Chapter 2: Varieties of Scrutability

David J. Chalmers

1 The Scrutability Schema

In this chapter I will formulate and discuss a number of specific scrutability theses. Scrutability theses can vary along a number of dimensions. They can vary in the scope of the class of truths that are held to be scrutable, and indeed in just what sort of thing truths are taken to be. They can vary in the specificity of the class of truths in the scrutability base. And they can vary in their understanding of the scrutability relation: that is, in their understanding of what it is for a truth to be scrutable from a class of truths.

In this chapter, I will focus especially on different ways of understanding the scrutability relation. The canonical understanding of this relation construes it in terms of a priori entailment, but there are also many other ways to understand it. These other ways are useful as a way of motivating the idea of scrutability for those who are skeptical of claims about apriority, and they can also be used (as in the next chapter) as part of an argument for the a priori scrutability thesis. The notions of scrutability are also worthy of attention in their own right.

First, however, I will briefly focus on other aspects of scrutability theses. For a start, scrutability theses can vary in their scope. Scrutability theses can be *restricted*, holding that truths in one specific class (say, the class of mental truths) are scrutable from truths in another class (say, the class of physical truths). They can be *unrestricted but specific*, holding that all truths are scrutable from truths in a specific class (say, the class of physical truths). And they can be *unrestricted and unspecific*, holding that there is *some* class of truths, satisfying a certain constraint (say, a compact class of truths), such that all truths are scrutable from truths in that class.

In what follows, my focus will be especially on unrestricted and unspecific scrutability theses, though I will sometimes focus on restricted and/or specific theses. In particular, I will focus on unrestricted and unspecific theses of the following sort, which I will call the Scrutability Schema.

Scrutability Schema: There is a compact class of truths such that all truths are scrutable

from truths in that class.

This still leaves three things to be clarified: "truths", "compact", and "scrutable from". Each of these yields a further potential degree of variation. I have discussed "compact" already in the previous chapter. Here I will briefly discuss some readings of "scrutable from", before discussing "truths" and these readings of "scrutable from" in much more detail in the remainder of the chapter.

We might say that a truth S is scrutable from a class of truths C if the truths in C *epistemically necessitate S*, in some sense of "epistemically necessitate". This requires a reasonably strong epistemological connection between the truths in C and S. This connection can be understood in many different ways, though.

We have already seen *A Priori*, *Analytic*, and *Definitional* Scrutability theses, according to which scrutability is understood as a priori entailment, analytic entailment, and definitional entailment. Strictly speaking the latter two sorts of entailment are not themselves epistemological notions, but analytic and definitional truths are commonly held to be epistemologically accessible in some strong way, in which case analytic and definitional entailment have strong epistemological consequences. In any case, I will set aside analytic and definitional scrutability in what follows.

The most important notions of scrutability for my purposes are a priori scrutability, along with two further notions of scrutability do not require the controversial notions of analyticity, apriority, or definition. These are what we might call:

> *Empirical Scrutability*: S is empirically scrutable from C if, were one to come to know the truths in C, one would be in a position to know S.

> *Conditional Scrutability*: S is conditionally scrutable from C if one is in a position to know that *if* the members of C are true, then S is true.

Before proceeding, it is worth saying something about what it is to know a sentence, or (equivalently) to know that a sentence is true. For present purposes, these notions should not be understood as invoking metalinguistic knowledge. To know 'All bachelors are unmarried' is not to know something about the sentence. Rather, it is to know something about bachelors. This is one place where things go more straightforwardly if scrutability is formulated in terms of propositions, whereas the formulation in terms of sentences requires some extra work.

2 Sentences or Propositions?

Many different sorts of things are sometimes said to be truths: true propositions, true beliefs, true sentences, true utterances. Which of these is most relevant to the scrutability thesis?

There is much to be said for understanding truths as propositions, abstract entities that are independent of any particular language, and that are the things that we assert and that we believe. My own view is that if propositions are understood correctly, then a scrutability thesis concerning propositions is true. However, the nature of propositions is strongly contested, and different theories of propositions will have quite different results for scrutability.

On a *possible-worlds* view of propositions, the proposition expressed by a sentence is the set of possible worlds where the sentence is true. On this view, all necessary truths express the same proposition (the set of all worlds), a proposition that is itself knowable a priori. If so, then if it is necessary that water is H_2O, it follows that the proposition that water is H_2O is itself knowable a priori. On a *Russellian* view of propositions, the proposition expressed by a sentence is a structure involving those objects and properties that are the extensions of parts of the sentence. On this view, not all necessary truths express the same proposition, but the proposition that Hesperus is a planet and the proposition that Phosphorus is a planet are identical. On a *Fregean* view of propositions, the proposition expressed by a sentence is a structure of senses expressed by parts of a sentence, where senses are fine-grained entities reflecting the epistemic and cognitive significance of various expressions. On this view, not all necessary truths express the same proposition, and the proposition that Hesperus is a planet and the proposition that Phsphorus is a planet are distinct.

It is clear that if we antecedently assume one of these views of propositions, there will be very different results for a scrutability thesis cast in terms of propositions. On the possible-worlds view, all necessary truths will automatically be a priori scrutable from any basis. On the Russellian view, necessary truths will not usually be a priori scrutable from arbitrary bases. But some, such as the proposition that Hesperus is Phosphorus (if they exist), will arguably be scrutable from any basis, and there will be no epistemological differences between propositions expressed by pairs of sentences involving 'Hesperus' and 'Phosphorus' respectively. On a Fregean view, necessary truths will not automatically be scrutable, and epistemological differences between sentences involving 'Hesperus' and 'Phosphorus' will be preserved.

My own purposes include the analysis of fine-grained epistemological differences, and for this purpose, a Fregean view of propositions is the most promising. But I cannot simply assume such a view at the outset: it is controversial whether there are Fregean propositions, and even among

sympathizers, it is controversial just what sort of thing they might be and how they behave. Further, one of my purposes is to use the scrutability thesis to develop a Fregean view of propositions. If I were to assume such a view at the outset, there would be some circularity here. Perhaps some support for the view would accrue from demonstrating coherence and power in the resulting picture, but a flavor of assuming the conclusion and of begging the question against opponents would remain.

At the same time, it will not do to assume one of the other views of propositions. And it does not make sense to cast things in terms of propositions but stay neutral between these views, as the views yield very different results in evaluating scrutability theses. One can certainly cast scrutability theses in terms of all three notions of propositions, but a scrutability base for all three might look very different.

So I will set aside propositions for now. Still, if one accepts Fregean propositions, or if one is at least prepared to allow that propositions are sufficiently fine-grained to reflect differences in cognitive significance as well as differences in reference, then it will not hurt to conceive of the scrutability thesis in terms of propositions on occasion.

It is not out of the question to cast a scrutability thesis in terms of mental states such as thoughts or beliefs, holding for example that for any belief one might entertain, the truth or falsity of the thought could be inferred from a certain class of (potential) basic beliefs. But the individuation of thoughts and beliefs is also nontrivial, and it is more awkward to speak of mental items than of linguistic items, so I will set these aside too.

Instead, I will take the truths at issue to be linguistic items such as *sentences*. Here one needs to give a slightly different treatment for sentences in the base class and the dependent class. Sentences in the dependent class must include sentences of natural language, as these are part of what we are trying to analyze. But sentences in the base class need have no such constraint: the project of the *Aufbau* does not require that there is a natural-language expression for recollected similarity, for example. So we can allow that the sentences in the base class be sentences in a natural language or a merely possible language. The sentences in the base class will be sentence types rather than sentence tokens, as these sentences may never be uttered in the actual world. The base vocabulary will plausibly need a small number of indexicals, such as 'I' and 'now', but beyond this it can exclude context-dependent expressions, such as (arguably) 'tall' or 'ready'.

The sentences in the dependent class must include sentences of natural language. For reasons tied to context-dependence, however, these cannot be restricted to sentence *types*: there may be no context-independent fact of the matter about whether 'John is tall' is true, for example. To

handle these cases (and for other reasons outlined below), the dependent class needs to include sentence tokens, uttered by specific speakers on specific occasions.[1] One might also allow it to include sentence types, as long as the sentences lack a certain sort of context-dependence. But for simplicity, I will initially assume that the dependent class includes only sentence tokens: that is, the class of sentence tokens that are produced sometime in the history of the world.

This restriction has the side effect of weakening the scrutability thesis somewhat, as it may be that various problematic sentences are never in fact uttered in the actual world. One can avoid this weakening by requiring that the scrutability thesis holds not just in our world, but in all nearby worlds, where nearby worlds include worlds in which arbitrary possible sentences of existing languages are uttered. That is, we can require that in all such worlds, there is a compact class of truths (sentences true in that world) from which all truths (tokens produced in that world) are scrutable. In fact, it probably does not hurt to require that the scrutability thesis be nomologically necessary, so that it holds in all worlds with the same fundamental properties and laws as ours. One could even require that the thesis be metaphysically necessary, true in all possible worlds including worlds with entirely alien properties and laws. I think that some such claim is plausible, but I do not intend anything this strong to be built into the scrutability thesis as understood here. In the first instance, I will concentrate on nomologically possible worlds, and later I will consider stronger theses.

A final issue is that of what it is for a sentence to be true at a world. For sentence tokens and context-independent sentence types one can adopt the usual understanding, as worries about contextual variability do not apply here (I set aside issues about relativism for now). A residual issue concerns base sentences involving indexicals such as 'I' and 'now'. These sentences are not true and false absolutely (in the actual world): 'I am hungry now' can be true relative to one subject

---

[1]Another alternative is to allow the dependent class to include sentences in context. However, for my purposes this would require a quite unusual sort of context-dependence that is not well-understood within this framework. For example, it happens sometimes that the a priori inferential role of a name vary between two speakers, although extension of the the name does not vary. It is not clear how this should be represented in a sentences-in-context framework. One could allow something like Fregean sense as a context-dependent semantic value, and the speaker's psychology as a feature of context, but this would require highly nonstandard technical apparatus from the start. Or one could retain extension as the relevant semantic value, while relying on contextual variability in the truth-value of sentences such as 'It is a priori that if D then S'. But here again it is not clear how the relevant context-dependence should be represented. The treatment in terms of sentence tokens has the advantage of proceeding in a somewhat more theory-neutral way. Once we have used this method to motivate a theory involving Fregean senses and the like, one might then revisit the question of how best to construct a theory of their contextual variability. Thanks to Jonathan Schaffer for discussion here.

at one time, false relative to another subject at another time. For this reason, we cannot require that base sentences be true absolutely. Rather, they must be true or false relative to a subject at a time. It follows that the relevant class of base truths will be different for different subjects or different times. On a moment's reflection, this is just what one should expect, if the base truths include a specification of the subject's place in the world.

The scrutability thesis must then make reference to subjects and times, saying for example that for all subjects S at all times t, there is a compact class of truths (true relative to S and t) such that all truths (relative to S and t) are scrutable from that class. If the dependent truths are restricted to sentence tokens, as above, however, this formulation threatens to trivialize the thesis. It may be that every true sentence token is scrutable from a *single* truth (namely itself, or its counterpart in the base language). If so, then as long as subjects can produce only a finite number of tokens at one time, there will be a finite class of truths from which the true tokens are scrutable.

In light of all this, it is better to formulate the thesis by saying that there is a compact class of *sentences* such that for all subjects S and all times t, all truths (relative to S and t) are scrutable from true sentences (relative to that subject and time) in that class. If dependent truths are restricted to sentence tokens, the thesis can be put more simply by saying that there is a compact class of sentences such that every true sentence token is scrutable from the true sentences (relative to the subject and time of utterance) in that class. When these details are relevant, this is how I will understand the scrutability thesis in what follows.

3   Sentences and Thoughts

Casting things in terms of sentences rather than propositions can be awkward and unfamiliar. So it is only worth doing if there are significant advantages. The major advantage is with respect to the dialectical situation discussed above: sentences provide a relatively neutral starting point, where propositions do not. In particular, sentences allow us to argue in a relatively neutral way for scrutability theses that might eventually support a Fregean approach to content.

Here an especially relevant opponent is a Russellian about content who holds that the proposition that Hesperus is Hesperus and the proposition that Hesperus is Phosphorus are identical, so that there can be no psychological or epistemological differences between them. But the *sentences* 'Hesperus is Hesperus' and 'Hesperus is Phosphorus' are not identical, and likewise tokens of these sentences are quite distinct. Furthermore, the Russellian cannot plausibly deny that there are any psychological or epistemological differences between typical *tokens* of the sentence 'Hesperus

is Hesperus' and typical tokens of the sentence 'Hesperus is Phosphorus'.

Suppose that Sue is astronomically ignorant. She does not believe that the morning star is the evening star, but she is aware that her interlocutor thinks this, and she wants to please him. She says 'Hesperus is Hesperus', sincerely, and then she says 'Hesperus is Phosphorus', insincerely. Call the first sentence token S_1, and the second S_2. Then there is a clear sense in which S_1 expresses *belief*, while S_2 does not. There is also a clear sense in which S_1 expresses *knowledge*, while the second does not. We can mark this distinction by saying that Sue *believes* S_1 but not S_2, and that she *knows* S_1 but not S_2. It is somewhat nonstandard to talk about knowing sentences rather than propositions, but the intuitive distinction here is clear enough.[2]

Of course if the Russellian about propositions is correct, then knowing S cannot be identified with knowing the proposition expressed by S. According to the Russellian, S_1 and S_2 express the same proposition, and Sue both knows and believes it, so that Sue knows and believes the proposition expressed by S_2. But the intuitive distinction between the utterances remains undeniable. So all that follows is that a Russellian should not analyze this distinction wholly in terms of propositions expressed by an utterance.

There are various alternative ways that a Russellian might analyze this notion. Russellians (such as Salmon 19xx) who allow that propositions are presented under guises might say that the speaker knows a sentence token S when they know the proposition expressed by S under the guise associated with S. Others might say that the speaker knows S when they know some ancillary proposition that is not semantically expressed by S but is otherwise associated with S. It could also be suggested that to know S is just to know that this S is true, but I do not think that this is implausible: one could know S in the intuitive sense above even if one has no beliefs *about* any sentence tokens, but merely certain beliefs about planets.

However, there is a natural way to draw the distinction which is available to theorists of many different stripes. To do this, we rely on connections between utterances and *mental states* of the speaker. Intuitively, when Sue utters S_1, her utterance is associated with a mental state—a specific act of entertaining the proposition that Hesperus is Hesperus—that is itself a state of belief and a state of knowledge. When Sue utters S_2, on the other hand her utterance is associated with a very different sort of mental state—a specific act of entertaining the proposition that Hesperus is Phosphorus— that is neither a state of belief nor a state of knowledge.

To make this more precise, let us say that a *thought* is a token propositional attitude with a

---

[2]One way to get at the point: insofar as knowledge and belief are norms on assertion, it seems clear that S_2 violates these norms.

mind-to-world direction of fit. So beliefs, suppositions, and acts of knowledge are all thoughts. When one believes that P, knows that P, expects that P, hypothesizes that P, supposes that P, and so on, one thinks that P. Some thoughts constitute beliefs. Some thoughts constitute knowledge (as with my current thought that 2+2=4). Other thoughts constitute potential knowledge, in that it is possible that on (perhaps idealized) reflection they can come to be justified, such that they will then constitute knowledge.

We can then appeal to the idea that utterances of truth-apt sentences typically *express* thoughts. Sincere utterances typically express beliefs, and even insincere utterances typically express thoughts of some different kind. Note that expression is here construed as a primitive relation between utterances and mental states, and should not be confused with the different notion of expression construed as a relation between utterances and propositions. However, there is a close relationship between the notion. Intuitively, an utterance and the thought it expresses have the same propositional content. I will not build this in as a definitional constraint, as there may be views on which thoughts and utterances have contents of quite different sorts. But it is at least a constraint that when an utterance expresses a thought, the utterance and the thought have the same *truth-value*, and it is natural to hold that they must have the same *truth-conditions* as well. These are not the only constraints on the notion: there must also be a causal link between the thought and the utterance, and an appropriate psychological relation. I will not try to define these things here, and will take the notion as an intuitive primitive instead.

It may be that there are some utterances that do not express thoughts. One might absent-mindedly utter a sentence by rote, without entertaining its content at all. If one is grasping for words, one might use a word whose content is not that of a thought. But at least for typical utterances, it is plausible that they express thoughts.

We can then say that a subject who utters a token S *believes* S iff S expresses a thought that constitutes belief. The subject *knows* S if S expresses a thought that constitutes knowledge. The subject knows S *a priori* if S expresses a thought can constitutes a priori knowledge. Then in the case above, it is plausible that Sue knows S_1 but not S_2: the thought expressed by S_1 constitutes knowledge (and belief, and a priori knowledge), but the thought expressed by S_2 does not.

This picture is available to many different theorists of propositions. As with sentence tokens, the claim that there are the relevant psychological or epistemological differences between the thoughts involved in S_1 and S_2 is hard to deny. At best, a Russellian may deny that for a thought to consitute knowledge is for the subject to know the proposition that is the content of the thought: in cases such as the above, the subject may have two thoughts with the same content, one of which

constitutes knowledge and the other one does not. The same goes for belief.

Here it might help to briefly adopt a common model according to which thoughts correspond to sentence tokens in the language of thought. Each token has some content. When such a sentence token is in the "belief box", it corresponds to a belief in the content. When the token is in the desire box, it corresponds to a desire in the content. When a token in the belief box was brought about by the right sort of process, it constitutes a justified belief. When further conditions are met, it constitutes knowledge. In a case such as the above, Sue's utterance S_1 might be triggered by the sentence 'H = H' in her language of thought, while S_2 might be triggered by the sentence 'H = P' in her language of thought. The first sentence will be in the belief box will meet the relevant further conditions, and will constitute knowledge. The second sentence will not be in the belief box, so it will not constitute belief or knowledge.

Of course the picture involving a language of thought and belief boxes may be a fiction. But it remains plausible that thoughts correspond to specific states of a cognitive system, playing specific functional roles. If the state plays the right sort of role, the corresponding thought will be a belief, and so on. Even without a language of thought, Sue's utterances of S_1 and S_2 are plausibly brought about by quite different states, one of which plays the functional role of a belief and the other one of which does not. So one states corresponds to a belief, and the other to a thought that is not a belief. All of this is so far quite compatible with a view on which these two thoughts have the same content.

We can extend the thought-based analysis to sentence types as well as tokens, by saying: a subject knows (or believes) a sentence type S if the subject has a thought that could be expressed by a fully competent utterance of S and that constitutes knowledge (or belief). Here, the idea is that at least some sentence types (in a given language) put strong constraints on the class of thoughts that they can express, at least when used with full competence (that is, used correctly and without deference to a linguistic community). In practice, my talk of knowledge of sentence types will be restricted to knowledge of certain context-independent sentences (typically in a scrutability base) for which these constraints are strong, along with 'I' and 'now' sentences for which the constraints are also strong.

We can also say that a subject is *in a position to know* a sentence token S if the thought expressed by S constitutes potential knowledge: that is, if it is possible that on (perhaps idealized) reflection the thought will come to constitute knowledge. A subject is in a position to know a sentence type S iff the subject is in a position to have a thought that could be expressed by an utterance of S and that constitutes potential knowledge. In both cases, the idea is roughly that

there is some possible process of reasoning that starts from the subject's initial psychological state, involving the initial thought, and that ends with the thought constituting knowledge. For present purposes, we need not suppose that this reasoning process involves a priori reasoning alone: the subject can appeal to any number of empirical background beliefs. But we can stipulate that it does not involve the subject's engaging in *further* empirical discovery, based on perception, testimony, and so on.

We can also stipulate a corresponding notion of potential a priori knowledge. We can say that a subject is in a position to know a sentence token *S a priori* if the thought expressed by *S* constitutes potential a priori knowledge: that is, if it is possible that on (perhaps idealized) reflection the thought will come to constitute a priori knowledge. In this case, one can say that the the sentence token *S* is *knowable a priori*, or simply *a priori*. I discuss the notions of apriority and a priori knowledge at much more length in the first section of Chapter 4.

The notion of potential knowledge (whether a priori or not), like other notions I will use, presupposes that thoughts can be reidentified over time. This notion of persistence over time might give rise to objections. What is it for a thought at t_2 to be the same thought as a thought at t_1? One might suggest that this is simply for the thought to have the same content. But then one's conclusions about potential knowledge and apriority will be hostage to one's theory of content, and it is not clear that they can then be used to ground a theory of content.

However, persistence should not be understood as sameness of content. On some views of content, such as a fine-grained Fregean view, persistence may not even entail sameness of content. For example, my thought that I am hungry now might persist as my thought that I was hungry then, which arguably has a different fine-grained Fregean content. On almost any view of content, sameness of content does not entail persistence: if I think that *p* at t_1, and at t_2 I have a causally independent thought that *p*, then the latter does not persist as the former. This applies even on a Fregean view, but is particular clear on a Russellian view of content: when Sue has thought T_2 shortly after T_1, these thoughts may have the same content, but T_1 certainly does not persist as T_2. It is plausible that for T_1 (the thought expressed by 'Hesperus is Hesperus') to persist, it must always have a trivial character quite distinct from the character of T_2.

Instead, persistence requires an appropriate continuity between thoughts over time.[3] It is plausible that this continuity requires either sameness of content or temporal relatedness of content along the lines above (perhaps with common or related guises or Fregean modes of presentation), but the reverse is not the case. It also requires more than this, including causal and psychological continuity. In any case, I will not define persistence in terms of sameness of content, but will leave

the notion as primitive for present purposes.

Most importantly, it is plausible that persistence is an intuitive notion that everyone needs, whatever they think about the theory of content. Almost everyone allows that a thought can come to be justified, or that it can come to be confirmed by evidence. Making sense of these notions requires the notion of persistence. The notion, along with the related notions of negation, disjunction, and conjunction in thought, are also crucial to understanding the notion of *inference* in thought. Suppose one reasons: *A*, *B*, therefore *A&B*. For this to be a valid inference in thought, conferring immediate justification of the conclusion, one's initial thought that *A* must be appropriately related to one's later thought *A&B*, intuitively acting as a conjunct of that thought. If one formed an independent thought with the same content, then this thought would not acquire the same sort of immediate justification. So an appeal to the notion of persistence does not presuppose commitment to any theoretical account of content, or to any technical notions such as apriority. The notion is already manifest in our ordinary notions of justification, confirmation, and inference in thought.

Some philosophers are doubtful about the very idea of token thoughts. For example, Lewis (19xx) suggests that "beliefs" is a "bogus plural". On this view, subjects can certain believe that P for various P, but there are no token entities called beliefs to undergird this believing, except perhaps for trivial derivative entities such as the instantiation of the property of believing that P. Presumably Lewis would take a similar view of thoughts. For someone with this view, the current definitions of key notions such as knowing a sentence will be problematic.

In response, it is worth noting that the thoughts that matter for present purposes are *occurrent* thoughts, and some reasons for doubts about token non-occurrent states do not apply to them. I think that it is a phenomenological datum that there are such mental events as acts of judging, or acts of entertaining a hypothesis. Given that there are such acts, it is hard to deny that there are thoughts.

Still, if someone does reject the very idea of token thoughts, the key notions such as that of knowing a sentence token S (and the apriority of S, discussed below) will have to be understood differently. If the theorist (like Lewis) takes a fine-grained view of propositions, so that 'Hesperus is Hesperus' and 'Hesperus is Phosphorus' express distinct propositions, it will suffice for present purposes for them to understanding knowledge of S as knowledge of the proposition it expresses.

---

[3]It is worth noting that for present purposes, the most essential applications of the notions of persistence can be restricted to occurrent thoughts within a single brief stream of thought: we can imagine an idealized thinker entertaining the thought, and coming moments later to justify it.

If the theorist takes a coarse-grained view of propositions but allows that there are guises, it will suffice for them to understand knowledge of a S as knowledge of a proposition under the guise associated with S, as above. If the theorist accepts none of these notions, an alternative approach will be needed. But in any case, the intuitive distinction in the status of various utterances is clear, and any theory that cannot explain it is an incomplete theory. So I will take it that even for theorists who reject all of the notions just mentioned, some way of understanding the key notion of knowing or believing a sentence can be found.

It is also not out of the question that one could take the notion of a thought as a useful fiction for motivating an intuitive idea of the apriority or scrutability of an utterance, and leave that notion unexplicated in the short term. Once one has motivated scrutability theses and the like in this way, one can use them to motivate a Fregean notion of proposition, and of the propositional content of a sentence. With this done, one could return to a characterization of apriority in terms of Fregean propositions.

One more terminological note: I will say that to know a *class* of sentences C is to know all the members of C. I will occasionally speak of knowing that the members of C are true, which should be understood to be saying the same thing. Likewise, from now on talking of knowing that a sentence S is true should be understood as equivalent to talk of knowing S as understood above, rather than being understood as invoking metalinguistic knowledge.

4   Empirical Scrutability

A truth S is empirically scrutable from a class of truths C, for a subject, iff were the subject to come to know the truths in C, they would be in a position to come to know S. The Empirical Scrutability Thesis says: there is a compact class of truths such that all truths are empirically scrutable from truths in that class. That is: there is a compact class of sentences such that for all truths S, if the subject were to come to know the truths in that class, they would be in a position to know S.

(One can formalize the thesis as follows, where it is understood that the variable C ranges of compact classes of sentences, S ranges over truths, $K(C)$ says that one knows all truths in C, $PK(S)$ says that one is in a position to know S, and $\Rightarrow$ is the counterfactual conditional: $\exists C \forall S (K(C) \Rightarrow PK(S))$. Here the variable C is understood to range over compact classes of sentences, S is understood to range over truths, and relativization to subjects and times is omitted for simplicity.)

One can illustrate the thesis by first selecting a scrutability base. Let us say that G includes

all truths in a vocabulary that comprises the following expressions: (i) microphysical expressions (in the language of a final physics); (ii) macrophysical expressions (in the language of classical physics), (iii) expressions for arbitrary mental states, including phenomenal states and propositional attitudes (to avoid trivialization we can avoid expressions for factive states such as knowledge), (iv) expressions for lawhood, causation, and counterfactual dependence, (v) logical and mathematical expressions, and (vi) the indexicals 'I' and 'now'. Not much will depend on this choice of basis (and one could add expressions for secondary qualities and normative expressions if one likes, along with others). But it is supposed to be a reasonably generous basis, so that there is at least a chance that many or all truths will be scrutable from it.

It seems clear that *if* a subject knew all truths in G, then they would thereby be in a position to know many more truths about the world. And it does not seem immediately out of the question (at least modulo the problems discussed below) that for any truth T, they would be in a position to know T. If so, all truths would be empirically scrutable from G.

The empirical scrutability thesis is useful for our purposes in part because it avoids technical notions such as apriority and analyticity, and in part because it lacks the rationalist flavor of some other scrutability thesis. If anything, this thesis has an empiricist flavor, saying all the things that one could come to know by knowing certain (largely empirical) truths. So it provides a good entry point to scrutability thesis for one who is skeptical of rationalism and the a priori.

As stated, the thesis requires considering a scenario in which the subject comes to know all truths in a compact class of truths C (I will henceforth put this by saying that the subject knows C), which may be a huge class of truths about all of spacetime. Of course to know all truths in such a class, the subject would need to have cognitive capacities greater than any actual human subject. So as before, we need to idealize to make sense of this scenario.

There is a worse problem than idealization, however. For the empirical scrutability thesis to be plausible, C will have to be sufficiently encompassing that its truths are true of this world and this world alone. But then, assuming that no-one actually knows C, it will be impossible to know C. Any world in which someone knows C would have to differ from the actual world, and therefore must be a world in which C is false. But there are no worlds in which someone knows C and C is false. So no-one can know C. If so, then the Empirical Scrutability thesis is vacuously true (if counterfactuals with impossible antecedents are vacuously true), or at least hard to assess (if not).

One can get around the immediate problem, as well as weakening the idealization, by weakening the notion of empirical scrutability. We can say that a truth S is empirically scrutable from a class C of truths if there is a knowable *subclass* of C such that if the subject were to come to

13

know the truths in that subclass, they would be in a position to know S. [$\exists C \forall S \exists C' \diamond K(C') \& (C' \subseteq C) \& (K(C') \Rightarrow PK(S))$]. The empirical scrutability thesis will then require that the subject knows only those members of the base class C that are needed in order to know S. In most cases, this will be a much smaller class, avoiding the need for a strong idealization in these cases. And for many such subclasses, there will be no problem with the subject knowing the subclass. Of course there may still be some truths for which empirical scrutability in this sense fails, for reasons below, but at least this notion of empirical scrutability has application in many cases.

A resulting worry is that by weakening empirical scrutability in this way, one weakens the force of the thesis significantly. For many truths S, there will be some C-truths such that knowing these truths justifies belief in S, but there will be other C-truths such that knowing these truths justifies rejecting S. In such a case, it is not out of the question that knowing all truths in the first class would put one in a position to know S, but that knowing *all C*-truths would not put one in a position to know S. If so, then one will have to be careful in making inferences from these weakened empirical scrutability claims to other scrutability claims.

In any case, a version of the second problem recurs. If a subject can utter a sentence S_1 that is true only in the world of utterance, without knowing S_1, then that sentence will be unknowable, for the same reason as before. Less exotically, say that S_2 is a true utterance of 'P and I do not know that P'. Then S_2 is true, but it cannot be known. If S cannot be known, then for any set of truths in C that one can know, knowing these truths will not put one in a position to know C. A third case: say that S_3 is 'I know no sentences in C'', uttered by a subject who has beliefs about some sentences in C but does not know them, and is not now in a position to know that she does not know them. Then S_3 is true, but the subject is not in a position to know it by coming to knowing any subclass of sentences of C. So it appears that the weakened empirical scrutability thesis is still false.[4]

---

[4]A tricky related case is a true utterance of S_4, 'The time is now t' (for some precise temporal specification t), at t, by a subject who does not know S_4. Then S_4 is true. An initial worry is that the subject cannot come to know S_4, as knowledge takes time, and by that time S_4 will be false. This worry is unfounded, as S_4 is a token that stays true forever, and the subject can come to know S_4 if the thought expressed by S_4 comes to constitute knowledge. Of course this requires an appropriate account of the persistence of thoughts, one on which the thought expressed by 'The time is now t' persists not as a later thought that would be expressed by 'The time is now t', but rather as a later thought that would be expressed by something like 'The time was then t'. A related worry applies to sentence types such as 'It is now t' that might occur in the base class C. After t, the subject cannot come to have knowledge apt to be expressed by that sentence type. So instead, we should understand "come to know" so that at t, the subject can come to know a sentence type S' iff the subject can come to have thoughts that constitute knowledge and that would have been apt to

---

14

Various fixes suggest themselves. The first is to allow that counterfactuals with impossible antecedents are not always vacuously true. Then one might allow (nontrivially) that *if* one, per impossibile, came to know all truths in C, then one could know sentences such as S above. This strategy is not uninteresting, but I will set it aside in what follows.

A second, less exotic fix stems from the observation that in all these cases, it is not ruled out that one can come to know *whether* the sentence S is true (where knowledge of whether S is true is the natural generalization of knowledge that S is true as explained above). It is just that the very process of coming to know whether S is true (by the procedure of coming to know C-truths) will render S false.

We might call truths like this *Fitchian* truths, because Fitch's unknowable truth 'P and I don't know P' is a paradigm. We can say that S is a Fitchian truth if attempting to know the truth-value of S will render S false. Of course there are different methods by which one might come to know the truth-value of S, so one might also say that S is Fitchian with respect to a method iff attempting to know the truth-value of S by that method will render S false. Then S_1 and S_2 above are Fitchian with respect to any method, while S_3 is Fitchian with respect to the method of determining truth-value via knowledge of sentences in C.

One might then suggest a modified thesis, saying that there is some subclass of C such that if one were to come to know this subclass, one would be in a position to know *whether S* is true. $[\exists C \forall S \, \exists C'((C' \subseteq C) \& (K(C) \Rightarrow PK^*(S)))]$ Or better, one can suggest that there is some subclass such that if one were to come to know *whether* the sentences in this subclass are true, one would be in a position to know whether S is true. $[\exists C \forall S \, \exists C'((C' \subseteq C) \& (K^*(C') \Rightarrow PK^*(S)))]$ This latter formulation allows for the possibility that the process of coming to know the sentences may also change the truth-value of sentences in C, as well as the sentence S.

This modified thesis is not threatened by Fitchian truths or by any of the cases above. There are some smaller residual worries. One worry is that in cases where one cannot know that S is true without knowing whether a very large subclass of C-sentences is true, worlds where one knows that subclass may be so different from our world (in the cognitive capacity they allow, for example) that they are nomologically impossible. In such a world, knowing whether S is true might require knowing about alien features of that world not described by C-sentences. If so, then no knowledge of C-sentences will put one in a position to know whether S is true. It is unclear whether this scenario can arise. But a milder version of the worry applies more generally. For any

be expressed by S' at t.

15

truths S such that one can know the truth-value of S only by knowing some (actual) truths in C to be false, then the empirical scrutability relation between S and C does not tell us directly about the status and grounds of S in the actual world. Correspondingly, one will not be able to use empirical scrutability to argue directly for the scrutability of the truth of S from the truth of C in the actual world. So the force of the scrutability thesis is weakened somewhat. Still, the modified empirical scrutability thesis remains interesting and important.

A third strategy is simply to exclude Fitchian cases, and require only that all non-Fitchian truths (with respect to the method of empirical scrutability from C-truths) are empirically scrutable from C-truths. (For some purposes one will need to modify this thesis somewhat further, for reasons I wil discuss in the next chapter.) This strategy loses the universal scope of the thesis, of course, and it may appear somewhat ad hoc. Nevertheless, for the purpose of at least motivating scrutability theses, then the thesis may have a role to play. If all non-Fitchian truths are empirically scrutable, if empirical scrutability entails some other sort of scrutability, and if that other sort of scrutability is not subject to worries about Fitchian cases, then there is at least a reasonable prima facie case for a universal scrutability thesis concerning the other sort of scrutability.

A fourth fix, and perhaps the most natural, is to move to conditional scrutability, below.

These imperfections in the Empirical Scrutability thesis will not matter too much for my purposes. For these purposes, the thesis is mainly of instrumental value. It is mainly valuable for its role in helping to argue for other scrutability theses, such as Conditional and A Priori Scrutability, and in providing initial motivation for these scrutability theses for those who may be skeptical about them. If it turns out that the only problem for Empirical Scrutability is the Fitch-style problem, and that this problem does not affect the other theses, then the Empirical Scrutability thesis can still play these roles reasonably well.

5   Conditional Scrutability

A truth S is conditionally scrutable from a class of truths C, for a subject, iff the subject is in a position to know that *if* the members of C are true, then S is true. The conditional scrutability thesis says that there is a compact class of sentences C such that all truths S are conditionally scrutable from the C-truths. This formulation immediately avoids the Fitchian problems above. Even if S is unknowable, there is usually no problem knowing that if some other sentence T is true, then S is true. And in the cases above, there seems to be no problem with the idea that the relevant subjects are in a position to know (on idealized reflection) that *if* the sentences in C are

true, then S_1, S_2, and S_3 are true.

This thesis uses the notion of conditional knowledge: that is, knowledge of conditionals such as 'If P, then Q'. Such claims are common in English: it would be natural to say that I know that if it rains today, then my car will get wet. Such claims are about as common as claims about conditional belief, as when I say that I believe that if Australia bats first, Ponting will score a century. It is natural to hold that conditional knowledge stands to conditional belief much as knowledge stands to belief.

The correct analysis of conditional belief and conditional knowledge is nontrivial. It is implausible that conditional belief and conditional knowledge, at least as ordinarily understood in English, simply involve belief in or knowledge of a material conditional. For example, I might know that it is not raining, and thereby know (and believe) the material conditional 'If it is raining, then my car is dry'. But it is intuitively incorrect to say that I know (or believe) that if it is raining, then my car is dry.

It is somewhat more plausible to say that that conditional belief and knowledge involve belief in and knowledge of an *indicative* conditional. For reasons analogous to those above, most theorists deny that indicative conditionals are equivalent to material conditionals: intuitively, for an indicative conditional such as 'If it is raining, then my car is dry' to be acceptable, then there must be a stronger connection (perhaps an epistemological connection) between the antecedent and the consequent than the material conditional requires. But it is still not entirely clear what belief in or knowledge of an indicative conditional involves. For example, Lewis (19xx) gives good reason to think that conditional belief cannot simply be a matter of believing or knowing a *proposition*, at least while preserving one's epistemological intuitions. If so, then if conditional belief is belief in an indicative conditional, indicative conditionals cannot be understood as propositions.

The most common view of conditional belief holds that a subject believes that if P, then Q iff the subject's *conditional credence* in Q given P, $cr(Q|P)$, is sufficiently high. Here we adopt a view on which subject's have credence between 0 and 1 in various propositions: $cr(P) = 1$ when the subject is certain of P, $cr(P) = 0$ when the subject is certain of $\neg P$, $cr(P) = 0.5$ when the subject is entirely agnostic between the two, and so on. If P is the proposition that the dice will come up double-six, then my credence $cr(P)$ might be 1/36. In cases where the subject believes that P, then $cr(P)$ will be well over 1/2. To a first approximation, we can say that a subject believes that P iff $cr(P)$ is sufficiently high. It is plausible that the threshold for belief is context-dependent, vague, and differs between different propositions: for example, a credence of 0.999 may suffice for belief in some cases (belief that it will rain today) but not in others (belief that one will lose the

lottery). But we can understand "sufficiently high" to be context-dependent, vague, and variable between propositions in a similar way.

Subjects can also have conditional credences in one proposition given another. For example, if Q is the proposition that the red die will come up six, then my conditional credence $cr(Q|P)$ might be 1/6. In cases where $cr(Q)$ is greater than zero and where the subject is fully rational, $cr(Q|P)$ will be equal to $cr(P\&Q)/cr(Q)$. But for familiar reasons (Hajek 2004), it is reasonable to hold that subjects can have a conditional credence $cr(Q|P)$ even in some cases where $cr(Q)=0$. For example, if P is as above, and Q is the proposition that a randomly thrown dart lands exactly at position π on an interval, then $cr(Q)$ might reasonably be 0, while $cr(Q|P)$ might nevertheless reasonably be 1/6. So a subject's conditional credence in P given Q should not in general be understood as deriving wholly from the subject's credences in P, Q, and $P\&Q$. Rather, it should be understood as capturing some more complex cognitive dependence between the subject's attitudes to P and to Q.

Just as we can say that a subject believes that P when their credence $cr(P)$ is sufficiently high, we can likewise say that a subject believes that if P then Q when their conditional credence $cr(Q|P)$ is sufficiently high. Of course, much more needs to be said about just what "sufficiently high" involves. As before, we should expect that the threshold for conditional belief will be context-dependent, vague, and will differ for different pairs of propositions. But it is not implausible that what goes for unconditional belief also goes for conditional belief.

What about knowledge? In the case of unconditional knowledge that P, the justification requirement on knowledge plausibly corresponds to a claim that the subject is justified in having a sufficiently high credence $cr(P)$. This requires the idea that credences can be justified, and are subject to normative assessment. Some radical subjectivists reject this claim, holding that all credences in nonlogical propositions are equally reasonable; but this path leads easily to skepticism. If one holds that beliefs can be justified, it seems reasonable to hold that credences can be justified too. Of unconditional knowledge also requires that the proposition be true, and that some sort of anti-Gettier condition be satisfied, requiring for example that one's justification for the proposition is appropriately connected to the truth of the proposition.

Conditional knowledge that if P, then Q also plausibly requires that the subject is justified in having a sufficiently high credence $cr(Q|P)$. Insofar as unconditional credences can be justified, it is also reasonable to hold that conditional credences can be justified. A difficult questions concerns whether there is a truth requirement on conditional knowledge. It is not at all clear what it means to say that the conditional 'if P, then Q' is true, as opposed to acceptable for a subject. Still, there

are plausibly cases in which the subject has a high justified credence $cr(Q|P)$, but does not know that if P, then Q. This can happen if P is true and Q is false, or if the subject infers Q from P only with the aid of a false but justified belief in R, for example. So the question of just what needs to be added to justified conditional belief to obtain conditional knowledge remains an open question.

For present purposes, it might suffice to rely on our intuitive understanding of conditional knowledge, just as philosophers often rely on their intuitive understanding of knowledge even without an analysis of what an anti-Gettier condition involves. But an alternative way to proceed is to stipulate that for the purposes of the Conditional Scrutability thesis, what matters is justified conditional belief, not conditional knowledge. That is, we can modify the definition of scrutability so that conditional scrutability of S from C requires only that the subject is in a position to have a justified conditional belief that if the sentences in C are true, then S is true. And we could understand this notion in turn by saying that the subject is in a position to be justified in having a sufficiently high conditional credence in S given C.

In what follows, I will move back and forth between these related conceptions of conditional scrutability. The official thesis will be cast in terms of conditional knowledge, but I will often analyze things in terms of conditional credence. This is justified in part by the plausible thesis that conditional knowledge *requires* a sufficiently high justified conditional credence. When moving in the reverse direction, we can explicitly attend to the possibility of high conditional credence without conditional knowledge when it is relevant.

I will also adopt the idea that for at least some subjects and some propositions, there is a *rational* credence for the subject to have in the proposition, the credence that the subject ideally *ought* to have in the proposition. Or better, I will assume that for some subjects and propositions, there is a rational range of credences, in that one or more credences are rational, and some credences are irrational. Of course a subject's rational credences in a proposition may differ from the subject's actual credence in that proposition, if the subject is not ideally rational. We can say that the rational credence for a subject in a proposition is high when only high credences in the proposition are rational for the subject. As before, if we deny that subjects have high rational credences in some propositions, it is not easy to avoid skepticism.

We can then say that S is conditionally scrutable from C, for a subject, if the subject's rational credence $cr'(S|C)$ is high. Here the notion of rational conditional credence in a pair of propositions is understood in a way parallel to the understanding of rational unconditional credence in a single proposition, above. We can extend this notion to sentences and to thoughts as follows.

As before, we can suppose that sentence tokens express thoughts, and that (non-context-

dependent) sentence types are apt to express certain thoughts. We can also suppose that at least some thoughts T had by a subject have an associated credence $cr(T)$. Perhaps not all do, as when a subject merely entertains a thought without forming an opinion, and perhaps many credences are approximate, but this is good enough for our purposes. We can then also say that for some pairs of simultaneous thoughts T_1 and T_2, the subject has a conditional credence $cr(T_1|T_2)$. The rational credence of a thought T, $cr'(T)$, is the credence (or credence range) that the subject ideally ought to have in T. The rational conditional credence of T_1 given T_2, $cr'(T_1|T_2)$, is the conditional credence (or credence range) that the subject ideally ought to have in T_1 given T_2.

For a sentence token S, we can say that the subject's credence in S, $cr(S)$, is identical to $cr(T)$, the subject's credence in the thought T expressed by S, when there is such a credence. We can say that the subject's rational credence in S, $cr'(S)$, is identical to $cr'(T)$. For a sentence type R, we can say that the subject's rational credence in R, $cr'(R)$, is the range of possible values for $cr'(T')$, where T' is a thought apt to be expressed by R. Finally, we can say that the subjects rational credence in S given R, $cr'(S|R)$, is the range of possible values for $cr'(T|T')$, where T and T' are understood as above.

Where S is a sentence token and C is a class of sentence types, we can say that $cr'(S|C)$ is $cr'(S|C')$, where C' is a conjunction of all sentences in C.[5] Alternatively, we can say that $cr'(S|C)$ is the range of possible values for $cr'(T|T')$, where T is the thought expressed by T, and T' is a conjunction of thoughts apt to be expressed by sentences in C.

We can now say that S is conditionally scrutable from C iff $cr'(S|C)$ is high. The conditional scrutability thesis, stated fully in terms of types and tokens, says that there is a compact class C' of sentence types such that for all true sentence tokens S, the speaker's rational conditional credence $cr'(S|C)$ is high, where C is the class of truths (relative to the speaker and the time of utterance of S) in C.[6]

It has taken a while to unpack this thesis, but I think there remains some prima facie plausibility to it. It remains plausible that given such a class as G above (or better, some disciplined subclass

---

[5]Of course this requires that the class C is sufficiently disciplined, and/or that the language is sufficiently infinitary, so that a conjunctive sentence C' exists. (If C itself contains arbitrary conjunctions, we can avoid regress by stipulating that C' is a conjunction of all nonconjunctive sentences that are either members of C or conjuncts of conjunctive sentences in C.) The same goes for conjunctive thoughts.

[6]Whereas other scrutability thesis trivially rule out the possibility that false sentences are scrutable from truths, this is not quite so trivial for conditional scrutability. So one might consider explicitly making the thesis a biconditional, saying that S is true iff it is conditionally scrutable from C. Still, this claim plausibly follows from the original version, along with the claim that when S_1 and S_2 express incompatible thoughts, $cr'(S|C)$ and $cr'(S_2|C)$ cannot both be high.

of it), then for at least many truths S, ideal reasoning would support a high conditional credence in S given the hypothesis that all the sentences in G are true.

In many cases, when S is empirically scrutable from C, it will be plausible that S is conditionally scrutable from C.[7] Suppose that empirical scrutability holds: if our subject were to come to know all the truths in C, they would be in a position to know S. Suppose in addition that in this process, the truths in C constitute the total additional evidence that the subject needs in order to know S. According to a standard Bayesian principle, if one can rationally come to have a high credence in P through acquiring total evidence E, then one's antecedent rational credence $cr'(P|E)$ must be high. It follows from this principle and our suppositions, along with the thesis that knowledge entails a rational high credence, that the subject's antecedent rational credence $cr'(S|C)$ is high. That is, it follows that S is conditionally scrutable from C.

Things are more complicated than this, as it is not the case that in all nearby situations where the subject comes to know C, the truths in C constitute the subject's total additional evidence E. A minor worry arises if E is a proper subclass of C: in this case, $cr(S|E)$ will be high, and as long as the rest of C does not undermine S (as presumably it does not, as C is itself grounded in E), $cr(S|C)$ will be high. More of a worry arises if E includes evidence not included in C. It may be that it is impossible to come to know C (in a nearby world) without acquiring further evidence E': perhaps E' is the basis needed to know C, or perhaps knowledge of E' always follows along from knowledge of C given the subject's psychology. In the former case, it remains plausible that $cr(S|C)$ is high for the reasons given earlier in the paragraph. The latter case is tricky, though.

The most obvious case in which coming to know C inevitably yields additional evidence E' is one in which E' is 'I believe C'. Knowledge of C, will typically lead to introspective evidence E', even if E' is not included in C. In such a case, E' itself will be empirically scrutable from C, but it may nevertheless be that the conditional credence $cr'(E'|C)$ is low. So this is a case in which empirical scrutability does not yield conditional scrutability.

Still, the core cases of empirical scrutability discussed earlier (empirical scrutability from G, for example) would remain plausible even if one strengthened the claim to: if one were to come to know the truths in C, and use no additional evidence other than the truths in C, one could thereby come to know S. In this case, it follows from plausible Bayesian principles that $cr'(S|C)$ is high. So this mildly strengthened empirical scrutability principle leads to conditional scrutability.

So there is good reason to accept that in the core cases in which empirical scrutability holds,

---

[7]I may drop these five paragraphs, as the next chapter goes over this ground in any case.

conditional scrutability holds. Furthermore, even in the Fitchian cases in which empirical scrutability fails, there is no corresponding reason to think that conditional scrutability fails. So if empirical scrutability is true of all non-Fitchian cases, then there is a good case for thinking that conditional scrutability holds in general.

One concern about the conditional scrutability thesis is that the idealization involved in it is not as easy to attenuate as with other scrutability theses. In the case of empirical scrutability, we attenuated the idealization by moving to a notion on which S is empirically scrutable from C if there is some *subclass* C' of C such that knowing C' would put one in a position to know S. One might correspondingly move to a notion on which S is conditionally scrutable from C if there is some subclass C' of C such that $cr'(S|C')$ is high. The trouble with doing this is that for almost any S, including false S, this definition will be satisfied. For most false sentence tokens S, there are some truths that taken collectively would constitute strong misleading evidence for S. If C contains such sentences, constituting a subclass C', then $cr'(S|C')$ will be high, and S will be conditionally scrutable from C even though S is false.

A version of this problem arose for empirical scrutability, but it was not nearly as acute: no false sentence token can be known, so no false sentence can be empirically scrutable. To avoid this worry in the case of conditional scrutability, it is best not to employ the weaker subclass-involving notion of conditional scrutability. Of course this means that to make an inference from empirical scrutability to conditional scrutability, one will also have to employ the stronger notion of empirical scrutability that avoids subclasses (although one might still use the weaker sort of empirical scrutability to help make the case for the stronger sort). Where conditional scrutability, there is no getting around the need for a strong idealization at some point.

Perhaps the most pressing worry for conditional scrutability arises from situations in which a subject's beliefs are unreliable. Say that John has just been given an *anti-arithmetic* drug that renders him entirely incompetent at doing arithmetic: any arithmetical belief of his will have at best a 50 percent chance of being true. Let M be a token of '57+65=122' uttered by John, and let D be 'I have just been given the anti-arithmetic drug'. If John came to know D, the rational thing for him to do would be to suspend judgment about whether M is true. And even if John does not know D, it is plausible that his credence in M *conditional* on D should be around 0.5. If so, then as long as D is itself scrutable from C (as it must be if conditional scrutability holds), then $cr'(M|C)$ will be around 0.5. But M is true. So M is a truth that is not conditionally scrutable from C, yielding a counterexample to conditional scrutability.

One might respond that the ideally rational credence $cr'(M|C)$ is 1, on the grounds that if John

were ideally rational, then even on the supposition that D, his ideal reasoning would allow him to know M with certainty. But this does not seem quite right. It is plausibly *irrational* to accept simultaneously that one's belief in M is unreliable and to be certain that M. If I were to learn D (and to acquire no other new evidence), then rationality would require that if I do not question D, I should suspend judgment about M. So even before learning D, my credence $cr(M|D)$ should not be high. Even for an ideally rational being, $cr(M|D)$ will not be high. An ideally rational being may well have some tiny positive credence in D, and that area of their credence space will be equally divided between M and $\neg M$. It follows that for each of us—whether or not we are ideally rational, and whether or not we have recently been given such a drug—$cr'(M|D)$ will not be high. The same goes for John. So the counterexample stands.[8]

This sort of problem affects much more than conditional scrutability. Christensen (forthcoming) has observed that a problem of this sort affects even our knowledge of logical truths. It is commonly held that if L is a logical truth, the rational credence $cr'(L)$ is 1. But it may well be rational to have a small positive credence in the thesis T that one is unreliable about logic: after all, one cannot exclude with certainty the hypothesis that one has recently been given an anti-logic drug. For the reasons above, $cr'(L|T)$ cannot be high. It follows that $cr'(L)$ cannot be 1.

The best way to handle this problem is to disentangle various different principles of rationality. It is clear that the principle of rationality invoked above is quite different in kind from ordinary principles of rational inference and the like. We might call it a level-crossing principle: it is a principle by which one's higher-order beliefs about one's cognitive capacity are used to restrain one's first-order beliefs about a subject matter. This is a sort of practical principle governing how one should reason, but one quite different from familiar practical principles that derive directly from theoretical reasons. In the case above, one has theoretical reasons in support of M, which

---

[8]One version of this response holds that when one reasons ideally, one can know with introspective certainty that one is reasoning well (thanks to Declan Smithies for this suggestion). If so, then ideal reasoning to M might also yield certainty that one's belief in M is based on good reasoning, and thereby might allow one to exclude the possibility that one is reasoning unreliably in this instance. This may be compatible with the hypothesis that one has taken the drug, as long as the drug allows some occasions of good reasoning. In this case, $cr'(M)$ will be 1 and $cr'(M|D)$ will be 1. If the drug hypothesis D rules out any good reasoning, then on this view $cr(D)$ will be 0. Still, this leaves open the question of what $cr(M|D)$ should be, especially once we allow credences conditional on hypotheses with credence 0, and once we acknowledge that the supposition of D trumps any empirical evidence (including introspective evidence) with which it conflicts. I think it is natural to say that under this supposition, the introspective evidence is rendered irrelevant, so that $cr'(M|D)$ should not be high. In any case, it is far from clear that ideal reasoning provides this sort of introspective certainty.

yield practical reason to accept M, but the level-crossing principle gives practical reason to suspend judgment about M, and in this case the second practical reason wins.

We can imagine a cognizer—call him Achilles—who is at least sometimes insensitive to this sort of level-crossing principle. On occasion, Achilles goes into the mode of *insulated cognition*. When in this mode, Achilles goes wherever his theoretical reasons take them, entirely insulated from higher-order beliefs about his cognitive capacity. He might acquire evidence that he is unreliable about mathematics, and thereby come to believe 'I am unreliable about arithmetic', but he will go on drawing conclusions about arithmetic all the same. We might say that in the insulated mode, his reasoning is *practically self-confident*, even if it is not *theoretically self-confident*. That is, any self-doubt manifests itself only in what Achilles believes, and not in how he goes about believing.

What if Achilles comes to believe that he has been taking a falsity pill, so that all of his mathematical beliefs are false? Then he will believe 'All my M-beliefs are false'. At the same time, through introspection he may figure out 'I have the M-belief that FLT is true'. From these he would infer, by ordinary theoretical reasoning, that FLT is false. To avoid this result, we need to stipulate that when in the insulated mode, Achilles is also incapable of introspection. In fact, to avoid indirect evidence of his beliefs through observing his behavior, we can stipulate that in the insulated mode, he is incapable of perception, too.[9]

There is plausibly some sense in which insulated cognition is irrational, but it is a limited sort of irrationality. Suppose that Achilles is otherwise fully rational. And suppose that at a certain point of time, he might either engage in insulated reasoning or fully rational reasoning, where we stipulate that in both cases this involves armchair reasoning (without perception or introspection) that may exploit existing beliefs. Then insulated cognition will yield at least as many true beliefs as the fully rational mode, and in some cases (those in which Achilles has misleading evidence for his irrationality) it will yield more. It is only if Achilles has independent sorts of theoretical irrationality that his practical irrationality will be a problem, causing him to keep forming false

---

[9]What if Achilles also has a prior belief, formed before he entered insulated mode, that he will believe FLT? One could attempt to exclude such beliefs by requiring that Achilles disregard all evidence and beliefs from before he entered insulated mode, or at least that he disregard all empirical evidence and beliefs. But this would be too close to restricting Achilles to a priori reasoning, which would have the undesirable effect of making the notion of conditional scrutability depend on the notion of apriority. Instead, one can rely on the observation that if Achilles believes or supposes that he has been taking the drug, this will have the effect of undercutting his prior grounds for believing that he will believe FLT, or at least of rendering any such belief uncertain. So his insulated reasoning in support of FLT will overwhelm these defeated grounds for denying FLT.

beliefs where a fully rational creature would be restrained.

Insofar as it is reasonable to postulate ideal cognizers at all, there seems to be no bar to postulating *insulated ideal cognizers*: cognizers whose rational processes are practically insulated from higher-order beliefs, as Achilles' processes are, but are otherwise ideal. As we have seen, insulated ideal cognizers are in some ways more successful cognizers than fully ideal cognizers, at least where non-empirical reasoning is concerned, because their cognition is never affected by misleading self-doubts. For example, a fully ideal cognizer may have some small positive credence in its own unreliability (it cannot exclude with certainty the hypothesis that it has recently been given the drug above), so it will correspondingly never be absolutely certain of anything, even of logical truths. By contrast, there is no corresponding bar to an insulated ideal cognizer's being certain of logical truths.

Rational idealizations need not to be cashed out in terms of ideal cognizers.[10] More fundamentally, they involve idealizations in normative terms, speaking of what one ideally ought to do or believe. Where an insulated idealization is concerned, one can likewise cash this out in normative terms, speaking of what one ideally ought to believe (starting from one's current state) if cross-level principles are set aside. We might even define a notion of what one "ought*" to believe that works in this way, and a corresponding notion of one's "rational*" credence in a proposition, $cr^*(P)$. For example, where L is a logical truths, then even if one's ordinary rational credence $cr'(P)$ is less than 1 for the reasons above, it may be that the insulated rational credence $cr^*(P)$ is 1.

We can then say that S is conditionally scrutable from C if the *insulated* rational credence $cr^*(S|C)$ (which is identical to $cr^*(S|C')$, where C' is a conjunction of the sentences in C) is high. Understood this way, the drug case poses no problem for the conditional scrutability thesis. In this case, although $cr'(M|D)$ is not high, $cr^*(M|D)$ is still plausibly high. The belief that one is unreliable about arithmetic has no impact on one's insulated rational credence in M, and likewise the supposition that one is unreliable has no impact one one's insulated conditional rational credence.

---

[10]Still, a bonus of the insulated idealization is that it overcomes one familiar problem in appeals to ideal cognizers. Typically, one cannot simply identify one's rational credence in P with the credence one would have in P if one were to become an ideal cognizer, as this would entail that everyone has an overly high rational credence that they are ideal cognizers, and so on. However, this problem does not arise on the insulated idealization, because of the bar on introspection. So this problem does not exclude the thesis that one's rational credence in P is the credence one would have in P if one were to become an insulated ideal cognizer (starting from one's current state). I will not rely on this thesis, in part because it is not obvious that there could be a truly ideal cognizer, as opposed to a series of more and more ideal cognizers. But the thesis may occasionally be a useful aid to thinking about insulated credences.

So conditional scrutability is unthreatened.[11]

I think that for many purposes involving theoretical rationality, insulated rational credence is often the most useful notion. Certainly, insulated rational credence seems to better reflect the sort of claims that theorists often make about rational credence. Where non-insulated rational credences are concerned, even tiny empirical self-doubts will infect the analysis of all sorts of otherwise well-behaved matters, in ways that are hard to regiment, and that will render many standard claims of formal epistemology false. (Logical truths will not have rational credence 1, Sleeping Beauty will not have rational credence 1/3, and so on.) The insulated idealization keeps the focus on truly theoretical rationality, allowing a more straightforward analysis of theoretical reason. Of course when it is relevant one can still invoke a noninsulated idealization, in order to see how practical and theoretical rationality interact, and to determine what it is rational (simpliciter) for a subject to believe.

The insulated idealization also allows us to at least entertain a stronger conditional scrutability thesis. We can define a notion of conditional scrutability so that S is conditionally scrutable from C iff $cr^*(S|C) = 1$. And we can define a corresponding scrutability thesis holding that $cr^*(S|C) = 1$ for all truths S and for the appropriate compact C. Where ordinary rational credences are concerned, a thesis as strong as this is out of the question for reasons discussed above. But for insulated rational credences, the thesis is not obviously out of the question, at least for some appropriate C. In fact I am inclined to think that a thesis of this sort is true. I will return to this matter in the next chapter.

6   A Priori Scrutability

The A Priori Scrutability thesis says that all truths are a priori scrutable from a compact class of base truths. In the last chapter, I said that the a priori scrutability of a sentence S from a class C requires that S can be logically derived from some sentences in C and some a priori truths, or that a material conditional from a conjunction of sentences in C to S is a priori. These definitions are coherent when S is a sentence type, but it is not obvious how to make sense of them when S is a sentence token. That is, it is not obvious what it is for a sentence token to be a priori derived from other sentences, or what it is for a sentence token to reappear in the consequent of a material conditional.

To handle this matter, we can define a priori relations between sentences in terms of a priori relations between thoughts. Recall that a thought is a priori when it constitutes potential a priori

knowledge. We can likewise say that for thoughts T_1 and T_2 had by the same subject, T_1 *a priori entails* T_2, or *implies* T_2, when a disjunction of T_1 with a negation of T_2 is a priori. This definition invokes the idea that one thought can be a negation of another, intuitively when it is formed by negating the former thought. And it invokes the idea that one thought can be the conjunction of some others, or the disjunction of some others, intuitively when it is formed by conjoining or disjoining those thoughts. I think we have a clear intuitive grasp of these notions, although as before they require something akin to the notion of persistence of thoughts over time.

We can then say that a sentence token S_1 a priori entails (or implies) by another sentence token S_2 when the thought expressed by S_1 implies the thought expressed by S_2. A sentence type S_1 implies a sentence token S_2 iff some possible S_1-thought implies the thought expressed by S_2, where an S_1-thought is one apt to be expressed by a fully competent utterance of S_1. A sentence type S_1 implies a sentence type S_2 iff some possible S_1-thought implies some possible S_2-thought.[12]

We can then say, as before that a sentence (token or type) S is *a priori scrutable* from a class of sentence types C if S is a priori entailed by a conjunction of members of C. Here, the rough idea is that supposing some conjunction of base sentences in C and reasoning a priori enables one to conclude that S is true.

Note that a priori scrutability, unlike empirical and conditional scrutability, requires that S be epistemically related to *some* conjunction of members of C, rather than to the conjunction of all members. We could define it the latter way, but the two definitions are near-equivalent. If S is a priori entailed by a conjunction of all members of C, it is trivially a priori entailed by some conjunction of members of C. If S is a priori entailed by some conjunction of members of C, then S is a priori entailed by a conjunction of all members of C, at least if such a conjunction exists. This parallels the observation that if a material conditional $A \supset S$ is a priori, then so is any material conditional $A\&B \supset S$, because the former conditional entails the latter. As before, this attenuates the required idealization somewhat: to establish a priori scrutability of S from C, we need not always consider the conjunction of all members of C. A priori entailment by some

---

[12]As discussed in chapter 4, the requirement of metaphysical possibility is probably dispensable here. For example, on a view of metaphysical modality that allows certain brute metaphysical necessities, it might be held to be metaphysically necessary that subjects can only entertain thoughts of a certain complexity. If so, then if S_1 has great complexity than this, there will be no possible S_1-thoughts and S_1 will not imply anything. On such a view, as before, it is better to sever apriority from the definition in terms of metaphysical possibility. Instead, one might define it in terms of epistemic norms: for example, there are epistemic norms that dictate that from an S_1-thought and a priori reasoning, one should conclude S_2.

proper subclass of *C* will suffice.

A priori scrutability is clearly not subject to the Fitchian problems that arose for empirical scrutability. It is also not subject to the self-doubt problems that arose for conditional scrutability. If *M* is a mathematical truth, then given that *M* is a priori, any material conditional $D \supset M$ will also be a priori. For similar reasons it follows that *M* is a priori scrutable from any base, even a base specifying a world where the subject has been given an anti-arithmetic drug. At worst, the Achilles worry suggests that a subject cannot be come to be *certain* of *M* by a priori reasoning, for reasons discussed earlier. But if one wants to require certainty, one might perhaps invoke an insulated idealization in one's definition of apriority.

In a number of respects, a priori scrutability is better behaved than the other sorts of scrutability, and it is the notion that I will concentrate on the most. The main downside of the notion compared to the others is that it invokes the more theoretical and controversial notion of the a priori. So it is useful to have the other notions too, to help motivate and argue for scrutability claims. Still, a priori scrutability will be the central focus.

7   Generalized Scrutability

Scrutability theses need not be restricted to the actual world. If the scrutability thesis is true, then it is plausible that it still would have been true if the world had turned out differently. To see this, note that we can evaluate the truth of various sentences even given *hypothetical* information about ways the world might be. For example, in the Gettier case, it is irrelevant whether Smiths case is actual: a subject can know that *if* Smiths case as described is actual, then Smith does not know that someone owns a Ford. Or in the case of water, given an appropriate specification of the distribution, behavior, and appearance of clusters of XYZ molecules (information analogous to the information we have about H_2O in the actual world), a subject is in a position to conclude that *if* the specification is correct, then water is XYZ.

One might formulate a stronger scrutability thesis by requiring that the A Priori Scrutability thesis be necessary, or better, by requiring that it be a priori. If A Priori Scrutability is itself a priori (in the strong sense outlined in the appendix to chapter 3), then it will be true however the world turns out. That is, for every epistemically possible scenario, it will be a priori that if the scenario obtains, then A Priori Scrutability is true. So for every scenario, it would be a priori that truths (with respect to that scenario) are scrutable from a compact class of truths (with respect to that scenario). This would require that there is a compact scrutability base for each scenario, although

there might be entirely different scrutability bases for different scenarios.

For my purposes, it is useful to formulate a somewhat stronger thesis still, according to which there is a single scrutability base that applies to all scenarios. We can formulate such a thesis as follows. Let us say that a sentence S is epistemically possible iff $\neg S$ is not a priori, and that a class of truths is epistemically possible if every conjunction of truths in that class is epistemically possible. Then we can formulate:

> *Generalized Scrutability*: There is a compact class C of sentences such that all epistemically possible sentences are scrutable from some epistemically possible subclass of C.

In effect, the generalized scrutability thesis says that there is a compact class of sentences that will serve as a scrutability base *however* the world turns out. The relevant compact class might be thought of as yielding a sort of scrutability base for the entire space of epistemically possible scenarios. Of course there are epistemically possible scenarios that involve all sorts of alien properties that are not present in the actual world. As a result, a scrutability base for the entire space of such scenarios may need to involve many more families of expressions than a scrutability base for the actual world. So for the purposes of generalized scrutability, the understanding of "compactness" might need to be weakened significantly, compared to the understanding that is required for actual-world scrutability. Perhaps we might even need to allow an infinite number of families in the base, for example if there are infinitely many sorts of fundamental alien properties. But the hope is that some interestingly limited class will suffice all the same.

In principle, one can formulate generalized versions of A Priori Scrutability, Conditional Scrutability, Empirical Scrutability, and various other scrutability theses. But as one might expect, it will be Generalized A Priori Scrutability that is most central for our purposes.

8   Postscript: Scrutability, Knowability, and Determinacy

A well-known thesis in philosophy is the following.

> *Knowability Thesis*: For any truth S, it is possible to know S.

This thesis is often doubted, for both intuitive and formal reasons. Intuitively, it seems that there may be truths concerning the distant past, the far away, and the very small, that it may be

impossible for anyone to know. Formally, the thesis above gives rise to what it often known as the Paradox of Knowability, first published by Fitch (19xx).

Fitch in effect gives a disproof of the Knowability Thesis, arguing from the weak assumption that some truth is not *known* to the conclusion that some truth is not knowable. Let P be a truth such that in the actual course of history, no-one ever knows P. Let Q be 'P and no-one knows that P'. Then Q is true, but Q is unknowable. If someone were to know Q, then they would know P, but if someone were to know P, then Q would be false. So no-one can know Q.

The scrutability thesis is closely related to the knowability thesis. It does not say that every truth is knowable, but it does say that every truth is scrutable, or derivable from a limited class of basic truths. One might thereby wonder whether scrutability theses are liable to similar problems. We have seen above that some versions of the Empirical Scrutability thesis are liable to problems akin to Fitch's. However, Conditional and A Priori Scrutability avoid both problems.

Concerning the intuitive problem: the base truths C may well include relevant truths about the distant past, including perhaps the spatiotemporal configuration of physical particles then, and so on. Even when S is an intuitively unknowable truth about the distant past, there is no corresponding intuitive problem with the idea that one can know that *if* the sentences in C are true, then S is true. Likewise there is no corresponding intuitive problem with the idea that one can know a priori a material conditional from C to S. Something similar applies to truths about the far away and the very small. So here is no intuitive objection to the scrutability thesis here.

As for the paradox of knowability: even though Q above is unknowable, there is no formal problem with the claim that one can know that *if* the sentences in C are true, then Q is true. Indeed, as long as P itself and claims about knowledge of P are both scrutable from C, then 'P and no-one knows that P' will be straightforwardly scrutable from C. This goes for both A Priori and Conditional Scrutability.

One might suggest that the Scrutability Thesis entails the Knowability Thesis, if we granted that D is itself knowable. By knowing D (empirically) and by knowing $D \supset Q$ (a priori), one could thereby come to know Q. However, there is no reason to believe that D is itself knowable. In fact, there is good reason to believe that it is not, both for intuitive and Fitch-style reasons. The intuitive reasons are obvious: D may involve information about the distant past and the far away that no-one will ever know. As for the Fitch-style reasons: assuming that no-one in the actual history of the world believes D, then D specifies a world in which no-one believes D. If someone came to believe D, they would live in a world quite different from ours, one in which their belief would be false. So no-one can know D.

One might think that one can define a factive operator "scrut", such that one scruts P iff one derives P from base truths. One might then try to generate a Fitchian paradox, by taking P to be any truth that one does not actually scrut, and taking Q to be 'P and I do not scrut that P'. By Fitch's reasoning, if scruting is factive, then Q is an inscrutable truth. However: the notion of scruting above is ambiguous. If to scrut P is to derive P from C, where C are the base truths of the actual world (or of any specific world) then scruting is not factive: there will be worlds in which P is scruted but false. If to scrut P is to derive P from the base truths of the world one is in at the time of scruting, then 'P and I do not scrut that P' is indeed inscrutable. But this does not yield a counterexample to the A Priori or Conditional Scrutability theses above, as these these require only that truths be derivable from the base truths of the worlds in which they are true, not the worlds in which they are so derived.

It may be that scrutability theses can do some of the work that knowability theses have been intended to do, or that they capture some of the intuitions that have led theorists to express sympathy with the knowability thesis. For example, Edgington (19xx) suggests that the intuition is that if P is true in the actual world, then it is possible that one can come to know, in some different world, that P is true in the actual world. Of course this raises questions about what it is to know in a different world that P is true in the actual world. One suggestion is that to do this requires specifying the actual world with a canonical sentence D, and coming to know that if D were the case, P would be the case. Transposing this counterfactual claim into an epistemic mode (if D is the case, then P is the case), the resulting claim is not too far from the conditional scrutability thesis.

Another problem for the knowability thesis concerns cases of indeterminacy. (This problem is raised by Hawthorne (2005) for the case of omniscient knowers, but the problem generalizes.) On most views of vagueness, there are some statements S that are neither determinately true nor determinately false. Such statements are usually said to be indeterminate. On some such views, the statement $S \lor \neg S$ will be true all the same. If so, one could reason disjunctively: if S, then S is true, so S is knowable; if $\neg S$, then $\neg S$ is true, so $\neg S$ is knowable. So either S is knowable or $\neg S$ is knowable. But if S is indeterminate, this conclusion is implausible. One can raise a parallel problem for the scrutability thesis, yielding the conclusion that for all S, either S is scrutable (from a relevant D) or $\neg S$ is scrutable. Once again, this conclusion is implausible when S is indeterminate.

One could resist this conclusion by rejecting the law of the excluded middle and refusing to accept that $S \lor \neg S$ is true when S is indeterminate, or by holding that when S is indeterminate, it is

likewise indeterminate whether S is scrutable. But perhaps the most straightforward way to avoid the problem is to understand the scrutability thesis as applying to *determinate* truths. That is, the thesis will say that when S is *determinately* true, or when det(S) is true, then D implies S. On the relevant sort of view, the disjunction $det(S) \vee det(\neg S)$ will not be true in cases of indeterminacy, so the problem here will be avoided.

One might worry about cases of higher-order indeterminacy, where it is indeterminate whether $det(S)$ or $\neg det(S)$. In such a case, the best thing to say is that it is indeterminate whether S is scrutable from D. Given the presence of vagueness in language, one should expect that scrutability can be vague too. On this view, implication by base truths goes along with determinacy, and vagueness of implication goes along with vagueness of determinacy.[13]

One can extend the scrutability thesis to the thesis that for all S, the truth-value of S is scrutable from D, whatever this truth-value may be. Here, one could simply apply the thesis to the statement 'S has truth-value T', or better, one could apply the thesis to a statement such as '$\neg S$', 'indet(S)', and other statements which are true iff S has a relevant truth-value. As in the cases above, then if one adopts the view of indeterminacy outlined above, these statements will be scrutable only when they are determinately true. So, for example, the claim will be that if indet(S) is determinately true, then D implies indet(S).

A final worry related to these matters arises from cases analogous to the Liar paradox. Say that S is 'This sentence is not scrutable from D'. Then if S is true, it is inscrutable, and if S is false, it is scrutable. Either way we have a counterexample to the thesis that a sentence is true if and only if it is scrutable.

This worry is an instance of a general worry for any thesis holding that a sentence is true iff it has property ϕ. Whether the sentence 'This sentence does not have ϕ' is true or false, it generates a counterexample to the thesis. I do not think it is reasonable to infer that no such thesis can be true, however, especially in light of the existence of the Liar Paradox. Instead, it seems best to say that sentences like this should be handled by whatever mechanism best handles the Liar Paradox. Indeed, one might take it to be a constraint on solutions to the Liar Paradox that they should also apply to sentences like this.

---

[13]There is an alternative view on which the vagueness of implication goes along with the vagueness of truth rather than the vagueness of indeterminacy: S is indeterminate iff it is indeterminate whether S is scrutable. If we accept the law of the excluded middle, this view will most naturally be combined with a view on which it is always the case that either S is scrutable or $\neg S$ is scrutable (cases apparently in the middle will be borderline cases of each).

The most obvious thing to say is that in cases like this, 'S does not have ϕ' is indeterminate. Given the discussion above, 'This sentence is not scrutable from D' is slightly more complicated, as the relevant thesis says that a sentence is *determinately* true iff it is scrutable. This renders the sentence at issue more closely analogous to the Strengthened Liar, 'This sentence is not determinately true'. So a proponent of the Scrutability Thesis should say that the sentence has the same truth-value of the Strengthed Liar, whatever that truth-value is (perhaps involving some sort of higher-order indeterminacy). Saying more requires an adequate treatment of Liar paradoxes in general, but that is a problem for everyone, and not for the scrutability thesis in particular.