The Paradox of Self-Consultation and A Theory of Epistemic Work

Bert Baumgaertner
University of Idaho
bbbaum@uidaho.edu

Bernard Molyneux
(formerly) UC Davis

January 16, 2025

Abstract

We introduce what we call the paradox of self consultation: This is the question of how apriori inquirers, like philosophers, mathematicians, and linguists, are able to (successfully) investigate matters of which they are initially ignorant by systematically questioning themselves. A related phenomenon is multiple grades of access: We find it extremely hard to think up analyses of our concepts that do not suffer from counterexamples; moderately hard to think up counterexamples to proposed analyses; and trivial to verify that a provided counterexample is genuine. We consider a range of potential explanations, including two-system approaches, and show why they are unsatisfactory, despite being on the right track. We then proceed to give a naturalistic solution to the paradox and multiple grades of access. In doing so, we present a novel theory of epistemic work, which we connect to formal learning theory.

Keywords: Intuition; Paradox of Self-Consultation; Paradox of Analysis; Philosophical Cognition; Complexity; Formal Learning Theory

1 Introduction

In this paper we will present a first theory of epistemic work. Our theory was initially motivated by the need to answer a couple of interrelated questions, but the endeavor of addressing what seemed like a small, relatively self-contained cluster of issues has expanded beyond its initial frontier. To begin with, our questions were as follows:

1. Why is it that, when we come to inquire after certain questions—e.g. the question of what is knowledge, or what is justice—we proceed by interrogating *ourselves*? If we already know the answer, then there is nothing further to discover. And if we do not, then by asking ourselves, we are asking the wrong people.

2. When we come to answer these questions, we generally engage in three kinds of epistemic activity: There is the primary activity of coming up with a theory or analysis, the secondary activity of finding counterexamples to such theories, and the tertiary activity of checking that a putative counterexample is genuine. But why is it that (as we shall see) the first activity is so much more difficult than the second, and the second more difficult than the third? If we know our target–if, e.g. we know what knowledge is–then shouldn't all three tasks be equally easy? And if we don't, shouldn't they be equally impossible?

We are not the first to ask the first kind of questions. For example, Kirk Ludwig writes: "How can one be expected to give a definition of a word unless one understands it, and if one understands it, must it not be that one already knows its definition?"[Lud07, p.131] These kinds of questions are not unrelated to the ancient paradox expressed in *The Meno 80d*:

> "And how will you enquire, Socrates, into that which you do not know? What will you put forth as the subject of enquiry? And if you find what you want, how will you ever know that this is the thing which you did not know?"

Here, Socrates is asked how he could be expected to recognize the answer he seeks, if he does not know the answer already.

Yet, despite this argument, there *is* something to be learned from armchair reflection, not just in philosophy, but in mathematics, linguistics, and to some extent all areas of inquiry. Progress has historically been made on various matters by asking oneself the right questions. The question of how this is possible is what we call "the paradox of self-consultation".

In rough preview, our answer to question (1), above, is that progress may be made through self-consultation because, though we need nothing outside of ourselves to access the information in question, we nevertheless access it only by doing some epistemic work. It is, as it were, *internally* available, but not *freely* available.

The rough preview of how we answer the second kind of question is that there is more work required of some epistemic tasks than there is of others. Specifically, this is because some problems are more complex than others. In a little more detail, the issue is that the hardest task (of coming up with a theory or analysis) involves some number of iterations of the next hardest task (of finding a counterexample), and accomplishing the latter, in turn, requires us to complete some number of iterations of the easiest task (that of verifying a counterexample). Thus the harder tasks are harder because they contain the easier tasks as component parts, and that means that the steps required of the harder tasks is some multiple of the steps required of the easier.

Before we offer a fuller characterization of this notion of epistemic work, two points are worth making. First, by 'epistemic work' we do not mean *reasoning* in anything like the traditional deductive or inductive sense. The processes we have in mind that fall under epistemic work are more akin to the work done in a posteriori inquiry, where one has to make a great number of observations to be confident a hypothesis holds universally, than they are to a priori inference and proof.

This brings us to a second point, which is that our account of epistemic work is surprisingly general. As may be expected, it is not just philosophical inquiries that engage in the kind of epistemic work we give an account of, but also others that traditional fall under the a priori category, such as mathematics, linguistics, and perhaps even some areas of artificial intelligence. But in addition to that, the epistemic processes involved in self-consultation exhibit the same structural features, at the appropriate level of abstraction, as those pertaining to a posteriori inquiry, as explicated within the rich literature on formal learning theory.

2 The Paradox of Self-Consultation

2.1 A Version of the Paradox

In what follows, it will be helpful to work with a concrete example of an inquiry in which progress is made through self-consultation. We choose the task of analyzing knowledge (henceforth, we refer to this analysandum as 'K').[1] One of the most interesting episodes in the history of this project was the publication, in 1963, of Gettier's paper "Is justified true belief knowledge?"[Get63]. Until then, the orthodox account, according to which K was justified true belief, had had a relatively peaceful reign from its first formulation in Plato's *Meno*, through its refinement in the twentieth century by C.I. Lewis [Lew46]. Gettier's paper is famous for raising two counterexamples to this orthodox account. They are familiar enough that we won't repeat them here. What's important is that, in the wake of them, support for the orthodox view crumbled, suggested replacements quickly fell victim to their own counterexamples and no analysis of K was able to gather a significant consensus.

Now suppose that some epistemologist, S, reads Gettier's famous paper for the first time and instantly recognizes the power of his counterexamples. Years later, having seen many failed attempts to meet Gettier's challenge, S comes to address the question herself, and proposes her own analysis of K. Months after publication, alas, her suggestion is shot down, felled by a counterexample she had not considered. S claps her hands to her head in frustration. How did she miss *that*? She is in no doubt that the counterexample is genuine. It clearly meets her analysis, yet is just as clearly *not* a case of K.

This case exhibits the prima facie paradoxical features. On the one hand, if S didn't know what was required for K, how was she able to recognize that the surprise counterexample was genuine? On

[1] To successfully analyze *knowledge* we may have to have *knowledge* of the concept, its application conditions, etc. To avoid confusion, we refer to knowledge in its former case—i.e. considered as a somewhat arbitrary analysandum—as 'K', and knowledge in the latter case—i.e. considered as something that facilitates analysis—as 'knowledge'.

the other hand, if she has access to the correct conditions, then why did she not only fail to provide the correct analysis, but publish *in good faith* an account that was incorrect?

Now suppose that you are the philosopher who came up with the counterexample. Let's suppose that you are a bit of a counterexample whiz who often comes up with clever cases like these. How are you doing it? It seems you simply *must* have intellectual access to the correct analysis of K! Otherwise, it is a complete mystery that you are able to reliably formulate counterexamples to other people's accounts. But if you have such access, why settle with shooting down other analyses? Why not provide the correct analysis yourself? Why is that so much more difficult?

2.2 Multiple Grades of Access

Perhaps we can answer this riddle quite easily, by observing that there multiple tasks here, and the epistemologist is not equally good at all of them. To begin with, we have the task of providing an account of K, one that suffers no counterexamples. We can safely surmise that this is the hardest task, since epistemologists have not succeeded in producing a consensus answer. In contrast are two tasks that epistemologists have done quite well at—namely showing (usually via counterexample) that one or another proposed analysis of K is incorrect, and recognizing such refutations when provided by others. Since these are different tasks, there is no paradox in being better at some than at others.

The problem, however, is that it seems that skill at these three tasks ought to jointly co-vary. If epistemologists have access to the conditions required for K, they ought not only to be able to provide counterexamples to wrong accounts, and recognize those counterexamples that are genuine, but also to produce an account of that concept for themselves. Because they cannot—and, worse, because they repeatedly produce the *wrong* account—we have reason to conclude that they do not have access. But in that case how are they able to produce counterexamples to other people's accounts?

So we may observe that if, indeed, there are three tasks that are not equally hard, then the problem is to explain how that can be so. How is it that we experience *multiple grades of difficulty*?

> [Multiple Grades] Inquirers find it very hard to think up analyses of our concepts that are not subject to refutation and moderately hard to think up refutations to proposed analyses, but (relatively) easy to recognize the validity of a genuine refutation.

The *prima facie* puzzle is that, if we have access to the criteria for (e.g.) K then we ought to be able to do all of these tasks quite easily whereas, if we don't, we ought not to be able to do any of them. The fact that some are harder than others requires explanation.

Obviously, if we are to explain *Multiple Grades*, we must say why there are (at least) *three* levels of difficulty associated with conceptual analysis. As we shall see in the next section, however, many otherwise appealing ways to approach the issue only predict two. Such accounts fail to meet the explanatory demand inherent in *Multiple Grades*.

3 Putative Explanations

3.1 The Ubiquity Response

One might meet the phenomenon above with a shrug, giving something like the following answer:

> There are many things that we have knowledge of that we cannot specify in very much detail. For example, we all know the faces of our loved ones, and are able to recognize familiar places, but few of us could draw a convincing picture of either of these from memory.

The problem with this response, however, is that pointing out that a phenomenon is widespread, and hence unsurprising, is not enough to adequately explain it. The question remains: *How* are we to sort cases if we don't have access to the right criteria? And *why*, if we do, are we not able to provide them?

3.2 Recognitional vs Criterial Concepts

Let's linger, nevertheless, on the distinction between recognizing a face and drawing it from memory. Could something similar explain the distinction in analyzing K? Could we make a similar distinction between *recognitional knowledge*, which permits us to categorize cases depending on whether they belong to the concept, and *criterial knowledge* which, rather like providing a drawing, requires us to specify the contours of the notion in question?

Such a distinction appears plausible and useful. Consider Supreme Court Justice Potter Stewart's observation that, though he could not provide a concrete definition, he knew pornography when he saw it. We can describe this as a case where the judge had recognitional, but not criterial, knowledge of pornography. Perhaps we could say, similarly, that epistemologists have recognitional knowledge, but not criterial knowledge, of K. The former is what allows them to recognize that (e.g.) the Gettier cases are not cases of K. The latter is what (if only they had it) *would* permit them to reduce K to its necessary and sufficient conditions.

The utility of this distinction, however, is limited. One problem is that it does nothing more than give the devil a name. 'Recognitional knowledge' is the kind of knowledge that allows us to instantly categorize a case, but which we struggle to express, 'Criterial knowledge' is the other kind. This does little to help us understand how it is that there *could* be different kinds of knowledge, or in what the difference consists. The question now becomes: How is recognitional knowledge possible in the absence of criterial knowledge? How can we sort cases reliably when we don't have access to the criteria we need to do so?

3.3 Binary Approaches

Even if that question could be answered, a more general objection to the prior approach is that, as noted in section 2.2, armchair inquiry admits of *three* levels of difficulty; but the distinction between recognitional and criterial knowledge only divides our epistemic access into two. So, unless it is further supplemented, it looks poorly equipped to fully explain the phenomenon.

This problem is quite general, and a number of initially plausible responses to the problem can be seen to fall into the same trap. Consider any explanation of the grades of difficulty that depends on a binary distinction between:

- Conscious versus unconscious information
- Explicit versus implicit (or 'unarticulated') knowledge
- Knowing *how* (to use a concept) vs knowing *that* it applies
- First vs second order knowledge, *e.g.* Knowing that p vs knowing *that* (or how) we know that p
- Knowing a proposition versus knowing how to express it
- Knowledge corresponding to the difference in 'Fast' versus 'Slow' thinking.

It is not that we think these approaches are entirely on the wrong track. On the contrary, we think that some of these approaches may point towards *parts* of the solution. Without further supplementation, however, these approaches are all insufficient, since they only provide us with the means to distinguish two grades of difficulty, where we are looking to explain three.

3.4 Three-Systems Accounts

The last of the suggestions listed above is associated with the work of Daniel Kahneman [Kah11], and has attracted much attention of late. According to this approach, human cognition makes use of two systems, a fast and automatic but error prone cognitive 'System 1' and a slow and deliberate but more reliable 'System 2'. It has been suggested that this approach could help illuminate some aspects of philosophical cognition (e.g. see [Nag12] and [Cap14]). To extend it to explain Multiple Grades, can we simply add a third system, 'System 3'?

Alas, no extension of Kahneman's specific approach looks promising. To see why not, note that any mere extension would, presumably, retain the fast but inaccurate System 1 and the slow but

accurate System 2. The new System 3 would presumably fall somewhere on the same continuum, being either faster and less accurate than 1, slower but more accurate than 2, or somewhere in between. However, it seems that with respect to conceptual analysis, speed and accuracy do not trade off the way a Kahneman-style account would predict. On the contrary, when it comes to formulating correct analyses of philosophical concepts like K we are both slow *and* prone to error. We have spent a couple of thousand years—with several recent decades of particularly intense work—trying to come up with a satisfactory analysis of knowledge, generating many erroneous proposals along the way. In contrast, when it comes to (successfully) refuting these incorrect analyses, we are both quicker *and*, by all appearances, *less* error prone. Indeed, the mere *recognition* of a correct counterexample is, by and large, something that occurs with extraordinary quickness, taking little more than the time required to properly comprehend the case. So in the cases we are concerned with, higher speed correlates with higher accuracy, which is the exact opposite of what we would predict from an extension of the Kahneman account.

Still, we might try to solve the problem of Multiple Grades by hypothesizing a 'three systems' account, albeit one that isn't an extension of Kahneman's. Instead, we would have to posit a powerful 'System A' that is both fast and accurate, an intermediate "System B" that is slower and less accurate and a weak 'System C' that is very slow and quite error-prone.

On its own, however, this suggestion barely does more than give the three devils names. Where we once had three different kinds of difficulty that needed explaining, we now have three systems that each have the right competence. Not surprisingly, given how little has been done, we can simply rephrase the original puzzle within such a framework, to wit: How is the fast and accurate System A able to quickly and reliably validate proposed counterexamples, if not by accessing internally encoded conditions? But if it has access to the conditions for the concept, then why can't it use that access to quickly and reliably perform the other two tasks?

3.5 Access Requires Work

We might distinguish between the sense in which inquirers have access to the relevant information and the sense in which they do not by appealing to the work involved. For example, it takes little work to determine that $2 \times 2 = 4$ but somewhat more work to check whether $48 \times 32 = 1536$. Perhaps the reason why some tasks associated with conceptual analysis are harder than others is that, in much the same way, some tasks require more work than others. And of course, there are indefinitely many gradations of work, so we could in principle explain however many levels of difficulty there may turn out to be. We could say that it takes a lot of work to find the right analysis, a medium amount of work to find a counterexample to a given analysis, and a trivial amount of work to check that a given counterexample is genuine.

We think that this response is on the right lines. Ultimately, we will suggest that the difference in difficulty between the three kinds of task consists in a difference in the number of algorithmic steps involved. However, it is important to observe a problem at the outset, namely that, where this may be obvious in the case of long multiplication, none of the three tasks involved in conceptual analysis presents itself as involving a series of explicit inferential steps. On the contrary, the task of checking a counterexample has the phenomenology of an instant judgment whereas the task of finding one, and of finding a theory without one, both appear to involve a mixture of creativity and intuitively guided trial and error. So if, despite this, they involve a number of algorithmic steps, then we will have to say something about why these steps are less than consciously obvious.

A second problem is that, though one instance of long division may be harder than another, all instances are of the same type. The task of multiplying 243 by 325 is the same kind of task as that of multiplying 234234 by 10909809. The difference is purely one of degree, rather than kind. The three tasks involved in conceptual analysis, in contrast, appear to be different kinds of epistemic task.

Lastly, it seems odd, if we are to treat these problems as similar, that the problems of analysis can be broken into three distinct categories, whereas long division problems cannot. On the contrary, for any long division problem, e.g. that of multiplying two four digit numbers, there is a harder one, e.g. that of multiplying five. *Prima facie*, this is what one should expect if the difficulty of the problem reduces to the number of steps involved. *Prima facie* we should not expect a trichotomous breakdown of the sort we have with conceptual analysis.

Similar considerations apply if we compare the different levels of difficulty involved in the three analytical tasks with the different degrees of difficulty involved in reasoning through different proofs.

For instance, consider Euclid's proof about the sum of a triangle's internal angles, which takes but a few lines in "The Elements" and can easily be grasped by school children. Contrast it with Andrew Wiles' extensive, hundred-page proof of Fermat's Last Theorem, which is challenging even to those with years of advanced study. Despite these differences in length and complexity, these proofs share a fundamental similarity in their nature. They both proceed, step by step, via a series of mathematically sound inferences, though Wiles' proof involves many more inferences than Euclid's.

If this perspective were to apply to distinguishing the analytical tasks, then developing an analysis of 'K' would require navigating a more extensive series of inferences compared to simply identifying a counterexample, which would require more inferential steps than just recognizing a counterexample as genuine. This would then explain why each task is more challenging than the next.

Differences in proof difficulty seems like the wrong comparison. When Euclid sought a proof, he embarked on the same kind of task that Wiles did when he sought a proof. When others came to check Euclid's proof, they embarked on the same kind of task that others did when, two millennia later, they attempted to check the reasoning of Wiles. Yes, there was more work involved in Wiles' proof than in Euclid's, but these were nevertheless tasks of the same sort.

Moreover, analyzing 'K' seems different. One instantly recognizes a counterexample as genuine, without any inference seeming to be involved. And where the process of developing a counterexample, or an analysis without one, seem to involve some kind of cognitive work, it seems to involve hazy intuitive reflection, creative experimentation, adjustments, improvements, and repeated iterations, rather than following a linear sequence of inferential steps.

A better observation is that it took more work to come up with a proof of Fermat's Last Theorem than it took to check the proof, once it was provided. Analogously it takes more work to come up with a theory of K than it does to verify it (using counterexamples), and more work to formulate a counterexample than to verify that counterexample is genuine. Thus we can observe a general distinction between the work involved in formulation vs verification, which seems like it might explain the differences involved in the various tasks of analysis.

Though we find this promising, we must nevertheless observe certain dissimilarities which might prove important. For unlike the case where a proof, neither checking a counterexample, finding a counterexample, nor finding a theory without counterexamples presents itself as involving a series of explicit inferential steps. On the contrary, the first has the phenomenology of an instant judgment whereas the latter two involve a mixture of creativity and intuitively guided trial and error.

There may be some minimal reasoning involved at certain points, of course, of the sort that goes "the analysis classifies such and such as a case, but it is not, therefore the analysis is wrong", or "this analysis has withstood all counterexamples thus far, therefore we have good reason to provisionally accept it". But note that there is no sign here of any great differences in the length of reasoning involved, so this does not promise to explain why some tasks are much harder than others. Moreover, this kind of reasoning only makes its appearance after a counterexample has been found and judged to be genuine, or after an analysis has been found and discovered to be robust. So this seems like reasoning that occurs after the tasks we are interested in, and therefore not the kind that could explain the intrinsic difficulty of the tasks that precede it.

Furthermore, the mere suggestion that different tasks *could* in principle involve different workloads is not very good reason to expect that they *will*. A full account, in contrast, would not only lead us to expect such differences, but would allow us to predict, on independent grounds, which task will be harder than which. It should explain, not just *that* there are multiple grades of difficulty, but why the tasks are ordered in difficulty the way that they are.

Lastly, how should we measure the difficulty involved? If these tasks turn out, at some appropriate level of description, to be algorithmic, then we could simply count steps. In that case it might turn out that it takes but a few steps for the easiest task, tens of thousands for the second, and zillions for the third. However, though that would explain the difference in time taken for each, it would do so at the expense of hypothesizing a structure with too many gradations of difficulty. On the face of it, such a model should lead us to expect a near-continuous spectrum of difficulty corresponding to the precise number of steps involved, rather than a coarse three way division. What if, instead, we proposed that the easiest task involves one step, the next easiest two, and the hardest three? Then that would posit a structure whose gradations of difficulty map tidily onto the phenomenon we are trying to explain, but at the expense of looking grossly unlikely. It would either (wrongly) predict that the hardest task takes triple the time taken for the easiest—three near-instants instead of one—or hypothesize that

some steps take longer than others, with the second and third steps taking (much, much, much) longer than the first. But that would just push the question back from why certain *tasks* are more difficult than others to why certain *steps* are more difficult than others.[2]

In brief, there are two shortfalls to the "access requires work" suggestion, even if it is (as we believe) on the right track. First, it suggests a spectrum of difficulty rather than a (three) tiered system. Second, it says nothing about why the tasks should be ordered in terms of difficulty the way they are. To address these shortfalls, we suggest that we need a more substantial account of what is meant by 'more work'.

4 An Account of Work

4.1 Binary distinction as a starting point

In the remainder, we will not attempt to explain how it is that humans have a conceptual ability to rapidly decide whether a putative counterexample to an analysis is genuine, but to explain why it is that, given this ability, humans are much slower at providing counterexamples to bad analyses and much slower still to provide the correct analysis.

We therefore begin by observing a binary distinction between what we can do very reliably and rapidly—namely certify genuine counterexamples—and what we cannot do with such ease, *i.e.* everything else. For the reasons we gave in section 3, no binary distinction of this sort can, on its own, be entirely adequate. But as an initial step towards a fuller account, the binary distinction is legitimate.

As we saw earlier, there are many ways in which this binary distinction might be undergirded. It is plausible, for example, that the brain, like an artificial neural net, encodes its dispositions via a myriad of reinforced connections with nothing resembling an explicit, readable memory where an exact, precise rule is stored. In fact, it may be that there is no explicit rule to look up. Our brains might follow the rule by dint of a brute causal process, the way that the planets follow the rules of physics. An account like this, according to which the rules are "implicit" or "unarticulated", has the potential to explain how we can sort cases without being able to introspect the classificatory rules we use to do so. Quite simply, there would be no explicit rules to introspect.

An alternative possibility is that there *are* explicit internal rules that are accessed by some sub-personal cognitive system, but they are out of the conscious, introspective reach of the agent as a whole. Such a state of affairs might be unsurprising from an evolutionary perspective. After all, as long as we subconsciously accessed these (inner, explicit) rules in a way that was good enough for the goals of communication, clear reasoning and whatever else might be relevant to our survival and success, there was no pressing evolutionary need for us to access to what was going on "under the hood".

We intend to remain neutral on what undergirds the binary distinction. It is enough for our purposes that, when it comes to explaining why we can't directly access the rules that govern our use of concepts, we have a number of plausible cognitive scientific hypotheses. We leave the question of which of them is correct to empirical science. We focus instead on the question of why, given only this binary distinction, we should predict there are three tiers of difficulty associated with philosophical analysis.

Foreshadowing to the section after our account of work, the solution to the paradox of self-consultation is identifying the mistake made by the riddle setup: just because you can query from the outside, doesn't mean you know how things get done on the inside.

4.2 Work even with a Black box classifier

In effect, we can treat the conceptual faculty as a black box that somehow encodes a pre-theoretical intuitive understanding of such concepts as K. For the reasons discussed in the previous section, we needn't commit to any hypotheses about how the black box works. Because the agent has no access to the criteria the box uses to come to a decision, all she can do, if she wishes to discover those criteria and make them explicit, is repeatedly provide the box with cases, record its decisions, and try to infer the criteria it is using from the judgments that it makes.

[2] And if we now break the steps themselves down into sub-steps, we veer back towards the earlier question of why these fine grained (sub)steps don't correspond to finer gradations of difficulty.

Thus we can think of the box as taking a single input—a case that needs sorting—and that the box returns the answer YES only if the case matches the (perhaps implicit and unarticulated) classification criteria the box uses. So if, for example, the box represents the concept K, then it returns YES only if the case is a genuine case of knowledge according to its hidden criteria. It returns NO only if it is not.[3]

Now consider the first kind of task, where the agent is given a theory and a putative counterexample. In that case, she need only decide how the theory classifies the case, test the case against the black box and compare the two decisions. If the theory differs from the box she can infer, at least, that the theory is not a match for the intuitive criteria used by humans.

How much work did this task involve? It will be a sum of the work done by the box to produce an output and the work done to compare that with the predictions made by the theory. For our purposes, we can just use the variable $w_{c,t}$ to stand for the total work involved to check whether a given case c falsifies a given theory t.

Now suppose our agent has a theory and wants to know whether it has a counterexample—*i.e* a case upon which the theory and the box disagree. Then, since she has no direct access to the inner criteria used by the box, the only way to proceed is to go through cases one by one. The time taken to find a counterexample, in that case, will depend greatly on which cases she chooses to test first. At one extreme of the spectrum, she might try out cases at random, but that will mean finding a counterexample, even if one is there, is likely to take a long time. A more efficient procedure, if she can find it, is to try to intelligently predict where the theory and the box are likely to disagree and try out such cases first. Either way, the agent repeatedly checks the classifications made by the theory against the classifications provided by the box. There is no other way to proceed.

How much work does this second task take? The workload involved in the second sort of task is equal to the sum over the workloads involved in the first. Specifically, where it takes n attempts to find a genuine counterexample, the work $w_t(x)$ required is equal to the sum over the work $w_{i,t}$ required to check each case up until the nth. Symbolically:

$$w_t(x) = \sum_{i=1}^{n} w_{i,t}$$

Just as a prospector could strike gold with the first swing of the pick axe, the first case the theorist examines might turn out to be the counterexample she seeks. In such a limiting case, the workload involved in the two tasks would be equal. But in all other cases, the second task requires more work—potentially *much* more work—and could even (for reasons we momentarily defer) be infinite.

The third task is to find a theory without counterexamples, so the theory is not given. Consider how our agent would go about completing this task with the black box. She must try theories, one after another, to find one that is robust against counterexamples. That means repeating the second task—the attempt to test the theory against cases—for each theory considered. So, just as the second task required the repeated application of the first, the third requires the repeated application of the second.

Again, there may be more efficient methods for structuring the search that try out more probable theories first. But such variations in efficiency only correspond to differences in the order in which the second task is repeated. There is no way (other than through brute good luck) to avoid repeating it. Accordingly, the work required for the third task is equal to the sum over all the required iterations of the second task. With this in mind, suppose it takes m attempts to find a theory with no counterexamples, then we have:

$$w(x, y) = \sum_{i=1}^{m} w_i(x)$$

What matters, here, is that the third task requires us to iterate through the second task, which itself requires us to iterate through the first. The workload is therefore likely to be a (potentially huge) sum over the workloads required for tasks of the easier sort.

[3]We use "only if" rather than "if and only if" because there is the possibility that the box does not return an answer.

There are two kinds of insights we gain from our black box example. One is that conceptual analysis can be fruitfully compared with the application of the scientific method. An analysis can be regarded as a hypothesis about our classificatory dispositions, which we then test against those dispositions via repeated exposure to cases. This insight, we contend, provides a resolution of the paradox of self-consultation. The second insight is that the example lends itself to a logical characterization that allows us to connect with a robust literature in formal learning theory. We discuss these in more detail in turn.

4.3 The a priori as a self-applied scientific method

This process does not look very much like a paradigm a priori process. It does not involve deductive reasoning from axiomatic or self-evident propositions. In fact it hardly appears to involve very much reasoning at all.

In our opinion, the process much more closely approximates an a posteriori search through empirical data. On an idealized, classical model of scientific investigation[4], the scientist proceeds by looking at cases and attempting to falsify her hypothesis. So if, e.g., the hypothesis is that all ravens are black, the scientist proceeds by looking at raven after raven to try to find one that isn't. Framed in this way, we can see the same tripartite division in difficulty that we found with respect to philosophical theorizing. Given the hypothesis and a white raven, it is fairly trivial for the scientist to falsify the hypothesis. She merely needs to look at the raven and make a color observation. But suppose that she is given just the black ravens hypothesis and asked to find a falsifier. Then she must search through the domain of ravens to find one that is non-black, performing the easier task (the mere observing of color) for each case. And this, in turn, is a mere step in the hardest task, which is to find a hypothesis that is not falsified by the data. To perform this task, she must search through the domain of hypotheses and run the easier task (the attempt to find a falsifier) for each case she considers.

We can think of a proposed philosophical analysis of a concept as a hypothesis about the application conditions of the analysandum. The hypothesis is tested against (hypothetical) cases, to see whether it classifies cases in the same way as the concept. The three levels of difficulty are the ones we find in any search task, whether it be an external search through real world objects to find (e.g.) a raven that is not black or an internal search through imaginary cases to find (e.g.) a justified true belief that is not a case of knowledge. In both cases, the task is easiest when we are given the case and the hypothesis, since no search is required; it is harder when we are given the hypothesis but not the falsifying case, since we must then search through cases; and it is hardest when we are given no hypothesis, since we must first search through hypotheses and, for each, perform a search through the cases.

With this in mind, should philosophical analysis count as an a priori endeavor? It depends on how that notion is defined, but it seems to us to be more like a posteriori research that proceeds (in large part) through self-experimentation. In effect, we treat our concepts as black boxes which we reverse-engineer by making and testing hypotheses concerning their precise application conditions.

The a posteriori character of the investigation is perhaps more overt when we observe that it need not involve *self*-consultation at all. Work in experimental philosophy has proceeded by testing analyses against the intuitive judgments of a large number of subjects. This is not a deviation from traditional a priori philosophical methods if those methods were never a priori to begin with. On the contrary, it is simply a refinement of the a posteriori methods philosophers have always used, one that takes advantage of statistical methods in the social sciences.

All that said, whether self-consultation is a priori or a posteriori is a digression from our main argument. While we find it helpful to think of the a priori as a kind of self-applied scientific method, our main focus is on the issue of multiple grades of access. As becomes clearer in the next section, our solution ultimately rests on the nature of the tasks and their relation to each other, not how we classify the cognitive processes by which we complete them. Before we present our account of work in more detail, we turn back briefly to the paradox of self-consultation and our solution to it.

4.4 Back to the Paradox of Self Consultation

The heart of the paradox of self-consultation is about the concept of access. If all kinds of access are equal and we have access to the criteria for, e.g., knowledge, then we ought to be able to do three kinds

[4]See, for example [Pop14] or [Kel96]

of tasks with the same level of ease: recognize a genuine refutation to an analysis, think up refutations to proposed analyses, and think up analyses of concepts that are not subject to refutation. If we don't manage to do these tasks with the same level of ease, then, as the puzzle goes, we ought not to be able to do any of them.

Our solution, like other putative ones, denies the part of the antecedent that claims that all kinds of access are equal. But unlike these other putative explanations, ours acknowledges that there are at minimum three distinct levels of access. We characterize these levels in terms of a containment or dependency relation, one that can be represented symbolically with alternating quantifiers.

This dependency relation is the key to solving the paradox of self-consultation. When we considered a "three systems" account above in Section 3.4 as a putative solution, we were left with the question of why some queries would be directed to slower and less accurate modules, while others to the faster and more accurate modules. The dependency relation between the tasks explains why any way of characterizing modules, whether one or many, there is simply no way to sidestep the complexity of the tasks. We might have three different modules for each task, or we might have one module do all three, or something else. It doesn't matter what the arrangement is because it is the *containment of the tasks* that generates the increasing complexity. We need not repeat how work escalates across the tasks to explain how we get an increase in the time it takes.

In addition to denying that all kinds of access are equal, our solution presupposes that there is some form of a 'black box' that can be consulted. We need not posit what precisely this is - it may be a module in the brain, some cognitive mechanism that corresponds to our intuitions, a neural network for the concept in question, etc. Whatever it is, what is relevant to our solution is that it takes less work to use a concept than it takes to understand how it works. We can use our hand without knowing how it does what it does, and as we build hypotheses and theories about how it works, we can use it to test them. Something like that is true of our minds - we can use them without knowing the inner workings. However, as we pointed out in our considerations of binary approaches above, this distinction by itself is not sufficient. We needed an account of work that provided us with at least three levels, which we have done.

So the solution to the paradox of self consultation is like solutions to other paradoxes: somewhere along the way a multifaceted concept masks itself as one. In our case, the relevant concepts, depending on how the paradox is described, could be *understanding*, *knowledge*, or *access*. Whatever it is, when these concepts conceal the fact that self consultation requires work we can generate the paradox, but when we make explicit our account of work, we cannot. We understand/know/can access the definition of a word to judge that a counterexample is genuine without understanding/knowing/accessing the criteria because the former requires less work (and most likely *much* less work) than the latter.

5 Generalizing: Alternating Quantifiers and Formal Learning Theory

The second of two insights we gain from our black box classifier example is that it connects with a robust literature on formal learning theory. By making this connection, we make the case that our account of work is not a post-hoc solution to the paradox of self-consultation. Although we came upon our account by thinking about the paradox, the account proves to be quite general.

To connect with formal learning theory, we characterize our black box classifier example with some common logical vocabulary. All three tasks have something to do with checking a theory or analysis for counterexamples, but they differ in the logical way they relate to that central idea. Let's define a two-place predicate, Falsify(x,y) to be satisfied if and only if x is a genuine counterexample to theory y. Then the first task the philosopher faces is to take a theory t and putative counterexample c and check whether Falsify(c,t) is true. The second sort of task arises when the theory is provided, but not the example that purports to refute it, and the question is whether the theory has any successful counterexamples. This kind of task can be represented, symbolically, by replacing c with an existentially bound variable x. The task that results is to check the truth of $\exists$x Falsify(x,t). The third task is to find a theory that does not succumb to *any* counterexamples. This task equates to checking the truth of the sentence $\exists y \neg \exists x$ Falsify(x,y).

From a logical perspective, the truth of the sentence $\exists y \neg \exists x$ Falsify(x,y) is equivalent to $\exists y \forall x \neg$Falsify(x,y) by Demorgan's Law for Quantifiers. An alternation in types of quantifiers yields an increase in com-

plexity.[5] To see this, note that there is an order of dependency in the evaluation of the quantified formulas. Evaluating the outside existentially quantified formula cannot be done without evaluating the inner universally quantified formula, but evaluating the inner universally quantified formula can be done without the outer. Furthermore, evaluating the universally quantified formula depends in turn on evaluating the predicate formula (the negation simply flips whether the predicate has been satisfied or not), but not the other way around.

The complexity we are pointing to is not a mere artifact of logic. It is supported by the rich and robust field of formal learning theory. To see this, there are some important qualifications that our account of work takes on that also undergird existing insights. First, we do not consider the work added through inefficiency, lack of access to resources and unnecessary elaboration to be intrinsic to the task. The lower bound, in contrast, is what we get when all such extrinsic additions are stripped away. As such, we consider it a feature of the task itself, not a feature of those doing the task. So when we say that a task *requires* a certain amount of work to complete, we are talking about a lower bound that is intrinsic to the problem itself, not about the particulars of the agents who try to solve the problem.

This is not to say that the agent and the resources they have access to are irrelevant. It is possible, for example, to formulate problems in such a way as to place constraints upon what resources are permissible and how accessing those resources counts towards the measurement of work. If, for example, we are asking how much work is required to multiply 48 by 32 using an electronic calculator then the answer depends on whether we only consider the work required to punch in the query or whether we also count the untold millions of computational operations performed by the calculator itself. Put generally, the question is whether to count all the steps in an outsourced sub-procedure as contributing to the work involved in the task, or whether to merely count the minimal number of steps required to do the outsourcing. Relative to different interests and concerns, either answer might be appropriate.

This brings us to a second qualification. There are several senses in which the multiplication performed by the electronic calculator required only a negligible amount of work. It didn't take much time, it didn't take much energy, and we ourselves didn't have to do it. However, we are interested in a more abstract notion of work according to which the work performed is measured in the number of operations, or steps, that must be required to see it through. Though it seems to us to happen instantaneously, the multiplication performed by the calculator involved a huge number of operations, and in that sense the machine did quite a lot of work.

With these qualifications in mind, consider again the logical characterization of the tasks. The alternation of quantifiers represents a hierarchy of work between levels that is absolute relative to work done within a level. We can characterize the situation using two black boxes. The first is the black box we referred to earlier, which takes a hypothesis and a case as input and verifies whether the case falsifies the hypothesis. The second—which we will imagine for easy reference to be a red box—is upstream of the first and determines which case the black box will be given next.

Now imagine that, in picking cases to feed to the black box, the red box selects at random. Then in most cases the chances of it presenting a genuine counterexample would be very small. At the other extreme, if it has an extremely intelligent strategy for selecting likely counterexamples to try, it is much more likely to produce a counterexample within a tractable number of attempts. On a particularly good run, for example, the first case it tries might be a near-hit; the second case is an amendment of the first patched up to fix its problems; which leads to new issues, which are fixed on the third, and successful, attempt.

On the one hand, we could think of the black box as a static entity to which the cases are delivered via a dynamic "data stream" (in the terminology of [Kel96]). This invites us to think of the black box as occupying a fixed position, looking at each case as it floats by, with the red box organizing the data stream, determining which case is next to be floated down the river. On the other hand, we could think of the cases as forming a static "search space" through which it is the black box that moves. This is the image most favored by the classical AI engineer. On this equivalent way of looking at things, the red box is a navigator, directing the black box to where in the space it should look next. Since the navigator plays the principal role in how quickly the black box finds what it is looking for—and whether it finds it at all—the question of how the red box works is central to any problem in classical

[5]Note that the application of Demorgan's Law for Quantifiers didn't introduce the complexity, it was already there but merely expressed in different notation.

AI.

But the red box might be thought of more generally, not just as the placeholder for a search algorithm in an intelligent program, but also for whatever procedure is used in a natural science to determine which case is studied next. In some sciences experimenters may exercise a great degree of control over what they will examine next. In others, the scientist is forced to study cases as they arise naturally, *e.g.* consider the scientist who studies eclipses or some other astronomical phenomenon that cannot be made to happen upon demand. In either case, the red box is just a placeholder for whatever process decides the order in which cases are to be considered, whether or not it is within intelligent control.

We can use the same two-box framework to model an armchair theorist looking for counterexamples to an analysis. On the one hand, she has something that tells her when a case is a counterexample. This is her black box. But she does not query her black box at random, nor does she exhaustively try out every possibility. She has some (presumably) sophisticated but (presently) obscure process for focusing only on the promising cases. This is her red box. Like the one described above, the armchair theorist's red box might work by trying out a promising but flawed case first, and then scrutinizing progressive refinements. Insofar as the two boxes are mere placeholders for whatever complicated process the theorist uses to do this, the model says too little to be controversial. It nevertheless permits us to make some pertinent observations. Firstly, to the extent that it is possible to voluntarily organize the data stream (i.e. to the extent that we are not looking at eclipses or some other phenomenon whose schedule is beyond our control) we can make a task easier by intelligently managing the order of inspection. There are, nevertheless, in-principle limits to this. As we will see, no matter how intelligent the red box is, it cannot perform a more complex task in the hierarchy with greater speed than it performs a less complex task. In fact, if we restrict ourselves to idealized processes that do not inspect the same case twice, we can make a stronger claim: That no matter how quickly and intelligently red box α performs the more complex task, and no matter how stupidly (short of repeating inspections) red box β performs the less complex task, α can never complete its task more quickly than β completes its.

We can say that advances in the design of the red box correspond merely to *in-level* reductions in the difficulty of the task, to be distinguished from the *cross-level* differences in difficulty that pertain to tasks arranged in a complexity hierarchy. The former are in principle remediable. This is the domain in which intelligence can improve. The latter are not, as they are difficulties intrinsic to the task itself.

To see why, suppose the black box is asked whether a given case is a counterexample to a given hypothesis. The answer is yes or no, and it is decidable in one inspection. This is the only case scenario, so it might seem misleading, though accurate, to describe it as the worst case scenario. If we do so, however, we will begin to see the outline of a pattern.

The higher level task of finding a genuine counterexample can, in the best case scenario, be completed in one step. That happens in the limiting case where the red box provides the black box with a genuine counterexample on its first attempt. So the harder task, in its best case scenario, is completed in the same time-frame—namely one step—as the easier task is completed in its worst (only) case scenario.

What about the worst case scenario for the more complex task? If we disregard the possibility of inspecting the same case more than once, the worst case is one in which every instance in the space of potential cases is inspected before either a counterexample is found (on the last possible attempt) or the search terminates without success. So the worst case scenario for this more complex task involves one complete tour of the case space.

Compare this with the yet higher level task of finding a hypothesis with no counterexamples. In the best case scenario, the red box selects such a hypothesis on its first attempt. However, it still needs to find out that it has done so by checking that there are no counterexamples, and to do that it must ask the black box to check the hypothesis for each case in the case space. So at this yet higher level, the *best* case involves one complete tour of the case space, which is the *worst* case for the less complex task.

We have been primarily interested in the three levels of difficulty discussed in the opening to the paper, but the point generalizes to all higher levels. That's because the levels in the hierarchy of difficulty are marked by the concatenation of alternating quantifiers. To see this in more detail, note that:

1. If a statement of the form $(\forall x)\phi$ is true, its truth value can only be determined by checking every

instantiation of ϕ. This is the *worst case scenario*, in terms of time, for a universally quantified query.

2. If a statement of the form $(\forall x)\phi$ is false, its truth value can be determined by checking just one instantiation of ϕ, if we are lucky enough to check the counterexample first. This is the *best case scenario*, in terms of time, for a universally quantified query.

3. If a statement of the form $(\exists x)\phi$ is true, its truth value can be determined by checking just one instantiation of ϕ, if we are lucky enough to check the verifying example first. This is the *best case scenario*, in terms of time, for an existentially quantified query.

4. If a statement of the form $(\exists x)\phi$ is false, its truth value can only be determined by checking every instantiation of ϕ. This is the *worst case scenario* in terms of time, for an existentially quantified query.

As we observed, any two consecutive levels in the hierarchy will take one of the following two forms:

- The harder question is of the form $(\forall x)\boldsymbol{(\exists y)\phi}$, where the easier question is the embedded boldened expression.
- The harder question is of the form $(\exists x)\boldsymbol{(\forall y)\phi}$, likewise.

Since in the first of these the harder question is universally quantified, the best case scenario must be one where the embedded expression is false (situation 2 above). But since the embedded expression is existentially quantified, it is false only if ϕ is false for every instantiation, which is the *worst* case scenario (situation 4) for the easier question. In the second of these, the situation is reversed but the argument, *mutatis mutandis*, goes the same way.

An important point, here, is that the harder question is guaranteed to take more steps to answer no matter how intelligent the red box is in organizing the higher level task, and no matter how stupid it is in organizing the lower level task. For no matter how smart it is at the higher level, it cannot do better than the best possible case scenario. And no matter how terrible it (or some rival red box) is at organizing the lower level task, it cannot do worse than the worst case scenario. So (excluding systems that repeat the same step) no possible system completes the harder task more quickly than some system completes the easier. In this way the boundaries between levels in the hierarchy are absolute.

References

[Cap14] Herman Cappelen. X-phi without intuitions. In Anthony Robert Booth and Darrell P. Rowbottom, editors, *Intuitions*, pages 269–286. Oxford University Press, 2014.

[Get63] E.L. Gettier. Is justified true belief knowledge? *Analysis*, 23(6):121–123, 1963.

[Kah11] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[Kel96] Kevin T Kelly. *The logic of reliable inquiry*. OUP USA, 1996.

[Lew46] C.I. Lewis. *An Analysis of Knowledge and Valuation*. The Open Court Publishing Company, 1946.

[Lud07] Kirk Ludwig. The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy*, 31(1):128–159, 2007.

[Nag12] Jennifer Nagel. Intuitions and experiments: A defense of the case method in epistemology. *Philosophy and Phenomenological Research*, 85(3):495–527, 2012.

[Pop14] Karl Popper. *Conjectures and refutations: The growth of scientific knowledge*. routledge, 1963/2014.