# SURPRISE, SELF-KNOWLEDGE, AND COMMONALITY

Frederic Schick

The point of a Festschrift is to honor the person to whom it will be presented. This calls for the papers written for it connecting somehow with the honoree's work. My paper connects only indirectly, by a sort of exclusion: it deals with topics outside the range of those that Peter has written about. Still, I hope it will interest him.

## I

Here is a familiar puzzle. A teacher announces on Monday that there will be a surprise exam on either Wednesday or Friday. Her students reason as follows. Say that the teacher's announcement is true. Then, if the exam were on Friday, we would know by Thursday, and so it wouldn't be a surprise. Therefore it won't be on Friday. That means it must be on Wednesday, and since we now know that, it won't then surprise us. There can't be a surprise exam on either of these days. The teacher's announcement is false: it contradicts itself.

The teacher gives the exam on Friday and everyone is surprised. Where did the students go wrong? This has been much discussed, and I want to discuss it once more. I want then to extend my discussion to some larger issues.

Let me bring it down to just a single student's problem. The announcement can be put as three suppositions. There will be an exam on either Wednesday or Friday:

(1) $W \vee F$

And it will come as a surprise to this student. If, that is, it will be on Wednesday, he won't, on Tuesday, believe it will be on Wednesday. And if it will be on Friday, he won't, on Thursday, believe it will be on Friday:

(2) $W \supset \sim B_t W$

(3) $F \supset \sim B_{th} F$[1]

Let us suppose also that both

(4) $\sim(W \cdot F)$ and

(5) $\sim W \supset B_{th} \sim W$

Assume now it will be on Friday. This begins a conditional proof:

(6) $F$

(7) $\sim B_{th} F$ (6) and (3)

(8) $\sim W$ (6) and (4)

(9) $B_{th} \sim W$ (8) and (5)

We want to proceed to $B_{th}F$, but (9) and (1) don't warrant that, for perhaps the student won't believe (1) on Thursday. We need to add the supposition that

(10) $B_{th}(1)$

This now gives us

(11) $B_{th} F$ (10) and (9)

and, by reductio,

(12) $\sim F$ (11) and (7)

We go on to

(13) $W$ (12) and (1)

(14) $\sim B_t W$ (13) and (2)

These two lines say that the exam will be on Wednesday and will surprise the student. No contradiction here.

But suppose also that, on Tuesday, the student believes all our suppositions above:

(15) $B_t(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 10)$

We then get

(16) $B_t W$ (15) and (1) to (13)

Here we do have a contradiction: (16) contradicts (14). But it doesn't follow that the teacher's announcement was false -- this is where the students went wrong. What follows is that either that *or something else the argument uses* is false. That is, what follows is

(17) $\sim(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 10 \cdot 15)$

This rests on a tacit assumption. It supposes that the student is deductively *thorough*, that he believes the deductive consequences of the conjunction of all he believes, that he is in this sense *thorough* every day of this week.[2] Suppose now also that his beliefs are *stable*, or at least stable this week, barring new relevant information -- that he is belief-*retentive*, that he gives up no beliefs unless he gets such information. This allows us to simplify. Let *m* be Monday, *today*. Then (15) can be replaced by

(15') $B_m(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 10)$

and since, by retentiveness, this implies (10) and (10) is therefore redundant, by

(15") $B_m(1\cdot2\cdot3\cdot4\cdot5)$

So (17) reduces to

(17') $\sim(1\cdot2\cdot3\cdot4\cdot5\cdot15")$

which is equivalent to

(18) $(1\cdot2\cdot3\cdot4\cdot5) \supset \sim B_m(1\cdot2\cdot3\cdot4\cdot5)$ and to

(19) $B_m(1\cdot2\cdot3\cdot4\cdot5) \supset \sim(1\cdot2\cdot3\cdot4\cdot5)$

Nothing is wrong with the teacher's announcement, nor indeed with 1·2·3·4·5; certainly the conjunction isn't contradictory. But there clearly *is* something wrong with $B_m(1\cdot2\cdot3\cdot4\cdot5)$. Let us describe a person who is both deductively thorough and belief-retentive (during this period) as *disciplined*. We have just seen that, if 1·2·3·4·5 is true, a disciplined student doesn't believe it (this is (18)). And also -- this says the same -- that if he believes it, it isn't true (19). That is, he *can't* believe it, not in good logical conscience: there is no possible world in which he believes it and it is true.

Here is a simpler scenario.[3] The teacher tells the student there will be a surprise exam on Wednesday. She tells him that

(20) W and

(21) $\sim B_t W$

No problem with 20·21; perhaps the exam *will* be on Wednesday and will surprise the student. But if

(22) $B_m(20\cdot21)$, then

(23) $B_t W$ (22)

This contradicts (21), so it follows that

(24) ~(20·21·22) and also that

(25) (20·21) ⊃ ~$B_m$(20·21) and

(26) $B_m$(20·21) ⊃ ~(20·21)[4]

Nothing is wrong with the announcement here either. Still, the student should not have believed it. As with 1·2·3·4·5, 20·21 is not contradictory, but the student could properly believe it only if he weren't disciplined.[5] For he would be believing a proposition that can be shown to be false if he believes it and is disciplined. Hintikka (1962) calls such propositions *doxastically indefensible* for this person.[6] I will call them *incredible* for him, or not properly *believable* by him. He cannot *properly* believe such propositions because he could defend his believing them only by admitting to a lack of discipline. Another way of putting it: if he believed such a proposition and also believed he believed it, he would have to believe a contradiction (like 14·16 and 21·23).

I will stretch the concept of discipline to cover noncontra-diction, the *non*belief of *x*-and-not-*x*; this now makes what is logically false incredible for all. Still, we have been speaking of incredible propositions that may in fact be true -- the teacher's announcement *is* true -- and these are incredible for some people only. Thus 20·21 is incredible for the student but isn't so *for me*, there being nothing in that announcement about *my* being surprised. *I* can properly believe it, the student involved cannot. Getting beyond our scenarios: I can predict

that this or that will surprise someone else, but no one can say about himself (can properly believe) that *x* will surprise him.

Let me touch on two small points. Is the above about *surprise*? I held that *x* would surprise the student if he didn't believe it beforehand. Perhaps this puts it badly. Perhaps a person is only surprised if he believed *not-x* beforehand. And he clearly isn't surprised unless he has come to know *x*. No problem in that for us here. Suppose that the teacher meant that her students would be surprised in the fuller sense. Strengthening her announcement to make that explicit cannot weaken its force, so our conclusion would still stand. The teacher's announcement would remain incredible for the student.[7]

Also, the title mentions self-*knowledge*, but the above has brought out only that we can't properly *believe* certain propositions. Still, if we cannot believe them, we cannot know them either. I am keeping here to *beliefs* because that shows why self-knowledge has limits. We can't have full self-knowledge because we can't have full self-belief. We aren't fully knowable to ourselves because we aren't fully credible to ourselves.

## II

The surprise exam conundrum is of no interest in itself. That is why I discussed it. I counted on no one's really caring how things came out in that situation to let the reader be receptive to my analysis of it. I want now to apply that analysis to a different situation.

My wife knows me very well. She knows me better than I know myself. Still, she sometimes is wrong. Say that I am agonizing

whether to do *x* or *y*, and she tells me the outcome is clear: I will choose to do *x*.

I now think about this. She says that, on Wednesday, I will choose *x*. This *x* is indeed an option for me. That means it is (*a*) an action I neither yet think I will take or that I won't. It is also (*b*) an action I think I would take if I chose it. So if I accepted my wife's prediction that I will choose *x*, I would (by *b*) now think would take it. But then (by *a*) *x* wouldn't be an option, and neither would *y* be an option. There would be nothing for me to choose. So -- aha! -- she is wrong. She has contradicted herself! And I go on agonizing until, on Wednesday, I choose to do *x*. Her prediction turns out to be right; it was *not* contradictory. Where did my reasoning fail?

I will stand by *a* and *b*. A set of options composes an *issue*, and an issue is a situation it makes some sense to agonize over. There is no sense in agonizing where we know what we will do -- for instance, in sweating out the question of whether or not to have dinner tonight. And even where we don't know what will happen, it makes no sense to agonize where we think we are weak, where we doubt we would follow through. There is no point in Romeo's asking whether he ought to leave Juliet: he doubts he would leave her even if he chose to, so leaving her isn't an option for him. I take this to mean that *a* and *b* are basic to optionality.[8]

Where did my reasoning fail? My wife predicted I would, on Wednesday, choose option *x* -- *x* is my doing this or that. She predicted

(27) $C_w x$

Since what I choose must have been an option,[9]

(28) $x$ will on Tuesday be an option for me (27)

(29) $\sim B_t x$ (28), by *a* and

(30) $B_t(C_w x \supset x)$ (28), by *b*

This alone is not troublesome. But if I *believe* the prediction, believe it now, on Monday -- if, that is,

(31) $B_m(27)$, then

(32) $B_t C_w x$ (31), and

(33) $B_t x$ (32) and (30)

Since (33) contradicts (29), something has to go. But nothing is wrong with (27) -- this is where my thinking failed. What I should have concluded is

(34) $\sim(27 \cdot 31)$ and so too

(35) $(27) \supset \sim B_m(27)$ and

(36) $B_m(27) \supset \sim(27)$

Suppose I am fully disciplined.[10] Then, if (27) is true, I can't believe it. (This is (35).) And if I believe it, it can't be true. (This is (36).) My wife may be right to believe (27), but (27) isn't credible *for me*. Generalizing again: everyone can say about other people that they will choose certain options they have, but no one can say about himself (can properly believe) that he will choose his option $x$.[11]

We can't predict our own choices any more than we can predict our surprises. But this is not to be wondered at, for

choices are a sort of surprises (in my weak sense of the word). In choosing, we surprise ourselves; we come in the end to do what, before, we didn't think we would. And there is more of the same. We can't predict our *learnings*. No one can properly believe he will learn (or conclude, or discover, or realize) *x* -- a point stressed long ago by Popper (1950). For suppose you *will* learn *x*, say on this coming Wednesday. You can't learn what you already believe, so you won't yet believe *x* that morning. Still, you will then believe that whatever you learn is true. No contradiction here; you may in fact learn *x*. But if you *believe* that you will, the argument continues as in (32) to (36). Thus your learning *x* on Wednesday is now incredible for you. Since the truth of *x* will surprise you, you can't predict your learning it.

We can take this beyond *prediction*, beyond our foreknowledge (our fore*belief*) of our own choices and learnings and the like. For if *k* isn't properly believable and *h* implies *k*, then *h* isn't believable either. Say that *x*, *y*, and *z* are your options, $c_1$, $c_2$, $c_3$, $c_4$... are their consequences in different states of nature, *P* is your probability distribution and *U* your utility distribution, and that you are rational. Let all this be *h*. Say that it follows from *h* that *k*: that you will choose *x*. Since *k* is not believable for you, neither is *h*. But *h* is not a proposition about some choice you will make. Nor does it speak of anything else that will come to surprise you.

Again, these are limits to *self*-belief and self-knowledge. *Other*-knowledge is not so restricted. Jill can know what Jack will choose. Still, she can't convey that knowledge she has

about Jack to Jack -- such knowledge is out of bounds for him. This reverses an old and familiar thesis on mental privacy. The thesis is that there is much about us that we can know but no one else can, unless we choose to tell them (for instance, that I now have an itch). The point here is the opposite: that other people can know things about us that *we* can't know, even if these others tell us. Privileged access goes in both directions.

The point is not wholly new. Suppose that

(37) Jack believes nothing

And suppose that Jack believes this, that

(38) B(37)

Then it is false that Jack believes nothing,

(39) ~(37) (38)

This contradicts (37), so it follows that

(40) ~(37·38) and also that

(41) (37) ⊃ ~B(37) and

(42) B(37) ⊃ ~(37)

Jack may in fact be a total skeptic: (37) may be true. Jill can believe that Jack is a skeptic. But he himself can't properly believe it, for he cannot rightly believe what must be false if he believes it. The moral of this Cretan-like case (The Cretan says that he always *lies*) has to do with self-reference. The argument shows that certain propositions referring to a specified person *p* can't properly be believed by that person. So also with our other arguments: these too involve some *p*'s believing some proposition referring to *p*. They too reveal the impropriety of certain beliefs that are self-referential. The

only novelty in the new cases is that *these* self-referential beliefs (seldom under any suspicion) are revealed as improper.[12]

## III

I can hear a question. OK, we can't predict our choices, but we can predict our *actions*. We can't predict that we will choose *x*, but we can predict we will *do* it, that we will act out *x*. Believing we will *do x* isn't improper. True, if we believe this, we can't then *choose x*. (Since $C_t x \supset \sim B_m x$, $B_m x \supset \sim C_t x$.) We can only think we will *x*, *properly* think it, if we won't choose *x*. But why should that concern us?

First, because it has to affect what we think of ourselves as thinkers. The point is that we sometimes choose, and that we can't then know beforehand *what* we are going to choose, can't even know what we will *do* -- or even properly *believe* we will choose it or that we will do it. Others may know this, but we ourselves can't. To that extent our knowledge is bounded, and bounded not by our mental limitations but by our self-discipline.

This may leave you unconcerned, so here is a second point. We can indeed deny ourselves the beliefs that I say are improper. We lose nothing of any importance if we avoid such beliefs. Still, we sometimes ascribe such beliefs, if not to ourselves then to others. Sometimes we even endorse ideas that oblige us to do that. We trip ourselves up where we do, so it is well to be cautioned against it.

Let Jack and Jill be in a Prisoners' Dilemma. They can either cooperate or not -- not cooperating is *defecting*. For each, defection is the dominant option. Both Jack and Jill are rational, so they both will defect. And since they both prefer the outcome of joint-cooperation to that of joint-defection, they will both be sorry.

Where they think they won't meet again, there is no way out for them. But say that they think they will meet again, that their present interaction is only the first of many just like it. Here it may seem that the prospect each faces of having to live with the other's resentment ought to deter defection. Still, it often is argued that, where the number of rounds is finite and known to both of the agents, if they are rational they both will defect from the start to the finish.

The argument is this. Suppose that the number of rounds is known by both Jack and Jill to be 100, and that they both are rational. In the 100th, each will know that there will be no further meetings, no need to guard against reprisals, so they both will defect. Let Jack think that Jill is rational. He will then think in the 99th round that Jill will defect in the 100th whatever *he* does in the 99th, that his cooperating in the 99th round wouldn't be rewarded by Jill in the 100th. He will therefore defect in the 99th, and Jill, thinking likewise about Jack, will too. The same in round 98: each expecting the other to defect in the round that follows, each will here defect, though we must now also assume that Jack thinks that Jill thinks *him* rational -- and must also assume that Jill thinks he thinks *her* rational. So we move stepwise back to round 1 (both agents

defecting all the way), though with a heavier load of assumptions at each preceding stage.

This backward induction rests on assumptions about these people's beliefs about each other. Sometimes these are put as follows: both Jack and Jill believe that they both are rational (that they will be rational throughout), and that, in each round, they each face a problem of the Prisoners' Dilemma sort, and that there will be 100 rounds. Both also believe that they both believe this, that both believe that both believe it, that both believe that both believe that both believe it, etc. All this together is called the *Common Belief Assumption* (*CB*), strictly: the assumption of the commonality of their beliefs in their rationality and about the structure of their interaction.

We can now see that this is too strong.[13] CB implies that Jack now believes that he is rational (and will be throughout) and that in the last, 100th round he will be in a Prisoners' Dilemma. It follows from what he now believes that he there will choose to defect, for in a Dilemma only defection is rational. Being disciplined, Jack *believes* he will choose this. He believes that *now*. But then he *cannot* choose it -- adapting (36): $B_1(C_{100}x) \supset \sim C_{100}x$. So also for every previous round; he can never choose to defect. By the same argument, neither can Eve.

This undermines our scenario. For the basic Dilemma premise is that Jack and Jill have options, that they each must choose. Given CB (and their both being disciplined), this basic premise cannot stand. Each of them knowing what they will do, there is no issue for either of them. Neither has any choice to make:

they are in no dilemma. If we want to endorse CB, we must give up the Dilemma story. We can't use CB in discussing dilemmas.

The remedy is clear: we need a suitably weaker thesis, a thesis that, unlike CB, ascribes (imposes) no improper beliefs. Say that *s* is some set of propositions. Suppose now that each person involved believes every item in *s* that he can *properly* believe (in our special sense of *propriety*), that each believes this of all the others, that each believes that all the others believe it of all the others, etc. Call this the *Mutual Belief Assumption* (*MB*), strictly: the assumption of the mutuality of these people's beliefs in the items of *s*. (CB with regard to these items implies MB with regard to them, but not vice versa.)

Where *s* is as above, Jack and Jill's mutual belief implies that Jack believes that Jill is rational (plus other facts about her). It implies that Jill believes that Jack is rational (plus other facts about *him*). It implies that Jack believes that Jill believes that Jack is rational, that Jill believes that Jack believes that Jill is rational, etc. It does *not* imply that either believes that he (she) himself (herself) is rational, or that the other believes himself (herself) to be rational, etc. No improprieties here, or ascriptions of them to others (or any ascriptions of any beliefs whatever to oneself or of any self-ascriptions of any beliefs to others...). But note that this suffices for the backward induction, for the outside observer's argument to the joint-defection prediction. The observer's backward induction doesn't need CB. It needs only the weaker MB -- and the observer/predictor's beliefs about both parties being rational, etc.[14]

This thinking can be extended. CB is often said to be essential to the theory of games, or at least to justifications of equilibrium solutions. If that were right, it would mean trouble, for it would mean that game solutions are justifiable only where they aren't solutions, where the parties have no options, where there is no game to play. I suggest that it isn't right, that the full CB is more than game theory needs, that the jobs it is asked to do can all be done by MB. Binmore and Brandenburger say that "any equilibrium notion that incorporates some measure of self-prophesying necessarily entails common [belief] requirements..." (1990, p. 106).[15] I am saying that equilibrium analysis calls just for *other*-prophesying, that a player isn't also (can't be!) the observer/ predictor of the game. The moral here is the same as above: the logic of proper, defensible belief denies us nothing we need, but neglecting that logic invites an undermining confusion.

The moral is not one for game theory only, for CB is sometimes endorsed in other contexts too. Rawls (1971) proposes what he calls a *publicity* condition. Initially, he speaks of that in connection with a system of rules: "A person taking part in an institution knows what the rules demand of him and of the others. He also knows that the others know this and that they know that he knows this, and so on" (Rawls 1971, p. 56). Still, he later goes beyond rules: "for the most part, I shall suppose that the parties [behind the veil of ignorance] possess all general information" (1971, p. 142).[16] This means that everyone knows that everyone has *x*, *y*, and *z* as options, options distributing primary goods in certain specified ways, that everone

prefers more such goods to less, and that everyone is a maximinner. If everyone had this information, everyone could predict his own choice, which isn't logically possible (it isn't possible for *any*one). Better: there would be no choice to make. There would be too much knowledge behind the veil of ignorance to allow for any choosing.

No problem here for Rawls, for he retreats from his supposition, though only on the grounds of the 'complexity' of some of the information behind the veil. I am arguing there are *logical* grounds for backing off from publicity, from CB-publicity, that if Rawls assumed it, he would undermine his own theory. Fortunately, he doesn't need it, he doesn't need to suppose that everyone behind the veil is that well informed. (He doesn't even need the weaker, MB-sort of publicity.)

Once more, the moral here. There is no purpose for which we need to hold any improper beliefs, and none for which we must suppose that other people hold any. And we can mess up our thinking badly if we ascribe such beliefs to people.

## IV

Let me briefly comment on some ramifications. We have spoken of the skeptic who believes nothing whatever. He can't believe this about himself -- can't *properly* believe it -- for if he believed it, it wouldn't be true. Consider now his sister, who has joined a cult that requires its members not to *want* anything. Can she want to comply with that? If she wanted not to want anything, her wanting this would defeat what she wants. I will say she can't *properly* want it: if it is true (that she

wants nothing), she doesn't want it (for she wants *that*), and if she wants it, it isn't true. Another way of saying this: if she did want it and wanted to want it, she would have to want a contradiction.

Take also the teacher who announces to her students that there will be a test on Wednesday and that its being on that day will be a disappointment to them. This means there will be a test on Wednesday and that the students will want beforehand (say on Tuesday) that it *not* be on Wednesday. Can the students *want* this to be true? Can a person ever want to be disappointed by a specified *x*? I have argued that no one can properly believe he will be surprised by *x*. A parallel argument (with *wanting* in place of *believing*) shows that no one can properly want to be disappointed by *x*: if he will be disappointed by it, he doesn't want that disappointment, and if he wants it, he won't be.[17] Just as we can't predict our surprises, so we can't hope for our disappointments.

What about desire publicity? Are the desire-analogues of CB and MB of any philosophical interest? The ethics of Kant comes to mind. Kant presented a formal criterion of the desires we could morally act on, one that abstracted from our actual situations and from those of all other people. His Categorical Imperative is this: "Act only on that maxim whereby thou canst at the same time will that it should become a universal law" (Kant 1949 [1785], p. 38), a *maxim* being a desire to take a certain action described as the agent *sees* or *understands* that action. The desires a person acts on *are* his maxims, so we might put it like this -- in terms of honoring *desire* commonality --

that no one ought ever to act on a desire that couldn't be held by all.[18] Still, nothing whatever would follow from that: there is no desire that couldn't be held by all.

But perhaps this sells Kant short. Perhaps he means that the maxim you act on must be one that could now be implemented by all logically disciplined people. This would move us to desire- *mutuality*[19] and to this different version of the Imperative, that no one ought ever to act on a desire that couldn't be *properly* held by all. If Jill couldn't properly want *x* true, Jack oughtn't to act on his wanting *x*. This again uses "properly" in our special, logical sense -- for Kant, morality looks only to logic. (The Imperative itself isn't logic, but its exclusions depend just on logic.)

Here is an appealing corollary: never act on any desire to disappoint someone by *x*. We are led to this by the fact that whoever you want to disappoint can't properly want to be disappointed by it. (The fact emerges in the proof of the impropriety of wanting that, so the 'fact' is mere logic.) This does not say we may never disappoint any person. We can't avoid disappointing people; we do that wherever we do what anyone wants us not to be doing. It says we ought never to act *on the maxim* of disappointing someone (by *x*), that we may never make that our purpose, never may set our minds on just that. Are there other such corollaries? Not as things now stand with us, but let us move a step further.

We have taken people to be deductively thorough with regard to beliefs and also with regard to desires. Let us say that being disciplined also implies a thoroughness with regard to

beliefs and desires together, though this has to be hedged a bit to keep it from being too strong. Let us here put it this way, that if a person is disciplined and the conjunction of all he believes is *h* and the conjunction of all he wants is *k*, then if *m* follows from *h*-and-*k* and no conjunct of *h* alone follows from *k*-and-*m*, he also wants *m*.[20] (If disciplined Jack wants the Democrats to win and he believes that their candidate is Jones, he wants Jones to win.) We now have another corollary, this one perhaps unexpected.

Say that Jill will, on Wednesday, be led to falsely believe not-*x*. And suppose she now, on Monday, wants this to happen then. Here we have

(43) $x \cdot B_w \sim x$

(44) $W_m(43)$

By desire-thoroughness and retentiveness, we have

(45) $W_w x$ (44)

Jill will on Wednesday want *x* (45) and will believe ~*x* (43). Since everything follows from *x*·~*x*, ~(43) follows. Belief-plus-desire thoroughness here gets us

(46) $W_w \sim(43)$ (45) and (43)

By desire-retentiveness, we have

(47) $W_w(43 \cdot \sim 43)$ (46) and (44)

Since by desire-noncontradiction, (47) is false, we have, by reductio,

(48) ~(43·44) and also

(49) $(43) \supset \sim W_m(43)$ and

(50) $W_m(43) \supset \sim(43)$

These last lines say that Jill can't properly want to be deceived about *x*. Thus the Categorical Imperative implies that Jack ought not to try to deceive her: he ought not to lie to her. Kant would have said this is paydirt. He wanted to establish that lying is wrong -- not that lying itself is wrong, but that an action taken on a lying-maxim is wrong. Those who agree with that principle can take this last proof as a vindication of Kant. Those who *dis*agree can take it as a reductio of the Categorical Imperative, of the injunction to honor desire mutuality, to never act on any desire that couldn't be properly held by all.

REFERENCES

Basu, Kaushik, "Information and Strategy in Iterated Prisoner's Dilemma," *Theory and Decision* 8 (1977).

Bicchieri, Cristina, *Rationality and Coordination*, Cambridge, 1993

Binmore, Ken and Adam Brandenburger, "Common Knowledge and Game Theory," in Binmore's *Essays on the Foundations of Game Theory*, Blackwell, 1990.

Hintikka, Jaakko, *Knowledge and Belief*, Cornell, 1962.

Kant, Immanuel, *Fundamental Principles of the Metaphysic of Morals*, Liberal Arts Press, 1949.

Pettit, Philip and Robert Sugden, "The Backward Induction Paradox," *Journal of Philosophy* 86 (1989).

Piller, Christian, Review of Frederic Schick, *Understanding Action*, *Erkenntnis* 41 (1994).

Popper, Karl, "Indeterminism in Quantum Physics and in Classical Physics," *British Journal for the Philosophy of Science* 1 (1950).

Quine, Willard Van Orman, "On a So-Called Paradox," *Mind* 62 (1953); reprinted in his *The Ways of Paradox*, Random House, 1966.

Rawls, John, *A Theory of Justice*, Harvard, 1971.

Schick, Frederic, "Self-knowledge, Uncertainty, and Choice," *British Journal for the Philosophy of Science* 30 (1979); reprinted in Peter Gärdenfors and Nils-Eric Sahlin (eds.), *Decision, Probability, and Utility*, Cambridge, 1988.

--------, *Understanding Action*, Cambridge, 1991.

--------, *Making Choices*, Cambridge, 1997.

1. Imagine a second subscript to the "B"s identifying the believer -- in this first case, the student.

2. We need this in moving from (9) and (10) to (11) and also in moving from (15) to (16).

3. This simpler case is in Quine (1953), as is the line I take just above.

4. This second exam situation recalls Moore's story of the person who says it is raining but that he doesn't believe it.

5. In the second situation, we are assuming discipline in the move from (22) to (23).

6. For Hintikka, being disciplined doesn't extend to being belief-retentive.

7. The proof would need some extra moves from B~*x* to ~B*x*, but such moves are warranted by noncontradiction.

8. For more on optionality, see Schick (1997), pp. 8-11.

9. Strictly, it must have been an option for me *as I understood it*; see Schick (1997), pp. 11-20. I won't press this refinement here, but we need it for getting around some objections.

10. We need deductive thoroughness to move from (30) and (32) to (33). We need belief-retentiveness to move from (31) to (32).

11. A person can't predict his choices only *in the terms in which he sees his options*; I note this in Schick (1979). Romeo may be able to predict that he will choose as Juliet would have chosen or as she now wants him to -- provided he can't infer from that which of his options *as he sees them* he will choose.

12. Some improper beliefs are not self-referential at all. Think of the anti-mentalist theorist who believes that there *are* no beliefs.

13. It has been noted that the "etc." goes too far: only 99 iterations of "both believe that..." are needed in a 100-round case. My point here is different.

14. It may be that even MB is too strong; this was argued by Basu (1977) and it is implicit in the critique of CB in Pettit and Sugden (1989) and Bicchieri (1993). If MB is too strong, the backward induction fails, but not now because the parties have no issues to resolve.

15. Binmore and Brandenburger are speaking not of common belief but of common *knowledge*, and there mutuality implies commonality. Since what is known must be true, if Jack knows that Jill *knows* he is rational, Jack himself knows he is rational. But game theory needs only to talk of beliefs.

16. Their all having that information is then itself a general datum, and they have to believe *that*, etc.

17. The argument appeals to a desire-extension of our concept of discipline. Deductive thoroughness carries over smoothly, and so does noncontradiction, but retentiveness needs some rethinking. (In default of what sort of new data must our desires be stable -- what is *relevance* here?)

18. This refers to commonality of degree 1, not to any CB-like iterations.

19. To mutuality of degree 1.

20. The "no conjunct of *h*..." proviso is meant to avoid the implication that a disciplined person wants whatever he believes. The proviso in Schick (1991) is too strong; this was noted by Piller (1994).