Grokking Pain

S. Yablo

draft of June 2, 2000

I. First a puzzle about a priori knowledge; then some morals for the philosophy of language and mind. The puzzle involves a contradiction, or seeming contradiction, among three extremely plausible-looking claims.

First, I know a lot of things a priori. There are many sentences S such that I know a priori that S.

Second, for every sentence S that I understand, I know a priori that *S if and only if "S" is true*; and so in particular that *if S, then "S" is true*. Here the word "true" means: true in my idiolect.

Third, I rarely if ever know a priori that a sentence "S" is true, where "true" once again means: true in my idiolect. To know a priori that "S" is true, I would need to know enough a priori about "S"'s meaning to rule it out that "S" means something false. And my knowledge of what "S" means is a posteriori.

These constitute an inconsistent triad if we assume, as people normally do, that a priori knowledge is closed under obvious consequence relations. A priori knowledge of a conditional and of the conditional's antecedent suffice in principle a priori knowledge of the conditional's consequent

II. The above looks like a paradox if, but only if, we think that a priori knowledge is deductively closed – in particular closed under modus ponens. The solution is to notice that, on a plausible and familiar explication of a priority, it isn't.

A point often made about a priori knowledge is that one isn't expected to know that P on the basis of zero experience. After all, experience may be needed even to see what "P" means. For me to know a priori that P, the most that's required is that I know that P based on my appreciation of what "P" means. (Note that my appreciation of what other sentences mean plays no role here.) Thus the familiar slogan that a priori knowledge that P is knowledge in virtue of understanding "P."

Suppose all that is right; suppose that knowing a priori that P is knowing that P just in virtue of grasping the meaning of "P." Then it's not to be expected that a priori knowledge that A, and that if A then B, should suffice for a priori knowledge that B.

Why not? Knowing a priori that A, and that if A then B, means that I know the former just in virtue of understanding "A", and the latter just in virtue of understanding "If A then B". These pieces of knowledge put me in a position to infer that B – so what's the problem? The problem is that that my subsequent knowledge that B is not guaranteed to be in virtue of my understanding of "B." It would seem on the contrary that my knowledge that B is based on my knowledge both of what "A" means and of what "if A then B" means. It would have to be based solely on my knowledge of what "B" means for me to count as knowing a priori that B.

III. So what? If the proposed solution is right, then no theory which makes a priori knowledge

out to be closed under modus ponens can be correct. Lots of theories however would appear to do just that. I will focus on so-called two-dimensionalism, or what Chalmers calls modal rationalism, because it has been the subject of so much recent attention.

Modal rationalists say that it is a priori that P if and only if "P" is true in, or with respect to, all worlds-considered-as-actual. If we define the primary intension of "P" as the set of worlds w such that "P" is true in w-construed-as-actual, then it is a priori that P if and only if the primary intension of "P" is the set of all (centered) worlds. From this analysis it clearly follows that if the premises of a modus ponens argument are a priori, so is the conclusion. There is no way for a conditional and its antecedent to have universal primary intension without the conditional's consequent having universal primary intension too.

IV. You've now heard the main argument of this paper, more or less: modal rationalism miconstrues the extent of a priori knowledge, so modal rationalism is mistaken. If we know a priori that sisters are siblings, then we know a priori too what "sisters are siblings" means – enough anyway to know a priori that that meaning is true. This is a fussy little objection as it stands, but it points the way to some bigger things as we'll eventually see. For one thing it points the way to a clearer picture of the Chalmers/Jackson objection to physicalism than we have had so far.

A good place to start is by recalling a principal teaching of Kripke about evaluating sentences in counterfactual worlds. When we evaluate "P" with respect to a world considered as counterfactual, how the inhabitants of that world understand the sentence is irrelevant. If it was the inhabitants' understanding that mattered, then it would be true rather than false that "if "tail" had meant leg, horses would have had four tails." The proper method is to take our sentence, understood our way, and ask whether it is true of the given world considered as counterfactual "True of" here expresses a transworld relation between our sentence and their world.

That was a point about evaluation with respect to worlds taken as counterfactual. Although it's less widely appreciated, similar points apply to worlds taken as actual. The proper method is to take our sentence, understood our way, and ask whether it so understood is true in the given world considered as actual. Chalmers would seem to agree with this:

> A primary intension specifies what it takes for an entity in the actual world to qualify as the referent of the concept: these conditions of application will often build in no requirements about the presenece of the concept itself. In evaluating the referent at an actual-world candidate, we retain the concept from the real actual world [TCM, 8]
>
> a world with an artichoke at its center is precisely the sort of actual-world candidate that is endorsed by my thought "I am an artichoke," even if the artichoke is not thinking. In these cases we can retain the thought from the real actual world and simultaneously evaluate its truth-value in other actual-world candidates without any loss of coherence. [Note: Doing things this also avoids a problem raised for Fodor's theory by Block (1991) and Stalnaker (1991). The problem is that of what must be "held constant" between contexts (a token in the language of thought? the physical/functional structure of the thinker?). On my account, nothing needs to be held constant, as we always appeal to the concept

from the real world in evaluating the referent at a centered world.] [TCM, 422]

Just as we saw with Kripke, then, Chalmers's theory gives the wrong results if evaluation is construed as intra-world rather than cross-world. He does not want it to come out a priori that, say, there are sentences. But then he needs a world which taken as actual makes "there are sentences" false. Such a world obviously cannot contain sentences; so the token of "there are sentences" that gets evaluated with respect to that world had better be here, not there.

V. All of that is fair enough, I think. The problem is to see how it's to be reconciled with the modal rationalist's principal claims. These I take to be

(E) it is a priori that S iff for all worlds w, "S" is true in w-considered-as-actual

and its apparent consequence

(A) if it is not a priori that S, then there are worlds which construed as actual make "S" false.

(A) is a key premise in the rationalist's argument against physicalism: it is not a priori that that if physics, then pain. I want to explore a certain way of constructing counterexamples to (A), and then see if the method can be extended to encompass the case of interest. (The case where "S" is "if physics then pain.") The idea in a nutshell is that you will have counterexamples to (A) when your sentence contains grokking predicates. By a grokking predicate I mean one that identifies its object in part by aspects of our experience of it that don't purport to be representational.

For example: A certain kind of line drawing will be seen by anyone who looks at it as a human face. I don't mean that everyone will judge it to resemble a face or to represent a face, just that we cannot easily stop ourselves from "seeing a face in it" and forming associated judgments, e.g., the face looks cruel or alarmed or what have you. I will call line drawings like this "facical." Now consider a world w about which all I'm going to tell you is that it contains Figure One:

Is "facical" true of this figure in w considered-as-actual? The answer is yes, and it remains yes no matter what I go on to say about the observers in w: that they see nothing in the figure, that it looks to them like a battleship, or whatever. The reason as before is that we evaluate the figure with respect to our word "facical," understood as we understand it. Our dispositions to see faces in presented figures figure crucially in that understanding, so they are part of what we (imaginatively) bring to bear on the figure in w.

Suppose we introduce G as an objective, third-personal, predicate applying to all and only line drawings with such and such geometrical properties, the properties exemplified above by Figure One. What is the truth-value of "if something is G then it's facical" in actual-world-candidates other than world w? Given that we know nothing about w beyond that Figure F occurs in it, and that the features of Figure One that contribute to its classification as facical are summed up in predicate G, there seems little option but to say that "if something is G then it's facical" is true in

any world considered as actual.

But if "Gs are facical" is true in all actual-world-candidates bar none, then principle (A) makes a prediction: it predicts that "Gs are facical" is true priori. This is a problem because intuitively, you have to do some experimenting to realize that Gs are facial: you have to expose yourself to Gs and ask yourself if you see faces in them. Reflecting on the geometrical property will not do it, no more than reflecting on the geometry of Muller-Lyer lines will reveal to you that one will appear longer.

I suggest then that "Gs are facical" is a counterexample to principle (A). More things are true in every candidate-for-actuality than are true a priori. The reason is simple enough. When we evaluate sentences in candidates-for-actuality, we are allowed to exercise any dispositions that inform our understanding of the relevant words; we are allowed in particular to check how various items considered-as-actual strike us or make us feel. When we ask about a sentence's status as a priori or not, we are not as generous about allowable methods of verification. A sentence that we cannot know to be true without self-experimentation is not considered a priori. (Otherwise it would be a priori which geometrical figures make for optical illusion!)

VII. The claim so far is that if we go by (A), then claims employing grokking predicates can be non-a-priori without a falsifying world. And now a conjecture: "pain" is a grokking predicate. Pain is picked out at least partly in terms of non-representational aspects of our first-personal experience of pain, viz. the hurting aspect. (Compare the seeing-a-face-in-it aspect of facicality.) If the conjecture is granted, then someone who thinks that pain = c-fiber-firings can defend her position as follows.

I agree with you that it is not a priori that all cases of c-fibers firing are cases of pain. The reason however is not that some world-considered-as-actual makes "there are c-fiber firings but there's no pain" true. That would be a world that contained c-fiber firings with the odd property that were we to expose ourselves to these c-fiber firings in a first-personal way, it wouldn't hurt. But, just as attending (from the right perspective, with our actual dispositions) to a line drawing with property G, we can't help but see a face in it, exposing ourselves (from a first-personal perspective, with our actual dispositions) to c-fiber firings, we can't help but feel pain. This is why I say that there are no worlds which considered as actual make "there are c-fiber firings but there's no pain" true.

How can the conditional be non-a-priori despite being true in every world considered as actual? The reason is implicit in the above. When, in the course of calculating primary intensions, we look for the "hurt" in an objectively given state, we allow ourselves to exercise actual-world dispositions on that state, including self-experimental dispositions. We ask ourselves: does that hurt when I appropriately expose myself to it? The rules change when we switch from calculating primary intensions to judging a priority. When in the course of judging a priority we look for the "hurt" in the same state, we do not allow ourselves the privilege of exercising semantically relevant dispositions.

So, yes, it is a posteriori that where there are c-fiber firings there is pain. But that doesn't make for a zombie-world any more than the a posteriority of "where there are G-figures, there is facicality" makes for a world physically like ours but in none of whose figures we can see faces.

Objection: You've misunderstood the rationalist claim. What does it take for a thing considered as actual to go into an expression's primary intension? Contrary to what you suppose, being able tell by virtue of our competence with the expression that it applies to X is not enough. The question is "what one can know justified independently of experience by virtue of one's competence. Reactive dispositions that proceed via self-experimentation don't yield experience-independent knowledge" (Chalmers, p.c.) Rather than (E) I would say that

(E*) it is a priori that S iff for all worlds w, we are justified independently of experience in thinking that in w-construed-as-actual, it is the case that S.

Reply: OK, but then (N) above no longer follows. All that follows from a failure of a priority is that is that there are worlds which we are not a priori justified in regarding as S-worlds. That is,

(N*) if it is not a priori that S, then for some worlds w, we are not justified independently of experience is thinking that w-construed-as-actual, it is the case that S.

Let it be then that zombies are a priori conceivable, in the sense that it is not a priori that if physics then pain. It no longer follows that there are zombie-worlds; all there have to be is worlds that might be zombie worlds as far as we can tell a priori. That much is true already of the actual world!

Objection: The rationalist theorist assumes that worlds are given in such a way that (vagueness aside) one can always tell a priori just in virtue of one's competence whether w is an S-world or not. A world w which we're not justified independently of experience in regarding as a pain-world is a world w that we are justified independently of experience in regarding as a zombie-world.

Reply: The assumption that worlds are given in such a way as to permit a priori determination of whether or not S is the case is a tendentious one in this context. The assumption pinches in one of two places, depending on whether we start from the world end of things or the meaning end. Starting from the world end, let's suppose that worlds are given in terms of their "lower-level" properties, which I assume means relatively natural properties and/or properties on which everything else supervenes. This is what Dave suggests when he says:

> The supervenience conditionals that we are considering.. have the form "If the low-level facts turn out like this, then the high-level facts will be like that." (TCM)

These conditionals are for the rationalist analytic in the sense of being true in virtue of meaning, indeed the apriori aspect of meaning. I hope the pinch here is clear. The "a priori determination" assumption entails that nothing counts as a meaningful expression unless its meaning can be given in terms of supervenience conditionals of the kind suggested. But then the modal rationalist is committed to something prima facie analogous to the analytic confirmation relations advocated by Carnap and rejected by Quine.

One might wonder whether this leaves room for any ordinary English expressions; a Quinean would say it did not, and I am tempted to agree. But leaving that aside, the a priori determination picture clearly leaves no room for expressions like "facical," or for that matter "graceful," or "thrilling," or etc. Grasp of the meaning of "facical" etc. does not carry with it an ability to

intellectually contemplate worlds-c.a.a. and to decide on the basis of that contemplation which things the predicate applies to. Grasp of the meaning of "facical" ("graceful" etc.) takes the form of an ability to tell what is facical when confronted in the right sort of way, not with a representation of the relevant circumstances, but with the circumstances themselves.

Starting from the meaning end, let's suppose that worlds are given in terms that always do make it a priori detectable whether they are S-worlds or not. The means that facts about what is facical – and so facts about what would be seen as facical by people like us if were to confront them -- have a place in lower-level world-descriptions.

This again pinches – facts about what we would see faces in don't seem terribly lower-level – but never mind; let them in. The problem now is that if we let them in here, we should let them into our description of the alleged "zombie"-world as well. A world physically like ours has not been fully described until it's specified whether subjects like ourselves would (if suitably positioned) grok c-fiber firings in a painful way. Since the identity-theorist presumably thinks that the only permissible specification here is that we would feel pain in the envisaged circumstance, it's not clear why she can't just deny the dualist's claim that there's a world physically like ours but with no pain. Because, to repeat, whether or not there is pain is a world-c.a.a. depends on whether or not "pain" in our mouths stands for anything there; and that depends on whether anything there would be experienced as pain were we given the right sort of first-personal access to it.

Objection: None of your problems arise if we see facicality as a response-dependent concept: x is facical in w-c.a.a. if it is such as to elicit a certain sort of perceptual reaction in w's population. On that reading of "facical" I can move a priori from lower-level facts about the w-folks' reactions to determinations of facicality. I appreciate that that's not the way you want "facical" read – but hear me out.

Suppose you're right that there are words like "facical" and "graceful" whose meanings are closely tied to our actual reactive dispositions; the concept of facicality is not to be identified with the response-dependent concept of eliciting face-seeing reactions. Then as you say, a world-c.a.a. in which our dispositions were different would be, not a world in which Gs failed to be facical, but one in which "facical" meant something different -- shmacical, say -- and Gs were not shmacical.

Surely though nothing much turns on how we individuate meanings: on whether we say that "facical" means something different in the contemplated world (as you insist) or means the same (as on my response-dependent interpretation). Either way I maintain that it is a priori that S if and only if for all worlds w (even those where S changes meaning) in w-taken-as-actual it is the case that S. That there are worlds where people don't see faces in Gs is enough to make it non-a-priori that Gs are facical quite independent of the question whether "facical" has changed meaning in those worlds.

Reply: Now we're drifting back towards the T-schema problem. You can't really mean that S is a priori iff for all worlds w -- even those where S means something different -- if w obtains then

S. For there are bound to be worlds which if actual make S come out meaning something false. (It's a priori conceivable that "sisters are siblings" means that prime numbers are always even.) The effect of your proposal is to make it a priori that S iff it's a priori that "S" means something true. And as we've seen, that would have the result that it is almost never a priori that S.

Objection: It would have that result only if we assume that it is never a priori that "S" means something true. I deny this. Or rather I distinguish two cases. If we're talking about deferential meaning, as in Burge, then you may be right that it is rarely a priori that "S" means something true. But that's OK, because if "S" is interpreted deferentially, then it is not going to be a priori that S either. (It is epistemically possible that arthritis is a disease of the joints, and that a fortnight is ten days.) If we're talking about personal, non-deferential, meaning, then it easily can be a priori that "S" is true. I know a priori that "sisters are siblings," is, at least as I interpret it, true. So there is nothing to prevent me from knowing a priori too that sisters are siblings.

Reply: Hmmm……your views have a strange consistency. It won't convince you, but let me just register my dissent on both counts. Start with the deferential case. It's not at all clear why the fact that I can be wrong about what I mean by "S" should prevent me from knowing a priori that S. This would be plausible if I had inferred S from a premise about what "S" means; or if I implicitly depended on such a premise; or if its meaning something different was a "relevant alternative" that I had to rule out to count as knowing. But I don't see that any of these things is true. That may be just a quibble about the meaning of "a priori." But consider now the supposedly non-deferential case. It seems to me that all meaning, however personal, is deferential to this extent. Rationality requires us to remain always open to the possibility that someone will come along and convince us that even our own usage is best understood on a meaning-hypothesis that we find initially counterintuitive and surprising. This to me suggests that it is not a priori for me that "S" means something true even if "S"'s meaning is understood to be bound to no one's usage but my own.

Of course the rationalist could respond that the epistemic possibility of my being improperly or unreasonably convinced that "S" means something false counts for nothing. Or, taking a leaf from our objection above, that it counts for nothing if the scenario in which I am convinced that "S" means something false is one where the quotation-name "S" has changed meaning and so we are talking about a different sentence. But it seems to me that my grasp of a sentence's meaning is never itself enough to rule out the possibility of criticisms free of these defects. That however is a topic for another time.