Mark Schroeder
University of Southern California
October 22, 2008

stakes, withholding, and the subject-sensitivity of 'knows'

In recent work, John Hawthorne [2004], Jason Stanley [2005], and Jeremy Fantl and Matthew McGrath [2002], [forthcoming] have each advocated the view that 'knows' is subject-sensitive – that two agents in possession of the same evidence, who believe the same thing, might nevertheless differ with respect to whether they know it. Hawthorne argues for this view on the basis of lottery considerations, Stanley on the basis of cases, and Fantl and McGrath on the basis of considerations about fallibility. In this paper I advance yet a fourth motivation this view – on the basis of considerations about withholding. I'll show that like the (closely related) view of Fantl and McGrath, this view yields correct intuitive predictions about cases on which Hawthorne and Stanley's approaches intuitively fail. These predictions explain how Hawthorne and Stanley could be right in spirit, though not detail. Then I develop a more detailed model, which makes even more predictions, which I'll argue are also confirmed. Finally, I go on to show how to explain the features of so-called 'ignorant high stakes' cases, which might be thought to be problematic for my style of explanation.

I harman's observation

I'll understand *evidentialism* to be the thesis that *X* has at least as much reason to believe that *p* as that *q* just in case *X* has at least as much evidence that *p* as that *q*. So, substituting, *X* has at least as much reason to believe that *p* as that ~*p* just in case *X* has at least as much evidence that *p* as that ~*p*.

Gilbert Harman [2002] has observed that evidentialism (which he presupposes) leads to an apparent asymmetry between practical and epistemic rationality: in the practical domain, it is rational for *X* to do *A* just in case *X* has at least as much reason to do *A* as to not do *A*. But if we substitute 'believe that *p*' for 'do A', we get the thesis that it is rational for *X* to believe that *p* just in case *X* has at least as much reason to believe that *p* as to not believe that *p*. *If* we equate reason to not believe that *p* with reason to believe that ~*p*, this turns into the thesis that it is rational for *X* to

believe that *p* just in case there is at least as much reason for *X* to believe that *p* as for *X* to believe that ~*p*. And by evidentialism, this is equivalent to the thesis that it is rational for *X* to believe that *p* just in case *X* has at least as much evidence that *p* as that ~*p*. But this, Harman observes, is false. It is not rational to believe things for which your evidence is no better, or even barely better, than for its contrary. You should only believe something when your evidence is *substantially* better than for its contrary. So, Harman concludes, epistemic rationality is importantly different from practical rationality.

2 withholding

I beg to differ. Harman's conclusion relies on the equation of reason to not believe that *p* with reason to believe that ~*p*. But this is a mistake. There are (at least) *two* ways to not believe that *p*. One is to believe that ~*p*, but the other is to *withhold*, neither believing that *p* nor believing that ~*p*. Hence, reason to withhold is reason to not believe that *p*, and consequently the reason to not believe that *p* is not exhausted by the reason to believe ~*p*. This gives us an answer to Harman's observation: epistemic rationality *is* like practical rationality, in that it is rational for *X* to believe that *p* just in case there is at least as much reason for *X* to believe that *p* as for *X* to not believe that *p*. But for it to be rational to believe that *p*, *X* must have substantially greater evidence that *p* than that ~*p* – enough to overcome the reason to withhold, in addition to overcoming the evidence that ~*p*.

It follows that evidentialism, as I have defined it, is compatible after all with the robust parallel between practical and theoretical rationality. It *is* rational for *X* to do *A* just in case *X* has at least as much reason to do *A* as to not do *A* – even where doing *A* is believing that *p*. But in any case in which *X* has reasons to withhold, it is rational for *X* to believe that *p* only if *X*'s evidence that *p* substantially outweighs *X*'s evidence that ~*p* – in proportion to the strength of *X*'s reasons to withhold.

This brings me to my logical point: if two agents differ in their reasons to withhold, then it may be rational for one to believe that *p* but not rational for the other to believe that *p*, even though each is in possession of the very same evidence, *and even though evidentialism is true*. This may happen because the latter has more reason to withhold than the former does. So *if* reason to withhold can vary from agent to agent, it follows that rational belief is subject-sensitive. And if that is so, then on the assumption that *X* knows that *p* only if it is rational for *X* to believe that *p*,

something similar may go for knowledge. Our question, then, is: *does* the reason to withhold vary from agent to agent? To answer that, we need to know what could give us reason to withhold.

3 what is the reason to withhold?

It is a well-known fact that the 'epistemic goals' of believing only truths and of believing all of the truth come into conflict. To believe *only* truths is to avoid *type-1* error: believing that *p* even though *p* is false. To believe *all* the truth is to avoid *type-2* error: failing to believe that *p* when *p* is true. By definition, withholding guarantees type-2 error. So intuitively, whenever type-1 error is worse that type-2 error, there will be net reason to withhold (I'll give a more detailed picture in a moment). Moreover, even among cases in which type-1 error is worse than type-2 error, in cases in which type-1 error is *worse*, in comparison to type-2 error, there will be more reason to withhold.

Jason Stanley's bank case illustrates a special case of this observation, by introducing a variation in *stakes*. Stanley's variation in stakes introduces a variation in the badness of type-1 error. His cases go like this:

> *Low Stakes*. Hannah and her wife Sarah are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. It is not important that they do so, as they have no impending bills. But as they drive past the bank, they notice that the lines inside are very long, as they often are on Friday afternoons. Realizing it isn't very important that their paychecks are deposited right away, Hannah says, 'I know the bank will be open tomorrow, since I was there just two weeks ago on Saturday morning. So we can deposit our paychecks tomorrow morning.'
>
> *High Stakes*. Hannah and her wife Sarah are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. Since they have an impending bill coming due, and very little in their account, it is very important that they deposit their paychecks by Saturday. Hannah notes that she was at the bank two weeks before on a Saturday morning, and it was open. But, as Sarah points out, banks do change their hours. Hannah says, 'I guess you're right. I don't know that the bank will be open tomorrow.'[1]

In Low Stakes, it is less important for Hannah to avoid type-1 error with respect to the proposition that the bank will be open on Saturday. This is because the worst thing that happens if she is wrong, is that she will end up making two trips to the bank, one on Saturday (when it is

[1] These two cases, as well as that of Ignorant High Stakes, below, I take verbatim from Stanley [2005]. The other cases below are mine.

closed), and one on Monday (to deposit her check). But in High Stakes, it is more important for Hannah to avoid type-I error with respect to the proposition that the bank will be open on Saturday. This is because if she is wrong, then she and Sarah will be unable to pay their impending bill.

Understanding reason to withhold in terms of the relative cost of type-I versus type-2 error therefore allows us to predict and explain the subject-sensitivity of 'knows' in Stanley's Low Stakes and High Stakes cases. But contra Stanley, the account developed here predicts that stakes are not all that matters, and indeed that stakes do not always matter.

4 new predictions

The role of Stanley's stakes is to raise the cost of type-I error. But our account so far predicts that cost of type-I error makes a difference only when type-I error is worse than type-2 error. In many normal circumstances type-I error *is* worse than type-2 error. But this is not always the case, and it is certainly not the case in cases of forced choice. Here are two such cases[2]:

> *Forced Choice, Low Stakes*: Hannah and her wife Sarah are out driving on Saturday morning, at twenty minutes to noon. Sarah remembers that they still haven't deposited their paychecks from Friday, but points out that just one of their bank's two branches is open until noon on Saturdays, but she can't remember which, and there is only time to try one. Hannah says, 'I know which one it is – I was at the branch on Chapala Street two weeks ago and it was open, then. Let's go there.'
>
> *Forced Choice, High Stakes*: Hannah and her wife Sarah are out driving on Saturday morning, at twenty minutes to noon. Since they have an impending bill coming due, and very little in their account, it is very important that they deposit their paychecks that day, but they have so far forgotten to do so. Sarah remembers that they still haven't deposited their paychecks from Friday, but points out that just one of their bank's two branches is open until noon on Saturdays, but she can't remember which, and there is only time to try one. Hannah says, 'I know which one it is – I was at the branch on Chapala Street two weeks ago and it was open, then. Let's go there.'

In the cases of Forced Choice, Hannah and Sarah need to make a decision about where to go, or they are guaranteed not to get their money deposited that day. As in the original versions of Stanley's cases, the costs of a type-I error in the High Stakes version are greater than in the Low

[2] Shaffer [2006] offers cases which exploit the same features diagnosed here as an objection to Hawthorne and Stanley, although he acknowledges in a note that they may not raise a problem for the view of Fantl and McGrath [2002], which is closely related to the one offered here.

Stakes version. But in the Forced Choice cases, the costs of a type-2 error are also greater in the High Stakes version than in the Low Stakes version. So our account predicts that in these cases, stakes should not make a difference. I believe that this prediction is borne out by the cases.

It is a side-effect of this diagnosis that it explains the appeal of attributions of knowledge in so-called 'game show' or 'exam' cases. The contestant must answer a multiple-choice question in a certain time limit; she is highly uncertain as to which answer is correct, but in the end, in time for the buzzer, she chooses answer 'B' and gets it right. Colloquially, we say that she knew the answer. Without weighing in, here, on whether it is possible to know that *p* even if one's evidence licenses credence of less than .5 in *p*, I note that game show cases are cases of forced choice, and if structured correctly, are cases in which type-2 error is as bad as type-1 error. If that is right, the use of 'know' in such cases may not be terribly different from uses of 'know' in other contexts.

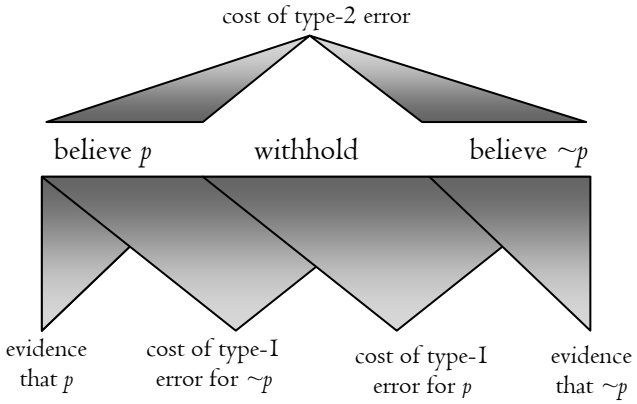
If these considerations are on the right track, then Stanley and Hawthorne are right in spirit, though not detail. Rational belief is subject-sensitive, but it does not, in general, vary with the stakes. The variation in the stakes is subsumed under a broader and more deeply explanatory generalization about the relative costs of type-1 and type-2 error with respect to a given proposition.

5 a more detailed model

So far, the picture I have been developing, in terms of the costs of type-1 and type-2 error, has I hope been intuitive, but it has also been suggestive rather than precise. In fact I think things are somewhat more complicated, and in this section I'll provide a simple model to illustrate those complications. The model will also allow us to make further correct predictions.

The first complication arises from the fact that there is, in fact, no such thing as the cost of type-1 error regarding some proposition. There is only the cost of type-1 error in believing it, and the cost of type-1 error in believing its negation. These two costs can easily come apart. For example, in High Stakes, the cost of wrongly believing that the bank is *not* open on Saturday is that Hannah and Sarah will have to stand in line, whereas the cost of wrongly believing that it *is* open on Saturday is that Hannah and Sarah will miss their impending bill. The latter cost can easily outweigh the former – for example, it could lead the bank to foreclose on their house, if we set the case up correctly.

Since there is strictly speaking no such thing as the cost of type-I error with respect to a proposition *tout court*, strictly speaking the costs of type-I error are not associated with reasons to withhold. Rather, the costs of type-I error of believing *p* are associated with a reason to *not believe p* – i.e., to either withhold or believe ~*p*. And similarly, the costs of type-I error of believing ~*p* are associated with a reason to not believe ~*p* – i.e., to either withhold or believe *p*. Meanwhile, the costs of type-2 error are associated with reasons to *not withhold* – i.e., to either believe *p* or believe ~*p*. This leads to the following picture, with shaded triangles illustrating which of believing *p*, withholding, and believing ~*p* is supported by each kind of reason:



The picture illustrates that evidence that *p* is reason to believe that *p*, evidence that ~*p* is reason to believe that ~*p*, the cost of type-2 error about *p* is reason to either believe *p* or believe ~*p*, the cost of type-I error for *p* is reason to not believe *p*, and the cost of type-I error for ~*p* is reason to not believe ~*p*. There are no direct reasons to withhold in this picture; only the net interaction effect of reasons to not believe *p* and reasons to not believe ~*p*.

Believing *p* is supported by the evidence that *p* (Ev_p), the cost of type-2 error (Err2), and the cost of type-I error in believing ~*p* ($ErrI_{\sim p}$). Similarly, believing ~*p* is supported by the evidence that ~*p* ($Ev_{\sim p}$), the cost of type-2 error (Err2), and the cost of type-I error in believing *p* ($ErrI_p$). Finally, withholding is supported by the costs of each type of error ($ErrI_p$ and $ErrI_{\sim p}$). So believing that *p* is at least as well supported by reasons as the alteratives just in case the total reasons which support believing *p* are at least as weighty as the total reasons which support believing ~*p* and at least as weighty as the total reasons which support withholding – that is, just in case:

$$Ev_p+Err2+ErrI_{\sim p} \geq Ev_{\sim p}+Err2+ErrI_p \quad \text{and} \quad Ev_p+Err2+ErrI_{\sim p} \geq ErrI_p+ErrI_{\sim p}$$

or, cancelling,

$$Ev_p + ErrI_{\sim p} \geq Ev_{\sim p} + ErrI_p \quad \text{and} \quad Ev_p + Err2 \geq ErrI_p.$$

This means that there are *two* ways in which having better evidence that *p* than that ~*p* can still fail to make it rational to believe *p*. This can happen if the costs of type-1 error of believing *p* exceed the costs of type-2 error by a sufficient amount – enough to outweigh the evidence that *p* and make the second conjunct fail:

Condition A: $ErrI_p - Err2 > Ev_p$

Or it can happen if the costs of type-1 error of believing *p* exceed the costs of type-1 error of believing ~*p*:

Condition B: $ErrI_p > ErrI_{\sim p}$

If either of these conditions are satisfied, then there is *net* reason to withhold, in the sense that it takes more than having better evidence that *p* than that ~*p* in order to make it rational to believe that *p*. If condition A holds, then it is more rational to withhold than to believe *p*, and if condition B holds, then it is more rational to believe ~*p* than to believe *p*.

Stanley's High Stakes and Low Stakes cases trigger both conditions, by raising the cost of type-1 error in believing that the bank will be open tomorrow, without raising the cost of type-1 error in believing that the bank will not be open tomorrow, and without raising the cost of type-2 error. The forced choice cases trigger neither condition, because since they are symmetric, they raise the cost of type-1 error about both propositions at the same time, and the nature of the forced choice also raises the cost of type-2 error to match the cost of type-1 error. So the more detailed model makes good on the intuitive claims I made in the last two sections.

6 two more predictions confirmed

If the model outlined in the last section is roughly correct, then we should expect there to be further cases which trigger one condition but not the other. In fact, it is easy to go either way:

> *Nasa Engineering.* Hannah and Sarah are engineers working on the design of NASA's next-generation shuttle, a multi-billion dollar project planned to operate

over several decades and ultimately carry hundreds of astronauts into space, where error means death. Currently they are trying to decide which materials to use for an important component, and are investigating two new alloys, to see which will be more appropriate for the component. Citing preliminary research, Sarah notes that the first alloy holds up better under temperatures under 300 degrees, and that most alloys which hold up well under 300 degrees also perform well at shuttle temperatures. Hannah objects, 'yes, but not all of them do, so we won't know until we do more research.'

Game Show. Hannah and Sarah are playing Go Big or Go Home, a successful game show on daytime television with a B-celebrity host. They have reached the final question, which is: 'will the bank be open tomorrow, on Saturday?'. The possible answers are 'yes' and 'no', and they must answer within the time limit, or they will lose all of their money (they have accumulated a very large sum so far). If they answer and get it right, they double their money, but if they answer 'yes' and get it wrong, they lose all of their money and if they answer 'no' and get it wrong, they keep what they already have. Fairly confident in the answer, Hannah tells Sarah, 'I know the answer is 'yes', because I was there two weeks ago on a Saturday morning.' Sarah objects that banks do change their hours. Hannah responds, 'I guess you're right – I don't really know that the bank will be open tomorrow.'

Nasa Engineering is a case in which the costs of each direction of type-1 error are equally high, and both are much higher than the costs of type-2 error, since there is no forced decision – the engineers can simply wait on more research before deciding which alloy will hold up better. So it triggers condition A but not condition B, making it more rational to withhold than to believe *p*, but not more rational to believe $\sim p$ than to believe *p*. Game Show, in contrast, is a situation of forced choice, which raises the cost of type-2 error, but it is constructed to keep the relative costs of the two directions of type-1 error different. It is a more controversial case, but if it works, it triggers condition B but not condition A, making it more rational to believe $\sim p$ than to believe *p*, but not more rational to withhold, than to believe *p*. So these cases confirm, I think, the predictions made by the model of the last section.

7 ignorant high stakes

So far, I have been offering an explanation of the subject-sensitivity of knowledge by way of explaining the subject-sensitivity of belief. On the assumption that rational belief is necessary for knowledge, this predicts that knowledge may also be subject-sensitive, across cases in which the other necessary conditions for knowledge are all satisfied. But it is well-known that not all cases in

which intuitions about knowledge vary with stakes are cases in which this explanation can be at work. So let's consider the case which raises this problem:

> *Ignorant High Stakes.* Hannah and her wife Sarah are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. Since they have an impending bill coming due, and very little in their account, it is very important that they deposit their paychecks by Saturday. But neither Hannah nor Sarah is aware of the impending bill, nor of the paucity of available funds. Looking at the lines, Hannah says to Sarah, 'I know that the bank will be open tomorrow, since I was there just two weeks ago on Saturday morning. So we can deposit our paychecks tomorrow morning.'

Ignorant High Stakes is like Low Stakes, so far as Hannah and Sarah can tell, but in terms of the real costs of type-I error, it is like High Stakes. Correspondingly, it is *rational* for Hannah to believe that the bank will be open tomorrow, but she is wrong to say that she knows it. So high stakes can affect knowledge, even when they don't affect rationality. (Similar cases show that ignorant high stakes affect knowledge in other cases in which either Condition A or Condition B is triggered, but not in cases in which neither is trigged, such as forced-choice situations.)

On the face of it, this looks like a problem for my proposed explanation of the role that stakes can play, on the model of thinking about rationality. If stakes can affect knowledge even when they don't affect rational belief, then maybe I am on the wrong track about what is doing the work, here. But I think this is a mistake. Familiar cases from the literature show that counterevidence of which the subject is unaware and undercutting defeaters of which the subject is unaware can affect knowledge without affecting rationality, in exactly similar ways:

> *Ignorant Counterevidence.* Hannah and her wife Sarah are driving home on a Friday afternoon. They plan to stop at the bank on their way home to deposit their paychecks. But as they drive past the bank, they notice that the lines inside are very long, as they often are on Friday afternoons. Not seeing the large sign standing outside the bank which says that it will be closed tomorrow, Hannah says, 'I know the bank will be open tomorrow, since I was there just two weeks ago on Saturday morning. So we can deposit our paychecks tomorrow morning.'

> *Ignorant Undercutting Defeat.* Hannah and her wife Sarah are driving home on a Friday afternoon. They plan to stop at the bank on their way home to deposit their paychecks. But as they drive past the bank, they notice that the lines inside are very long, as they often are on Friday afternoons. Not realizing that the desk editor of the Press-Gazette has been filling it with lies lately in the hope of getting fired, Hannah says, 'I know the bank will be open tomorrow, since I remember reading it in the Press-Gazette.'

In both Ignorant Counterevidence and Ignorant Undercutting Defeat, it is rational for Hannah to believe that the bank will be open tomorrow – she doesn't know about the counterevidence provided by the sign out front, or about the undercutting defeater provided by the desk editor's erratic behavior. But plausibly, in neither case does she know that the bank will be open – even if it turns out to be true.

In all three cases involving ignorance, there is a proposition which *would* make it irrational for Hannah to believe that the bank will be open tomorrow, if she was aware of it. In all three cases, its truth defeats her knowledge, even though she is unaware of it. This pattern is a *general* phenomenon about the relationship between knowledge and rationality, not an idiosyncratic feature of high stakes. I believe that this constitutes not just a response to the problem posed by Ignorant High Stakes, but collateral evidence for the picture offered here, on which facts about what it is rational to believe drive the subject-sensitivity of knowledge.

9 open questions

In this short paper I haven't sought to establish *what* the relationship is between knowledge and rationality or knowledge and reasons; nor have I sought to defend subject-sensitivity from other kinds of attack. I have just tried to provide a model, in terms of reasons for withholding, that explains when and why we should expect stakes to make a difference in what it is rational to believe, and when and why we should consequently expect stakes to make a difference in what we know, *given* some general facts about the relationship between knowledge and rationality.

The argument has been simple: given evidentialism, if rational belief is like rational action, in that it is rational to do what is supported by at least as weighty reasons as the alternative, then rational belief must depend on reasons to withhold. The obvious place for reasons to withhold to come from, is from the relative costs of type-1 and type-2 error, which can vary, in principle, from person to person. Indeed, a look at the nature of these costs not only allows us to capture intuitions about existing cases, but to make confirmable predictions about further kinds of case.

For now, I'll leave it to the reader to determine whether this is compatible with his or her favorite non-reason-implicating account of knowledge, though I hope to be able to take this up on another occasion.[3]

[3] Special thanks to Jake Ross for many fruitful discussions.

references

Fantl, Jeremy, and Matthew McGrath [2002]. 'Evidence, Pragmatics, and Justification.' *Philosophical Review* 111(1): 67-94.

_____ [forthcoming]. *Knowledge in an Uncertain World.*

Harman, Gilbert [2002]. 'Practical Aspects of Theoretical Reasoning.' In Al Mele and Piers Rawling, eds., *The Oxford Handbook to Rationality.* Oxford: Oxford University Press.

Hawthorne, John [2004]. *Knowledge and Lotteries.* Oxford: Oxford University Press.

Shaffer, Jonathan [2006]. 'The Irrelevance of the Subject: Against Subject-Sensitive Invariantism.' *Philosophical Studies* 127: 87-107.

Stanley, Jason [2005]. *Knowledge and Practical Interests.* Oxford: Oxford University Press.