accept Lewis's account of linguistic reference, does THYME have bearing after all on the search for the truth about the property **knowledge**? THYME does offer a systematic account of our usage patterns of "know". For if THYME and its corollary SAGE are correct, then as stated above, the following is a true *ceteris paribus* statement about our knowledge ascription practices:

> Modern-day speakers are willing to assert "K knows that *p*" just in case they don't register (implicitly or otherwise) an aversion to conveying the message, "I would advise anyone in a similar position to do just as K did in believing that *p*".

Yet as we saw in Chapters 3 and 4, K's "position" is multiply ambiguous between (at least) (1) each subset of K's reason-relevant mental states at t within a given level of recall effort, (2) the entire set of reason-relevant facts about K's situation, and (3) K's relevant mental states plus all relevant facts which are accessible to her at t. Consequently, our usage patterns of "know" are actually quite complex, once we consider not only "everyday" cases but also those cases considered in the epistemology literature (see Chapters 3 and 4). If we take to heart Lewis's equal emphasis on usage and naturalness, then our usage patterns in Gettier-type cases might well have no effect on the referent of “know”, since they only add complexity and account for only a miniscule proportion of our uses of “know”.[105]

If THYME and SAGE don’t help illuminate the nature of the property **knowledge**, might they at least help us determine the truth conditions for knowledge ascriptions? For at least three decades after Gettier’s 1963 paper, the question of the truth conditions for “K knows that *p*” did seem to be the dominant concern of epistemologists. Yet it appears that nearly every epistemologist, at least in the previous decade, subscribes to the following principle:

> (TRUTH) “K knows that *p*” is true just in case K’s belief that *p* instantiates the property **knowledge**.[106]

If we accept TRUTH along with KNORM and/or Lewis’s account of reference, then we likely won’t find THYME or SAGE helpful for discerning the truth conditions of knowledge ascriptions—precisely because they don’t seem to be helpful for illuminating the nature of **knowledge**.[107]

Perhaps other epistemologists will find ways in which THYME and SAGE can contribute to the traditional concerns of the discipline. But if my hypotheses in the end don’t bear on these traditional concerns, are they still contributions to philosophy? Would they not simply be speculative—albeit plausible—theories about our evolutionary biology and cognitive

[105] In fact, Weatherson (2003) argues precisely along these lines: He suggests that on Lewis’s account of reference, we should probably accept that “knowledge” refers simply to the property “justified true belief”.

[106] Or, for contextualists: “K knows that *p*” is true in context c just in case K’s belief that *p* instantiates the property **knowledge$_c$** (that is, the determinate, objectively significant property to which the word “knowledge” refers in context c).

[107] Do THYME and SAGE give epistemologists some reason to rethink their confidence in KNORM or TRUTH? We might wonder whether we really have good evidence for either claim. And there might be convincing THYME-based explanations for why most epistemologists find these principles appealing—explanations which don’t require the truth of either principle. I think it’s worth asking whether to retain our confidence in these principles (and not, perhaps, be agnostic about them), but I don’t have space here to begin defending an answer.

psychology? I have two responses to this concern. First, THYME and SAGE are theories about our *social*-cognitive psychology. Albert Bandura, the founder of social-cognitive psychology, defines a social-cognitive theory as one which "explains psychosocial functioning in terms of triadic reciprocal causation. In this causal model, behavior, cognitive and other personal factors, and environmental events [including social context] all operate as interacting determinants that influence each other bidirectionally" (Bandura 1988, p. 276). This is precisely the type of explanation which THYME and SAGE propose for our knowledge ascription behaviors: Our ascriptions and denials of knowledge are joint products of our cognitive processing of the believer's situation and our implicit concerns about how our assertions will be perceived socially—in particular, whether our listeners or audience will perceive us to endorse certain messages that we don't mean to endorse. I emphasize this point because the crucial element in THYME's and SAGE's explanations is that our word "know" originated in a word used to *give advice to others*. This insight might be helpful for future investigations of other linguistic practices which, like our "know"-related practices, don't *at first sight* have anything to do with our social context.[108]

So okay, are THYME and SAGE contributions to evolutionary biology and *social*-cognitive psychology, but not contributions to philosophy? In fact, I beg to differ: These theories offer compelling explanations for a range of puzzling features of our lives. Admittedly these puzzles aren't typically salient to those outside the discipline of epistemology. But in the absence of an

[108] It seems possible to me that our uses of such words as "cause", "(un)just", "(un)fair", and "(morally) right/wrong" have their origins in the desires of our ancestors to convey messages of the form "I advise acting in such-and-such ways in thus-and-such future similar situations, at least if your/her/our goal is thus-and-thus". Future work might explore these possibilities.

explanation, it *is* puzzling why many of us won’t ascribe knowledge to the protagonists in Gettier’s cases and similar ones. (“What more could knowledge ask for?”, we might ask.) And once we get to thinking about it, why do we require justification, or certainty, or truth before we agree that someone knows something? Why does the word “knowledge” get used so much that it’s one of the most common words in the English language? Why not just have words for “belief” and “true belief”, and leave it at that?

By offering resolutions to these admittedly small but still intriguing puzzles, THYME and SAGE help us make some small amount of progress in the tremendous task of understanding and examining our lives. Moreover, we often *value* understanding how we got to be the way we are—not just because this sometimes helps us achieve other goals, but also because it can be intrinsically satisfying. I do think that the understanding we gain from THYME and SAGE is intrinsically valuable, whatever else might come of it. At least speaking for myself, I find it significant to know why “know” has all the properties it does, even though of course there are more significant things that we can and do know. THYME and SAGE then arguably *do* contribute to philosophy. After all, we philosophers have traditionally dedicated our work to seeking out the things in our lives that have intrinsic value, and in particular to seeking out understanding for its own sake. Let’s agree, then, that THYME and SAGE are pieces of “social-cognitive epistemology”. They just might add some new flavor to the study of knowledge.

BIBLIOGRAPHY

Alston, W. P. (1988). The deontological conception of epistemic justification. *Philosophical Perspectives* 2(Epistemology): 257-299.

— (2002). Plantinga, naturalism, and defeat. In J. Beilby (ed.), *Naturalism Defeated? Essays on Plantinga's Evolutionary Argument against Naturalism*. Ithaca, NY: Cornell University Press.

Ballantyne, N. (2011). Anti-luck epistemology, pragmatic encroachment, and true belief. *Canadian Journal of Philosophy* 41(4): 485-504.

Bandura, A. (1988). Organisational applications of social cognitive theory. *Australian Journal of Management* 13(2): 275-302.

Beebe, J. R. & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind and Language* 25(4): 474-498.

Beebe, J. R. & Jensen, M. (2012). Some surprising connections between knowledge and action: the robustness of the epistemic side-effect effect. *Philosophical Psychology* 25(5): 689-715.

Berridge, K. C. (2004). Motivation concepts in behavioral neuroscience. *Physiology and Behavior* 81: 179-209.

Buckwalter, W. (2011). Gettier made ESEE. Unpublished manuscript. City University of New York Graduate Center.

— (2012). Non-traditional factors in judgments about knowledge. *Philosophy Compass* 7(4): 278-289.

Carter, J. A., Gordon, E. C., & Jarvis, B. W. (eds.). (forthcoming). *Knowledge-First: Approaches in Epistemology and Mind*. Oxford: Oxford University Press.

Chisholm, R. M. (1942). The problem of the speckled hen. *Mind* 51(204): 368-373.

— (1966). *Theory of Knowledge*. Englewood Cliffs, NJ: Prentice Hall.

Christensen, D. (2007). Epistemology of disagreement: the good news. *Philosophical Review* 116(2): 187-217.

Colaço, D., Buckwalter, W., Stitch, S., & Machery, E. (2014). Epistemic intuitions in fake-barn thought experiments. *Episteme* 11(2): 199-212.

Collier, K. W. (1973). Contra the causal theory of knowing. *Philosophical Studies* 24(5): 350-352.

Craig, E. (1991). *Knowledge and the State of Nature: An Essay in Conceptual Synthesis*. Oxford: Oxford University Press.

Davies, M. & Gardner, D. (2010). *A Frequency Dictionary of Contemporary American English*. London: Routledge.

Dawkins, R. (1989). *The Selfish Gene*. 2nd ed. Oxford: Oxford University Press.

DeRose, K. (1992). Contextualism and knowledge attributions. *Philosophy and Phenomenological Research* 52(4): 913-929.

— (1996). Knowledge, assertion and lotteries. *Australasian Journal of Philosophy* 74(4): 568-580.

Elga, A. (2007). Reflection and disagreement. *Noûs* 41(3): 478-502.

Fantl, J. & McGrath, M. (2009). *Knowledge in an Uncertain World*. Oxford: Oxford University Press.

Feit, N. & Cullison, A. (2011). When does falsehood preclude knowledge? *Pacific Philosophical Quarterly* 92: 283-304.

Feldman, R. (2003). *Epistemology*. New Jersey: Prentice Hall.

Feltz, A. & Zarpentine, C. (2010). Do you know more when it matters less? *Philosophical Psychology* 23(5): 683-706.

Foley, R. (1993). *Working Without a Net: A Study of Egocentric Epistemology*. New York: Oxford University Press.

Gendler, T. S. & Hawthorne, J. (2005). The real guide to fake barns: a catalogue of gifts for your epistemic enemies. *Philosophical Studies* 124: 331-352.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis* 23(6): 121-123.

Gibbard, A. (2003). *Thinking How to Live*. Cambridge, MA: Harvard University Press.

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable. *Journal of Personality and Social Psychology* 59(4): 601-613.

Goldman, A. I. (1967). A causal theory of knowing. *The Journal of Philosophy* 64(12): 357-372.

— (1976). Discrimination and perceptual knowledge. *The Journal of Philosophy* 73(20): 771-791.

— (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

Gopnik, A. & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: Bradford, MIT Press.

Greco, J. (2007). The nature of ability and the purpose of knowledge. *Philosophical Issues* 17(1): 57-69.

Harman, G. (1973). *Thought*. Princeton, NJ: Princeton University Press.

Hawthorne, J. (2004). *Knowledge and Lotteries*. Oxford: Oxford University Press.

Hawthorne, J. & Lasonen-Aarnio, M. (2009). Knowledge and objective chance. In P. Greenough & D. Pritchard (eds.), *Williamson on Knowledge*, Oxford: Oxford University Press.

Hawthorne, J. & Stanley, J. (2008). Knowledge and action. *The Journal of Philosophy* 105(10): 571-590.

Horowitz, S. (2013). Epistemic akrasia. *Noûs* 48(4): 718-744.

Hume, D. (1739/1978). *A Treatise of Human Nature*. Ed. L. A. Selby-Bigge. 2nd ed. Oxford: Oxford University Press.

Kelly, T. (2005). The epistemic significance of disagreement. In J. Hawthorne & T. Gendler (eds.), *Oxford Studies in Epistemology, Volume I*, Oxford: Oxford University Press.

Kornblith, H. (2002). *Knowledge and its Place in Nature*. Oxford: Oxford University Press.

— (2007). Naturalism and intuitions. *Grazer Philosophische Studien* 74: 27-49.

Kvanvig, J. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge: Cambridge University Press.

Lackey, J. (2009). Knowledge and credit. *Philosophical Studies* 142(1): 27-42.

Lasonen-Aarnio, M. (2010). Unreasonable knowledge. *Philosophical Perspectives* 24(1): 1-21.

— (2013). Disagreement and evidential attenuation. *Noûs* 47(4): 767-794.

— (2014). Higher-order evidence and the limits of defeat. *Philosophy and Phenomenological Research* 88(2): 314-345.

Lehrer, K. (1971). How reasons give us knowledge, or the case of the gypsy lawyer. *The Journal of Philosophy* 68(10): 311-313.

Leplin, J. (2009). *A Theory of Epistemic Justification*. Springer.

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: a typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes* 76(2): 149-188.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs* 13(4): 455-476.

— (1983). New work for a theory of universals. *Australasian Journal of Philosophy* 61(4): 343-377.

— (1996). Elusive knowledge. *Australasian Journal of Philosophy* 74(4): 549-567.

Loftus, E. F. (1974). Reconstructing memory: the incredible eyewitness. *Jurimetrics Journal* 15: 188-193.

Lycan, W. G. (1977). Evidence one does not possess. *Australasian Journal of Philosophy* 55(2): 114-126.

— (2006). On the Gettier problem problem. In Hetherington, S. (ed.), *Epistemology Futures*, Oxford: Clarendon Press.

Makinson, D. C. (1965). The paradox of the preface. *Analysis* 25: 205-207.

May, J., Sinnott-Armstrong, W., Hull, J. G., & Zimmerman, A. (2010). Practical interests, relevant alternatives, and knowledge attributions: an empirical study. *Review of Philosophy and Psychology* 1(2): 265-273.

Montgomery, D. E. (1992). Young children's theory of knowing: the development of a folk epistemology. *Developmental Review* 12: 410-430.

Myers-Schulz, B. & Schwitzgebel, E. (2013). Knowing that p without believing that p. *Noûs* 47(2): 371-384.

Nagel, J. (2010). Knowledge ascriptions and the psychological consequences of thinking about error. *Philosophical Quarterly* 60(239): 286-306.

— (2011). The psychological basis of the Harman-Vogel paradox. *Philosophers' Imprint* 11(5): 1-28.

— (2012). Mindreading in Gettier cases and skeptical pressure cases. In J. Brown & M. Gerken (eds.), *Knowledge Ascriptions*, Oxford: Oxford University Press.

Penrod, S. & Cutler, B. (1995). Witness confidence and witness accuracy: assessing their forensic relation. *Psychology, Public Policy, and Law* 1(4): 817-845.

Pinillos, A. (2011). Some recent work in experimental epistemology. *Philosophy Compass* 6(10): 675-688.

— (2012). Knowledge, experiments, and practical interests. In J. Brown & M. Gerken (eds.), *Knowledge Ascriptions*, Oxford: Oxford University Press.

Pollock, J. L. (1986). *Contemporary Theories of Knowledge*. London: Hutchinson.

Pritchard, D. (2005). *Epistemic Luck*. Oxford: Clarendon Press.

— (2010). Knowledge and understanding. In D. Pritchard, A. Millar, & A. Haddock (eds.), *The Nature and Value of Knowledge: Three Investigations*, Oxford: Oxford University Press.

Putnam, H. (1981). *Reason, Truth, and History*. Cambridge: Cambridge University Press.

Railton, P. (2014). Reliance, trust, and belief. *Inquiry* 57(1): 122-150.

— (ms.). The value of truth and the value of belief.

Russell, B. (1948). *Human Knowledge: Its Scope and Limits*. New York: Simon and Schuster.

Schafer, K. (2014). Doxastic planning and epistemic internalism. *Synthese* 191(12): 2571-2591.

Schaffer, J. (2006). The irrelevance of the subject: against subject-sensitive invariantism. *Philosophical Studies* 127: 87-107.

Schauer, F. (2009). *Thinking Like a Lawyer: A New Introduction to Legal Reasoning*. Cambridge, MA: Harvard University Press.

Schwarz, N. & Clore, G. L. (2003). Mood as information: 20 years later. *Psychological Inquiry* 14(3-4): 296-303.

Seligman, M. E. P., Railton, P., Baumeister, R. F., & Sripada, C. (2013). Navigating into the future or driven by the past. *Perspectives on Psychological Science* 8(2): 119-141.

Shin, J. (2014). Time constraints and pragmatic encroachment on knowledge. *Episteme* 11(2): 157-180.

Smith, M. (2010). What else justification could be. *Noûs* 44(1): 10-31.

Sosa, E. (1999). How to defeat opposition to Moore. *Philosophical Perspectives* 13(Epistemology): 141-153.

— (2007). *Virtue Epistemology: Apt Belief and Reflective Knowledge, Volume I*. Oxford: Oxford University Press.

Stanley, J. (2005). *Knowledge and Practical Interests*. Oxford: Oxford University Press.

— (2008). Knowledge and certainty. *Philosophical Issues* 18(1): 35-57.

Starmans, C. & Friedman, O. (2012). The folk conception of knowledge. *Cognition* 124(3): 272-283.

Sudduth, M. (2008). Defeaters in epistemology. *Internet Encyclopedia of Philosophy*. http://www.iep.utm.edu/ep-defea. Accessed April 22, 2016.

Turri, J. (2010). On the relationship between doxastic and propositional justification. *Philosophy and Phenomenological Research* 80(2): 312-326.

— (2013). A conspicuous art: putting Gettier to the test. *Philosophers' Imprint* 13(10): 1-16.

— (2014). The problem of ESEE knowledge. *Ergo* 1(4): 101-127.

Turri, J., Buckwalter, B., & Blouw, P. (2015). Knowledge and luck. *Psychonomic Bulletin and Review* 22: 378-390.

Turri, J. & Friedman, O. (2014). Winners and losers in the folk epistemology of lotteries. In Beebe, J. R. (ed.), *Advances in Experimental Epistemology*, London: Bloomsbury Academic.

Unger, P. K. (1975). *Ignorance: A Case for Scepticism*. Oxford: Oxford University Press.

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

— (2011). Defending interest-relative invariantism. *Logos and Episteme* 2(4): 591-609.

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics* (29): 429–460.

Williamson (2000). *Knowledge and its Limits*. Oxford: Oxford University Press.

— (2009). Reply to John Hawthorne and Maria Lasonen-Aarnio. In P. Greenough & D. Pritchard (eds.), *Williamson on Knowledge*, Oxford: Oxford University Press.