*JOURNAL OF PHILOSOPHICAL RESEARCH*
VOLUME 34, 2009

# RELIABILISM AND PRIVILEGED ACCESS

**KOURKEN MICHAELIAN**
INSTITUT JEAN-NICOD (CNRS-EHESS-ENS)

**ABSTRACT:** Reliabilism is invoked by a standard causal response to the slow switching argument for incompatibilism about mental content externalism and privileged access. Though the response in question is negative, in that it only establishes that, given such an epistemology, externalism does not rule privileged access out, the appeal to reliabilism involves an assumption about the reliability of introspection, an assumption that in turn grounds a simple argument for the positive conclusion that reliabilism itself implies privileged access. This paper offers a two-part defense of that conclusion: the reliabilist account of privileged access is defended both against arguments in favor of the rival content inheritance strategy and against an argument turning on empirical considerations concerning the individuation of the belief-producing process of introspection.

## I. INTRODUCTION

There has of late been much debate[1] over the implications of mental content externalism,[2] by now a virtual orthodoxy in the philosophy of mind, for the pretheoretically plausible epistemological doctrine that subjects have privileged access (of some sort) to their own mental states.[3] While portions of the literature in which this debate has been conducted are relevant to the concerns of this paper, and while content externalism serves as a fixed point of the paper, my focus here is not on the question of the implications for privileged access of externalism in the philosophy of mind, but rather on the less frequently posed question of the implications for privileged access of externalism in epistemology;[4] my focus, in particular, is on the question of the implications of reliabilism,and specifically of process reliabilism, for the doctrine of privileged access.[5]

Those discussions of the question that have occurred usually have not made contact with the potentially relevant empirical (or empirically informed) literature.[6] In this paper, I attempt to bring to light the relevance to the question of two broad streams of such literature: on the one hand, there is the now vast psychological literature concerning the systematic and pervasive errors committed by normal subjects when attempting to introspect their own mental states;[7] on the other hand, there is the equally large literature in which the debate between theory theorists and simulation theorists over the nature of the mechanism(s) subserving our "mindreading" capacities has been conducted.[8] No serious survey of either body of literature is feasible here; my goal, rather, is the modest one of determining what sorts of questions reliabilists concerned about privileged access should bear in mind as they encounter the literature. It will, nevertheless, prove possible tentatively to say something about the implications of reliabilism for privileged access, given the current results from the relevant subfields of psychology and cognitive science.

The thesis of this paper is that, while process reliabilism is incompatible with certain strong forms of privileged access, it (probably) is not only compatible with, but even implies, a certain weak form of access.[9] The variety of privileged access implied by process reliabilism, while no doubt weaker than some would like,[10] should satisfy most contemporary philosophers (or at least those of a naturalistic orientation). Just as the results of the ongoing debates over content externalism and privileged access should be of interest both to content externalists (since these potentially constrain what they consistently can say about privileged access) and to those committed to some more or less definite view about privileged access (since they potentially constrain their general philosophy of mind), even such a qualified answer to the question of reliabilism and privileged access should be of interest both to reliabilists (since it limits the range of privileged access theses they consistently can endorse) and to those wedded to some particular privileged access thesis (since it constrains their general epistemology).

The plan of the paper is as follows. The remainder of §I is devoted, first, to stating some relevant privileged access theses, ranging in strength from radical Cartesianism (a view, I believe, that now has no proponents),[11] through intermediate Cartesianism (roughly the position of Gertler 2002), to modest Cartesianism (a view more congenial to contemporary naturalistic sensibilities), and second, to reviewing the standard versions of process reliabilism and counterfactual or truth-tracking reliabilism discussed in subsequent sections.

§II provides an overview of the debate over the compatibility of content externalism and privileged access: after setting out a standard "slow switching" argument for incompatibilism, I discuss a standard causal response to it. My interest here is in neither of these arguments as such, but reviewing them enables me to isolate the key claim about the reliability of introspection on which the causal strategy depends. If this claim is correct, then a straightforward case can be made for the compatibility of process reliabilism and modest Cartesian access.

In §III, I discuss an argument of Gertler's that, were it successful, would establish that reliabilism cannot accommodate privileged access, even given the key claim of the causal strategy. I concede that Gertler's argument establishes that reliabilism cannot accommodate intermediate Cartesian access, but argue that her favored alternative content inheritance strategy fares no better with respect to intermediate Cartesianism, and indeed that there is good reason to reject the latter view.

Intermediate Cartesianism, then, poses no threat to process reliabilism. But considerations made evident by the generality problem for reliabilism might yet mean that the theory is unable to accommodate even modest Cartesian access. In §IV, after reviewing the generality problem, I construct an argument from a set of psychological results for the conclusion that introspection is an unreliable belief-producing process; the implication of this conclusion is that process reliabilism will not underwrite even modest Cartesian access.[12] I go on to argue, however, that, given that one of a range of current theories of mindreading (including versions of the theory theory and the simulation theory, as well as a recent hybrid view) will turn out to be correct, there is reason to individuate the belief-producing process of introspection in such a way that it turns out, after all, to be reliable. I conclude that there is reason, at present, to suppose that process reliabilism implies modest Cartesian access.

## A. ACCESS

Before stating the privileged access theses discussed in §§II–IV below,[13] I want to say something to allay the worry that, because even the strongest of these posits only quite a weak or narrow variety of privileged access, the question which of them is compatible with reliabilism is of rather limited interest. I restrict my attention to narrow forms of access precisely because any plausible privileged access thesis will have to be narrow—some such form of privileged access, that is, is the most for which any of us is entitled to hope.

The forms of privileged access considered here are narrow, first, in that they are forms of infallibility.[14] In addition to infallibility, 'privileged access' has sometimes been used to refer to various other relations subjects have been supposed to have to their own mental states; besides infallibility, the standard list includes indubitability (where *S*'s belief that *P* is indubitable, roughly, just in case it is impossible that there be evidence that would justify *S*'s rejection of her belief that *P*), incorrigibility (where *S*'s belief that *P* is incorrigible just in case no one (else) who knows that *S* believes that *P* warrantedly can challenge that belief), and transparency (where *S*'s belief that *P* is transparent just in case if *P*, then *S* believes that *P*).[15] To survey here the various available forms of indubitability, incorrigibility, and transparency is neither possible (since there are too many)[16] nor necessary (since contemporary discussions focus largely on forms of infallibility).

The privileged access theses considered here are narrow, second, because they explicitly concern properly epistemic relations between subjects and their own mental states. 'Infallibility' is sometimes[17] used to refer to the relation described by (INF*).

(INF*) *S*'s belief that *P* is infallible if and only if, if *S* believes that *P*, then *P*.

(INF*) describes a relation of perfect reliability; since the question I am interested in here is an epistemological one, I will instead be concerned with relations of the sort described by (INF).[18]

(INF) *S*'s belief that *P* is infallible if and only if, if *S* believes that *P*, then *S* knows that *P*/*S* is justified in believing that *P*.

The forms of infallibility considered here thus are epistemic relations, not mere reliability relations. Even given some sort of reliabilism, the distinction matters: while something in the neighbourhood of INF-infallibility, given reliabilism, entails something like INF*-infallibility,[19] INF*-infallibility, even given reliabilism, does not straightforwardly entail INF-infallibility.[20]

A third sense in which the privileged access theses considered here are narrow is that they pertain only to beliefs resulting from introspection, where introspection is some special, non-observational belief-producing process.[21] Theses of privileged access have been propounded which consist of conjunctions of claims about subjects' infallibility and claims about the introspective availability of mental states; e.g:

(PA*) (i) *S*'s beliefs about her own mental states are infallible, and (ii) *S* can introspect her own mental states.

The theses considered here, however, resemble (PA).

(PA) (i) *S*'s beliefs about her own mental states, if produced by introspection, are infallible, and (ii) *S* can introspect her own mental states.

The reason for the restriction in clause (i) is simply that theses, like (PA*), consisting simply of the conjunction of an infallibility claim and a claim about the availability of mental states to introspection are subject to certain obvious counterexamples (involving, e.g., beliefs about their mental states formed by subjects not via introspection, but instead on the basis of unreliable testimony). Clause (ii) is retained in order to give the thesis some teeth: without the clause, the thesis would be compatible with the possibility that, as a matter of fact, no subject ever has privileged access to her own mental states (because no subject can ever introspect).

A further sense in which the relevant privileged access theses are narrow is that they pertain only to those introspective beliefs produced when the mechanism subserving introspection (whatever its nature) is functioning properly; they thus avoid counterexamples turning on various sorts of interference with, or malfunctioning of, the normal process of introspection. The final sense in which the theses are narrow is that they concern only subjects' access to their current mental states: any thesis on which subjects are (even approximately) infallible about their past mental states, too, is subject to various straightforward counterexamples.[22]

The strongest form of access consistent with these restrictions is that posited by radical Cartesianism:

(RC) Necessarily, every subject *S* is such that (i) if *S* believes that she has current mental state *M*, her belief was produced by introspection, and

> the mechanism subserving that process was functioning properly, then *S* knows that she has *M*, and (ii) *S* sometimes has introspective access to her current mental states.[23]

There is an obvious question regarding the range of '*S*' here (and throughout); it should suffice to say that '*S*' ranges over "rational, cognitively well-developed persons" (Gertler 2003a, xi–xii). Though no one any longer endorses a privileged access thesis as strong as radical Cartesianism, it is nevertheless useful to have the thesis before us, as the theses discussed below fruitfully can be thought of as weaker descendants of the view.

Bearing this in mind, three features of radical Cartesianism merit emphasis. First, the thesis is a necessity claim: the radical Cartesian supposes both that in all possible worlds the relevant subjects are (in the specified sense) infallible and that in all possible worlds those subjects can at least sometimes introspect their own current mental states. Second, radical Cartesianism is consistent with the possibility that some belief-producing process other than introspection is, with respect to the production of beliefs about one's mental states, epistemically just as good as introspection; it counts as a thesis of privileged access simply because it says that introspection is especially epistemically good. Finally, the radical Cartesian is not, in virtue of clause (ii) of (RC), committed to any particular view about the workings of the process of introspection or of the mechanism(s) subserving it: the clause simply expresses the thought that subjects (sometimes) can know their own current mental states directly or non-inferentially; such an ability is compatible with various sorts of subpersonal processing, only ruling out the possibility that the sole basis of our knowledge of our own current mental states is inference from (inter alia) behavioral evidence of the sort we rely on when forming beliefs about the mental states of others. The first and third of these features are inherited by intermediate Cartesianism, and the second and third are inherited by modest Cartesianism.

A series of modifications turns radical Cartesianism into a more plausible and interesting thesis. While the suggestion that subjects are infallible with respect to their own current phenomenal states is somewhat plausible, the infallibility theses of interest in the context of a discussion of the implications of reliabilism for privileged access pertain to propositional attitudes only. And since even current propositional attitudes presumably are dispositional, it is necessary in order to bring a thesis of privileged access to propositional attitudes up to the level of plausibility to restrict it to occurrent propositional attitudes only. It has become usual further to restrict the scope of infallibility theses to access to attitude contents: as Gertler (2002, 126) points out, attitude-modes themselves should perhaps be understood dispositionally, and so restricting infallibility in this way enables us to circumvent potential worries about the possibility of introspecting attitude-modes, so understood.

Modifying (RC) appropriately produces (RC*).

> (RC*) Necessarily, every *S* is such that (i) if *S* believes that she has some occurrent propositional attitude with content *C*, her belief was produced

by introspection, and the mechanism subserving that process was functioning properly, then *S* knows that she has an occurrent propositional attitude with *C*, and (ii) *S* sometimes has introspective access to the contents of her occurrent propositional attitudes

(RC*) is in two ways still stronger than is wanted. First, there is a need, given reliabilism, to have a contingent access thesis on the table.[24] Second, given that privileged access is a contingent matter, a strict infallibility claim along the lines of (RC*) becomes implausible: since knowledge entails truth, clause (i) of such a thesis would entail that if *S* satisfies the appropriate conditions, then in fact she has an occurrent propositional attitude with the relevant content, i.e., that (properly functioning) introspection is perfectly reliable with respect to occurrent propositional attitude contents; in order to allow for the possibility that introspection is strictly fallible, it is more plausible to suppose more weakly that, if *S* satisfies the appropriate conditions, then she is justified in believing that she has an occurrent propositional attitude with the relevant content.

Making the appropriate modifications to (RC*) gives us modest Cartesianism:

(MC) Every *S* is such that (i) if *S* believes that she has some occurrent propositional attitude with *C*, her belief was produced by introspection, and the mechanism subserving that process was functioning properly, then her belief is justified, and (ii) *S* sometimes has introspective access to the contents of her occurrent propositional attitudes.

On modest Cartesianism, that a higher-order belief about one's own occurrent propositional attitude contents results from a process of (malfunction-free) introspection does not guarantee the truth of the belief. It is thus potentially misleading to refer to modest Cartesianism as an infallibility thesis, but, since the view is a recognizable descendant of radical Cartesianism, and for ease of exposition, I will continue so to refer to it.

Before reviewing reliabilism, I want to remark on two features of the—to my mind—curious compromise between radical and modest Cartesianism endorsed by Gertler (2002). On Gertler's view, first, though that a higher-order belief results from introspection does not guarantee its truth, nevertheless (fallible) privileged access is necessary. Second, while radical and modest Cartesianism count as theses of privileged access simply in virtue of positing that subjects have a special, non-observational method of acquiring information about their own mental states, intermediate Cartesianism claims also that this method (viz., introspection) necessarily is epistemically better than any alternative route to knowledge of the relevant mental states. This hybrid view, I will take it, can be formulated as follows:[25]

(IC) Necessarily, every *S* is such that (i) if *S* believes that she has some occurrent propositional attitude with *C*, her belief was produced by a process of introspection, and the mechanism subserving that process was functioning properly, then her belief is both justified and better justified

than is any type-identical belief held by *S*' (*S* ≠ *S*'), and (ii) *S* sometimes has introspective access to the contents of her occurrent propositional attitudes.[26]

Intermediate Cartesianism is, at any rate, the only compromise between radical and modest Cartesianism I will entertain here. Note that, unlike her radical and modest counterparts, the intermediate Cartesian is committed to the view that, when it comes to higher-order beliefs about the contents of one's own occurrent propositional attitudes, no belief-producing process is epistemically as good as introspection.

## B. RELIABILITY

Standard defences of content externalism against slow switching arguments for its incompatibility with privileged access (more or less explicitly) invoke reliabilism. Because of this, and because[27] different reliabilisms might have different implications for privileged access, I pause at this point quickly to distinguish between process reliabilism (Goldman 1979) and counterfactual reliabilism (the truth-tracking theory) (Nozick 1981).[28]

Where *S* tracks the truth with respect to *P* just in case *S* would believe that *P* if it were that *P* and *S* would not believe that *P* if it were not that *P*, (CR) is a standard version of counterfactual reliabilism.

(CR) *S* knows that *P* if and only if *P*, *S* believes that *P*, and *S* tracks the truth with respect to *P*

Since, in order for a belief that *P* to amount to knowledge, counterfactual reliabilism requires that the believer be reliable just with respect to *P*, counterfactual reliabilism might be thought of as a sort of local reliabilism. There are various well-known problems for counterfactual reliabilism,[29] but, since I am interested in the theory just to the extent that it "overlaps" with process reliabilism, I ignore these here.

Process reliabilism, in contrast to counterfactual reliabilism, might be thought of as a sort of global reliabilism, since it requires, in order for a belief that *P* to be justified, not that the believer track the truth with respect to *P*, but instead that the process by which the belief was produced be reliable. Where a belief-producing process is reliable, roughly, just in case it would tend to produce more true than false beliefs, (PR) is a standard formulation of process reliabilism.

(PR) *S*'s belief that *P* is justified if and only if it is produced by a reliable process.

The chief problem for process reliabilism is the generality problem (Feldman 1985; Conee and Feldman 1998), to which I return in §IV.A below.

Whereas counterfactual reliabilism is a theory of knowledge, process reliabilism is a theory of justification. The source of this difference between the theories is the concept of reliability at work in (PR), which is, in a sense, both more generous and more stringent than that invoked by (CR). Process reliability is more generous

in that it does not require that a believer track the truth with respect to *P* in order for her to be justified in believing that *P* (and, if *P* is true and un-Gettierized, so to know it); it is more stringent in that it requires that the process by which her belief that *P* was produced be reliable with respect to some broader range of propositions in order for the believer to be justified in believing that *P* (and so, potentially, to know that *P*).

Two consequences of this difference between the theories are salient here. Suppose, on the one hand, that introspection is unreliable with respect to the relevant range of propositions; even so, if it is reliable at least with respect to occurrent propositional attitude contents, then counterfactual reliabilism, but not process reliabilism, might be able to accommodate modest Cartesian access.[30] Suppose, on the other hand, that introspection is reliable, but not perfectly reliable, with respect to the relevant range of propositions; then process reliabilism will be able to accommodate modest Cartesian access, but counterfactual reliabilism might still be unable to do so (since some introspective beliefs about occurrent propositional attitude contents might still fail to track the truth).[31]

## II. CONTENT

With this background on privileged access and reliabilism in place, I turn to the question of the implications for privileged access of process reliabilism. It will be useful, as a first step in working those implications out, to review the debate between compatibilists and incompatibilists about content externalism and privileged access.

Mental content externalism (which I accept) is the view that

(E) A subject's mental content supervenes not on her intrinsic properties alone, but only on these together with her relational properties.[32]

Given that we accept content externalism, there remains the question whether we should in addition to the wide (externalist) content described by (E) acknowledge also a narrow (internalist) variety of content.[33] Even should we acknowledge both wide and narrow content, however, it is clear that the sort of content relevant to privileged access is wide: if I have privileged access to my own occurrent thought contents, then presumably I have such access to the wide content of, e.g., my thought that water is wet.[34]

Boghossian (1989) and others have argued that (E), together with the possibility of slow switching, rules privileged access out. I set out, in §II.A, a standard version of this argument for incompatibilism and, in §II.B, a standard response to it; though my concern here is not with the compatibilist/incompatibilist controversy per se, reviewing the debate enables me to isolate an assumption about the causal sources of higher-order beliefs about one's own occurrent propositional attitude contents that is key to the standard response. If correct, this assumption seems to imply that reliabilism entails modest Cartesian access (see §II.C); I show in §IV.B, however, that things are not quite so simple.

## A. SLOW SWITCHING

McLaughlin and Tye (1998, 351–352), drawing on Boghossian (1989),[35] state the slow switching argument as follows:

> Oscar, without his knowledge, has been traveling back and forth between Earth and Twin Earth. With each move, he stays long enough to acquire the concepts of the locals. [I.e., he is slowly switched.] So, when he utters the sentence 'Water is a liquid' on Twin Earth, he comes to express the thought that *twater* is a liquid, just as the members of the indigenous population do. Oscar is, however, completely unaware that such shifts in his thought and speech occur.
>
> The difficulty this switching case is alleged to present for compatibilism is the following. Suppose that on a particular occasion while residing on Earth, Oscar occurrently thinks that water is a liquid. Given his travels, that *he is thinking that twater is a liquid* is a relevant alternative to *his thinking that water is a liquid*. The introspective evidence available to Oscar is compatible with its being the case that *he is thinking that twater is a liquid*; and, so, his introspective evidence does not exclude the relevant alternative. Thus, Oscar cannot know by introspection that he is thinking *that water is a liquid*.

Regimenting this reasoning and taking as the dual targets of the argument modest and intermediate Cartesianism produces the following standard formulation of the slow switching argument for incompatibilism.

(S1) Suppose that content externalism is true.

(S2) Suppose that Oscar has been slowly switched from Earth to Twin Earth and back.

(S3) Suppose that, after having been switched back to Earth, Oscar has an occurrent propositional attitude with the content [that water is wet].

(S4) Suppose that, via introspection (subserved by a properly functioning mechanism), Oscar then forms the second-order belief that he has an occurrent propositional attitude with the content [that water is wet].

(S5) The possibility that Oscar has an occurrent propositional attitude with the content [that twater is wet] is a relevant alternative to the fact that he has an occurrent propositional attitude with the content [that water is wet]. (from S1–S4, together with an account of the relevance of alternatives)[36]

(S6) Introspection does not allow Oscar to rule out the possibility that he has an occurrent propositional attitude with the content [that twater is wet]. (from S1–S4)

(S7) Oscar does not know that he has an occurrent propositional attitude with the content [that water is wet]. (from S5, S6, and an assumption about relevant alternatives)

(S8) If modest Cartesianism is true, then Oscar knows that he has an occurrent propositional attitude with the content [that water is wet]. (from MC, S3, S4, and the fact that the belief is not Gettierized)

(S9) If intermediate Cartesianism is true, then Oscar knows that he has an occurrent propositional attitude with the content [that water is wet]. (from IC, S3, S4, and the fact that the belief is not Gettierized)

(S10) Therefore, if content externalism is true, then modest Cartesianism is false. (from S7 and S8)

(S11) Therefore, if content externalism is true, then intermediate Cartesianism is false. (from S7 and and S9)

It might be objected—this would amount to a rejection of (S8)—that the standard slow switching argument does no damage to modest Cartesianism, since (MC) is a contingent thesis and the case of Oscar is merely hypothetical. Naturalistically oriented philosophers, in particular, might be inclined to attempt to counter the slow switching argument with something like the following line of reasoning:

> At worst, the slow switching argument demonstrates that subjects who have been slowly switched do not have privileged access to the contents of their own occurrent propositional attitudes. But that just shows that privileged access is not necessary. Since, as far we know, our world is not one in which slow switching occurs, we still have been given no reason to suppose that we ourselves do not have privileged access to the contents of our own occurrent propositional attitudes. And such (contingent) access is enough.[37]

The problem with this sort of objection to the slow switching argument is that it overlooks the possibility that we ourselves are sometimes slowly switched, that is, that we find ourselves in circumstances relevantly similar to those of Oscar. If there is reason to think that we ourselves are actually slowly switched (as Ludlow 1995 argues that we are), then, though the case of Oscar described by (S1)–(S4) is merely hypothetical, naturalists should want to say that, if we have privileged access to the contents of our own occurrent thoughts, then so does Oscar (they should treat the case as if it were actual). We should not have to wait to develop a view on privileged access until we have settled the question whether we ourselves are slowly switched; and so we had better take the case seriously. Naturalists (or at least those who are content externalists), then, should not succumb to the temptation just mentioned.

That the slow switching argument can be blocked, of course, tells us neither that modest Cartesianism is true nor that intermediate Cartesianism is true. It will turn out, however, that, given the soundness of the pure causal response to the argument, we can formulate a straightforward argument from reliabilism to modest (though not to intermediate) Cartesianism.

## B. CAUSATION

The pure causal response to the slow switching argument invokes reliabilism; and, though it does not matter to the logic of the response whether process reliabilism or counterfactual reliabilism is invoked,[38] the argument of §IV shows that process reliabilism at minimum raises an additional concern about the soundness of the response.

The response amounts to a purely causal account of self-knowledge, an account, that is, on which privileged access derives entirely from certain causal relations between introspected and introspective states.[39] The strategy is to grant (S1)–(S5) (content externalism, the description of the case of Oscar, and the inference from these to the relevance of his thinking that twater is wet), but to argue that (S6) (the claim that Oscar cannot introspectively rule out the possibility that he is thinking that twater is wet) does not follow from these assumptions.

(C1) Suppose that content externalism is true. (S1)

(C2) Suppose that Oscar has been slowly switched from Earth to Twin Earth and back. (S2)

(C3) Suppose that, after having been switched back to Earth, Oscar has an occurrent propositional attitude with the content [that water is wet]. (S3)

(C4) Suppose that, via introspection (subserved by a properly functioning mechanism), Oscar then forms the second-order belief that he has an occurrent propositional attitude with the content [that water is wet]. (S4)

(C5) Then the possibility that Oscar has an occurrent propositional attitude with the content [that twater is wet] is a relevant alternative to the fact that he has an occurrent propositional attitude with the content [that water is wet]. ((S5) from C1–C4, together with an account of the relevance of alternatives)

(C6) If *S* has an occurrent propositional attitude with content *C*, then, if *S* introspects, that attitude will reliably cause *S* to form the belief that she has an occurrent propositional attitude with *C*. (R)

(C7) Had Oscar had an occurrent propositional attitude with the content [that twater is wet], then, had he introspected, Oscar probably would have arrived at the belief that he has an occurrent propositional attitude with the content [that twater is wet]. (from C6)

(C8) If counterfactual reliabilism is true, then Oscar probably knows that he has an occurrent propositional attitude with the content [that water is wet]. (from CR, C1–C4, and C7)

(C9) If process reliabilism is true, then Oscar knows that he has an occurrent propositional attitude with the content [that water is wet]. (from PR, C1–C4, C6, and the fact that the belief is not Gettierized)

(C10) If counterfactual reliabilism is true, then the Oscar case does not show that if content externalism is true, then modest Cartesianism is false. (from C1–C4, C8, and MC)

(C11) If counterfactual reliabilism is true, then the Oscar case does not show that if content externalism is true, then intermediate Cartesianism is false. (from C1–C4, C8, and IC)

(C12) If process reliabilism is true, then the Oscar case does not show that if content externalism is true, then modest Cartesianism is false. (from C1–C4, C9, and MC)

(C13) If process reliabilism is true, then the Oscar case does not show that if content externalism is true, then intermediate Cartesianism is false. (from C1–C4, C9 and IC)

(C14) Counterfactual reliabilism or process reliabilism is true.

(C15) Therefore, the case of Oscar does not show that if content externalism is true, then modest Cartesianism is false. (from C10, C12, and C14)

(C16) Therefore, the case of Oscar does not show that if content externalism is true, then intermediate Cartesianism is false. (from C11, C13, and C14)[40]

If this pure causal strategy is successful, then the slow switching argument is blocked at the point of the inference to (S6): because lower-order thoughts, under appropriate circumstances, tend to cause higher-order thoughts with similar contents, introspection can in fact rule out the possibility that Oscar's lower-order attitude has the content [that twater is wet].[41]

Note that the conclusions (C15) and (C16) of the argument are negative, in that they say only that the slow switching argument fails to establish that content externalism is incompatible with the given forms of privileged access, not that content externalism in fact is compatible with either of those forms of access. In §III, I defend the pure causal strategy against an argument designed to show that it fails.

## C. INTROSPECTION

Aside from (C14) (which I will of course not challenge here), the key undefended assumption of the pure causal strategy is (C6) or (R). (R) is not only key to the success of the pure causal strategy, but also serves as one premise of a simple argument for the conclusion that reliabilism implies modest Cartesian access.[42] Since (R) says nothing about introspection being more reliable than relevant other belief-producing processes, no such argument is available for the conclusion that reliabilism implies intermediate Cartesianism.

(R1) If *S* has an occurrent propositional attitude with *C*, then, if *S* introspects, that attitude will reliably cause *S* to form the belief that she has an occurrent propositional attitude with *C*. (R)

(R2) Introspection is usually a locally reliable belief-forming process. (from R1 and the notion of local reliability)

(R3) Introspection is a globally reliable belief-forming process. (from R1 and the notion of global reliability)

(R4) Therefore, if counterfactual reliabilism is true, then modest Cartesianism is true. (from CR and R2)

(R5) Therefore, if process reliabilism is true, then modest Cartesianism is true. (from PR and R3)

In §IV, I consider a challenge to the inference from (R1) to (R3) on the basis of a worry that the process of introspection cannot be individuated so as to legitimate the move. Note that, if the inference to (R3) fails, then so does that to (C9), so that the pure causal strategy then also fails.

## III. INTERMEDIATE CARTESIANISM

Gertler has recently argued that, because "the relation between a thought content and an introspective state isn't merely a causal relation," the pure causal strategy fails (2002, 127). Her argument invokes intermediate Cartesian access; those who would be content with modest Cartesian access might thus in principle simply disregard the argument. It is, however, instructive to consider the argument, for, in doing so, we learn, first, that reliabilism cannot accommodate intermediate Cartesianism, but also that, absent an apparently innocent but ultimately implausible assumption, the content inheritance strategy favored by Gertler fares no better.

### A. ASYMMETRICAL ACCESS

Gertler's argument against the pure causal strategy appeals to a thought experiment (Fig. 1) involving Nick, an alcoholic, and Nora, a neuroscientist of the future and Nick's sponsor in Alcoholics Anonymous. Nora knows which of his brain states realizes Nick's thoughts about his favourite drink, gin, as well as which of her brain states realizes her thoughts about Nick thinking about gin. She programs this information into a monitoring device, and hooks it up to both Nick and herself. The device scans Nick's brain for the "gin" state and, whenever he enters that state, reliably causes Nora to form a belief that Nick is thinking about gin. It also scans Nora's brain for (precursors of) the "Nick is thinking about gin" state and, whenever she is about to enter that state, reliably prevents her from entering it until Nick is actually thinking about gin.[43] Suppose that Nick has a thought with the "gin" content, *C*1, so that the monitoring device causes Nora to have a higher-order belief that Nick has an occurrent propositional attitude with *C*1. Suppose also that Nick introspects (using a properly functioning mechanism), and as a result himself forms a higher-order belief that he has an occurrent propositional attitude with *C*1.

As Gertler points out,

> according to causal accounts of self-knowledge, Nick's access [to his gin thought] is exclusively explained by the causal process which resulted in his self-attributing belief. But Nora's causal access to Nick's gin thought appears to parallel Nick's access to it, epistemically speaking. . . . Nora's belief that Nick is thinking about gin appears to meet the epistemic requirements of reliabilism, since the device establishes an appropriate counterfactual-supporting link between Nick's gin thoughts and Nora's belief. And Nora's method of detecting Nick's gin thoughts is no more prone to error than is Nick's own method. (2002, 129–130)

If the pure causal strategy succeeds, then, it appears that Nora's higher-order belief that Nick has an occurrent propositional attitude with *C*1 is precisely as well justified as is Nick's own higher-order belief that he has an occurrent propositional attitude with *C*1.

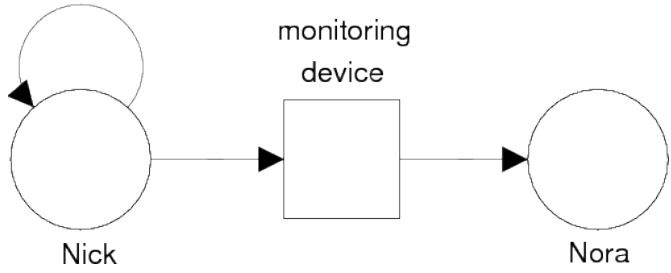


***Figure 1**: Gertler's thought experiment*

Taking aim at this implication of the pure causal strategy, Gertler gives an argument that can be regimented as follows.

(G1) Suppose that the case of Nick and Nora (described above) occurs.

(G2) Intermediate Cartesianism is true.

(G3) If intermediate Cartesianism is true, then Nora's second-order belief that Nick has an occurrent propositional attitude with *C*1 is less justified than is Nick's second-order belief that he has an occurrent propositional attitude with *C*1. (from G1 and IC)

(G4) Nora's second-order belief that Nick has an occurrent propositional attitude with *C*1 is less justified than is Nick's second-order belief that he has an occurrent propositional attitude with *C*1. (from G2 and G3)

(G5) If the pure causal response to the slow switching argument is successful, then Nora's second-order belief that Nick has an occurrent propositional attitude with *C*1 is precisely as well justified as is Nick's second-order belief that he has an occurrent propositional attitude with *C*1. (from G1 and the pure causal response)

(G6)Therefore, the pure causal response fails. (from G4 and G5)

The appeal to intermediate Cartesianism (G2) is crucial here, since the implications of modest Cartesianism (and, for that matter, those of radical Cartesianism) with respect to the case of Nick and Nora are consonant with those of the pure causal response.

Now, (G5) is a claim about the implications of the pure causal strategy with respect to the case of Nick and Nora. The key undefended assumptions of that strategy are (R) and (C14) (that either counterfactual or process reliabilism is true). Nothing in the thought experiment compels us to reject (R).[44] Thus the only apparent way of understanding the argument is as being directed against (C14), that is, as being directed against reliabilism. Though Gertler does not explicitly say so, then, the argument (G1)–(G5) is to be understood as moving from a certain privileged access thesis (intermediate Cartesianism), via a thought experiment (the case of Nick and Nora), to the conclusion that reliabilism is false.

Unless we can find fault with the logic of the argument (and I can find none),[45] then, we are bound to admit that reliabilism cannot accommodate intermediate Cartesian access: of the privileged access theses considered here, modest Cartesianism is thus the strongest that is (perhaps) compatible with reliabilism.

## B. INHERITANCE

Its inability to accommodate intermediate Cartesian access, however, should not be taken to count against reliabilism, for, Gertler's argument to the contrary notwithstanding, the content inheritance strategy[46] she favors likewise is unable to accommodate such access.[47] In this section, I review Gertler's case for the content inheritance strategy, and, in §III.C, I argue that it fails.

On the content inheritance strategy, "the relationship between introspected and introspective states [is] not merely causal but, rather, [is] a relationship of *inclusion*: introspective thoughts *embed* or *contain* introspected thoughts" (Gertler 2002, 135). As Burge puts it,

> by its reflexive, self-referential character, the content of the second-order judgement is logically locked (self-referentially) onto the first-order content which it both contains and takes as its subject matter. (1988, 122)[48]

According to this approach, then, the content of an introspected lower-order thought is itself a proper part of the content of certain introspective higher-order thoughts about it. Thus, the key difference between the content inheritance strategy and the pure causal response to the slow switching argument is, I take it, that the latter replaces (R) with something like (C).

(C) If *S* has an occurrent propositional attitude with *C*, then, if *S* introspects, that attitude will reliably cause *S* to form the belief that she has an occurrent propositional attitude with *C*; the content *C* of *S*'s lower-order attitude is moreover itself a proper part of the content of *S*'s higher-order belief that she has an occurrent propositional attitude with *C*.

In slogan form: causation by introspection suffices for content inheritance.[49]

The content inheritance strategist goes on to argue that the fact that Oscar's higher-order belief that he has an occurrent propositional attitude with the content [that water is wet] inherits the relevant part of its content (viz., [that water is wet]) from his lower-order thought suffices to ground his privileged access to the content of that thought. I will not attempt to set out the remainder of the content inheritance strategy in any detail here: the idea, roughly, is that since, normally, when a subject introspects, the content of the resulting higher-order belief contains the content of the introspected lower-order thought as a proper part, introspective thoughts normally get the relevant parts of their content determined in precisely the wide, externalist way in which lower-order thoughts have their content determined, so that introspection does, contra the slow switching argument, permit subjects to know their own occurrent thought contents, even given content externalism.[50]

I am prepared to suppose that the content inheritance strategy, however, precisely, it is to be worked out, succeeds in defusing the slow switching argument, but I want to challenge Gertler's argument for the claim that the content inheritance strategy, unlike the pure causal strategy (that is, unlike reliabilism) can accommodate intermediate Cartesian access. The argument for this claim has two parts: the first is designed to establish that content inheritance entails a certain metaphysical disparity between self- and other-knowledge (e.g., between Nora's belief about Nick's "gin" thought and Nick's own belief about that thought); the second is supposed to show that this disparity is epistemically significant. I will not discuss the second part of the argument here, since I think that the first part can be blocked, and since if content inheritance secures no metaphysical disparity between self- and other-knowledge, we have been given no reason to suppose that it can secure an epistemic disparity between these two sorts of knowledge.[51]

The argument for the metaphysical disparity can be reconstructed as follows.

(G7) If the content of a lower-order thought is a part of the content of a higher-order thought about it, then, if *S* instantiates the higher-order thought, *S* instantiates the lower-order thought.

(G8) No occurrent thought (i.e., no thought-token) is instantiated by more than one person.

(G9) No higher-order thought of *S*'s about a lower-order thought of *S*"'s ($S \neq S$") could embed the content of any lower-order thought of *S*"'s. (from G7 and G8)

(G10) Nora's second-order belief that Nick has an occurrent propositional attitude with *C*1 does not (and could not) embed the content *C*1 of Nick's first-order thought. (from G9)

(G11) Nick's second-order belief that he has an occurrent propositional attitude with *C*1 does embed the content *C*1 of his first-order thought. (from C and the description of the case)

(G12) Therefore, there is a metaphysical disparity between Nora's second-order belief that Nick has an occurrent propositional attitude with *C*1 and Nick's second-order belief that he has an occurrent propositional attitude with *C*1. (from G10 and G11)

This argument is clearly valid (or, rather, easily could be made valid); I argue in the next section, however, that it is unsound.

## C. OVERLAPPING THOUGHTS

(G8), the claim that no occurrent thought content can belong to more than one subject, presumably amounts to something like (G8').[52]

(G8') If *C* is the content of a belief of *S*'s, then, while *C* might be part of the content of some other belief of *S*'s, it could not be part of the content of any belief of *S*''s ($S \neq S'$).

Gertler notes that this is a "highly plausible" assumption, but provides no explicit reason for endorsing it (2002, 138). (Note that nothing in logic of the proper part-hood relation mandates it: *C* might, as far as that logic goes, be a proper part of the contents of non-identical beliefs *B* and *B*', no matter whether the subjects to whom *B* and *B*' belong are identical. And note that (G8) does not follow from (C), which only states a sufficient condition for content inheritance, and so by itself tells us nothing about whether Nora's belief about Nick's "gin" thought inherits the content of the latter thought.) The assumption is indeed plausible, but there is reason to reject it.

In order to secure (G8), Gertler will have to reject the following plausible principle stating a general sufficient condition for content inheritance.

(A) If occurrent propositional attitude *A* causes beliefs *B* and *B*', the causal relations between *A* and *B*, on the one hand, and those between *A* and *B*', on the other hand, are (otherwise) epistemically equivalent, and *B* inherits the content of *A*, then *B*' also inherits the content of *A*.

In slogan form: appropriate causation suffices for content inheritance. The notion of one causal relation being epistemically equivalent to another admittedly requires some unpacking, but we can make sufficient sense of it to proceed here. The causal relation between Nick's lower-order attitude and his higher-order belief about that attitude and that between his lower-order attitude and Nora's higher-order belief about it, in particular, are (ignoring for the moment the matter of content inheritance and its epistemic significance) clearly epistemically equivalent. Hence, by (A), if Nick's higher-order belief inherits the content of his lower-order attitude, then Nora's higher-order belief inherits that same content; we might, then, say that their higher-order beliefs overlap.

It might be objected that, if we suppose that, in violation of (G8), thoughts can overlap in this way (that is, if it is not merely that distinct subjects might have type-identical thoughts, but rather that distinct subjects might actually share a thought),

then, given (G7) (which I do not dispute), we are committed to the strange view that, e.g., some of Nora's thoughts (that is, some of the thoughts she instantiates) are located outside of her head; what is worse, we then seem to be committed to the view that some of her thoughts (some of the thoughts she instantiates) are located in Nick's head.

Fortunately for the defender of (A), it is far from clear that it is problematic to say that Nora can instantiate a thought that is nevertheless located in Nick's head. Clark and Chalmers, inter alia, have argued for an "active externalism," on which some of a subject's beliefs can indeed fail to be located in her head, on which "*beliefs* can be constituted partly by features of the environment, when those features play the right sort of role in driving cognitive processes," so that "the mind extends into the world" (1998, 12).[53] If some of Nora's beliefs are in Nick's head, then we might say (in the jargon of Clark and Chalmers) that Nora and a part of Nick's brain compose a coupled system.

Now, the suggestion that there is a coupled system here is no doubt more outlandish than are the analogous claims about most of the cases discussed by Clark and Chalmers;[54] compare it, e.g., to the case of Otto the Alzheimer's patient who (arguably) stores some of his beliefs in a notebook. But perhaps it does not differ in kind. Clark and Chalmers draw our attention to several features which render it reasonable to suppose that the information in the notebook constitutes some of Otto's beliefs: first, "the notebook is a constant in Otto's life"; second, "the information in the notebook is directly available without difficulty"; and third, "upon retrieving information from the notebook he automatically endorses it" (1998, 17).[55] The relevant part of Nick's brain is similar to Otto's notebook in the first two of these three respects: in virtue of their respective connections to the monitoring device, it is normally available to Nora; and the thoughts in question are directly available to Nora without difficulty. Nora, of course, does not automatically endorse those of Nick's thoughts she reliably detects; but perhaps this only means that there is no clear answer as to whether Nora's brain and the relevant part of Nick's brain constitute a coupled system.[56] The objection, then, at least is not decisive.

Perhaps anticipating this line of argument, Gertler points out that Clark and Chalmers argue only that one's dispositional beliefs can be located outside one's head, whereas what is at issue here is whether occurrent beliefs can be located outside one's head; she claims that because she "[applies] the inclusion strategy to occurrent states only, this claim is immune to the Clark and Chalmers argument" (2002, 145).[57] But the claim is puzzling, for it would seem on the face of it that if one's dispositional beliefs might be located outside one's head, then so might one's occurrent beliefs—one's occurrent beliefs, that is, need not occur in one's head.

Consider again the case of Otto the Alzheimer's patient. If Clark and Chalmers are right, the information in the notebook constitutes part of the set of Otto's dispositional beliefs. Under what conditions does one of these dispositional beliefs become occurrent (for Otto)? Gertler, apparently, would take it that, in order for it to become occurrent for Otto, the belief must occur in his head. The supposition

is sensible enough, for it is natural to describe the situation along the following lines: when Otto wants to "call up," e.g., his belief about the location of his home, he opens his notebook to the page on which the relevant information is recorded, looks to see what it says, and then (in his head proper) forms the occurrent belief that his home is located at such-and-such coordinates. In effect, the information is copied from one location in the coupled system to another.

While natural enough, this description of the situation misses the point of treating Otto and his notebook as a coupled system. It would seem to be sufficient for a dispositional belief of Otto's, located in the notebook, to become occurrent for Otto that it become occurrent for the coupled system composed by Otto and his notebook, which would presumably be a matter of the notebook (with Otto looking at it) being opened to the page on which the relevant information is recorded. The system itself has the dispositional belief that Otto's home is located at such-and-such coordinates; in virtue of being (the principal) part of the system, Otto has the dispositional belief that his home is located at those coordinates. Similarly, if the system itself has the occurrent belief that Otto's home is located at such-and-such coordinates, Otto, in virtue of being the principal part of the system, has the occurrent belief that his home is located at those coordinates. No copying would seem to be required. Clark and Chalmers go so far as to write that "Otto *himself* is best regarded as an extended system, a coupling of biological organism and external resources" (1998, 18), that is, that Otto just is the coupled system; and if so, then it is trivially true that if a belief is occurrent for the system, then it is occurrent for Otto.

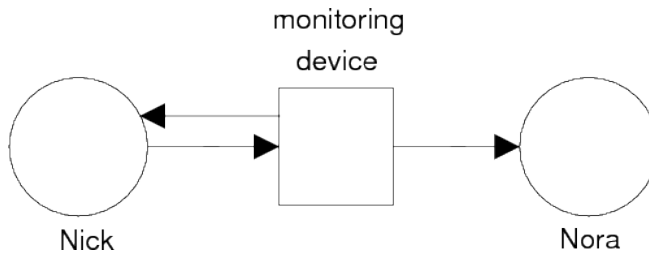


***Figure 2**: First variation on Gertler's thought experiment*

In order to see that Gertler is bound to reject (A), consider the following variation on the case of Nick and Nora (Fig. 2). Suppose, as before, that Nick and Nora are hooked up to the monitoring device in such a manner that Nick's having a thought with *C*1 reliably causes Nora to believe that Nick is having a thought with *C*1. Suppose, further, that Nora has effected an additional connection between Nick and the device, such that, when the device detects that Nick is having a thought with *C*1, it reliably causes him to believe that he is having a thought with *C*1, if it detects that he is not already entertaining that belief. (Were this the whole story,

Nick's higher-order beliefs about his "gin" thoughts would be overdetermined.) Suppose also that Nora has disabled the mechanism (whatever it might be) that subserves Nick's capacity for introspection. Suppose, finally, that Nick has a thought with the "gin" content *C*1, so that the monitoring device both produces in Nora a higher-order belief that Nick has an occurrent propositional attitude with *C*1 and produces in Nick himself a higher-order belief that he has an occurrent propositional attitude with *C*1.

Gertler will have to say that, in this variant scenario, not only does Nora's higher-order belief not inherit the content of Nick's lower-order attitude, but also Nick's own higher-order belief does not inherit the content of that attitude, for Nick and Nora are, in all relevant respects, symmetrically situated with respect to the target attitude. But the only salient difference between this case and the original of which it is a variant is the absence of introspection (or, rather, of the mechanism normally subserving Nick's introspective capacity):[58] the causal relations between Nick's higher-order belief and his lower-order attitude in the two cases are epistemically on a par (the question of content inheritance aside); Nick has precisely the same evidence available to him in both cases, and so on.[59] In short, it seems that, on penalty of incoherence, Gertler must maintain that causation by introspection is a necessary condition for content inheritance.

Combining this claim with (C) produces the view that, roughly, causation by introspection is both a necessary and a sufficient condition for content inheritance. But the view is highly implausible. Consider, e.g., a world just like ours, but for the fact that a mischievous (and nearly omnipotent) counterpart of Nora's has removed the mechanisms subserving the introspective capacities of its inhabitants and connected each of those inhabitants to a monitoring device in the manner in which Nick is connected to the device in the second of the thought experiments described above (Fig. 3). (Suppose, if you like, that the monitoring devices are quite small and are implanted in the inhabitants' skulls.) If introspection is necessary for content inheritance, then the world just described is a world without content inheritance. But to accept this is to make introspection into a sort of magic.[60]

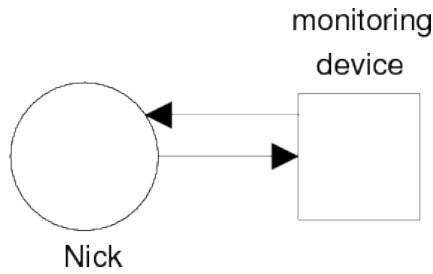


***Figure 3**: Second variation on Gertler's thought experiment*

I, therefore, contend that we should reject the view that introspection is necessary for content inheritance. I contend, moreover, that we should adopt the plausible principle (A). Absent the dubious principle about introspection, there is no obvious reason in favor of (G8). And given (A), we must reject (G8). Without (G8), the argument from content inheritance for a metaphysical disparity between self- and other-knowledge does not go through. And absent such a metaphysical disparity, we have no way of accommodating the sort of epistemic disparity between self- and other-knowledge posited by intermediate Cartesianism.

In fact, I suspect that the original case of Nick and Nora actually turns out to be a counterexample to intermediate Cartesianism. In that case, Nora's beliefs about Nick's "gin" thoughts are caused precisely as reliably as are Nick's own beliefs about those thoughts. Moreover, we may suppose, Nick's evidence that he is thinking about gin is precisely as good as Nora's evidence that he is thinking about gin.[61] These facts together mean that any reasonable epistemology, whether internalist or externalist, will imply that Nora's second-order belief that Nick has an occurrent propositional attitude with *C*1 is precisely as well-justified as is Nick's second-order belief that he has an occurrent propositional attitude with *C*1. Intermediate Cartesianism, of course, is incompatible with this implication—recall (G3).[62] If the case of Nick and Nora is, as it appears to be, a genuine possibility, then, since intermediate Cartesianism is a necessity claim, that privileged access thesis is false. Hence, that reliabilism is unable to accommodate intermediate Cartesian access is no evidence against reliabilism, and is no evidence that reliabilism finally will be incompatible with some tenable privileged access thesis.

Before turning to the problem about generality, which threatens to defeat the argument (R1)–(R5) for the conclusion that reliabilism can accommodate in particular modest Cartesian access, I should deal with an additional objection to (A). It might be objected that, if we adopt (A), the content inheritance strategy collapses into the pure causal strategy: appropriate causation does all the work; content inheritance just "comes along for the ride." I grant that, given (A), the content inheritance strategy collapses into the pure causal response. But this, I suspect, is precisely what is wanted.

To see why, consider a final variant of the Nick and Nora thought experiment. The connections among Nick, Nora, and the monitoring device this time are just as they were in the initial version. But now Nick has been slowly switched from Earth to Twin Earth and back, with neither Nick nor Nora being aware that this has taken place. As it turns out, not only is water on Twin Earth replaced by twater, but also gin is replaced there by tgin. During his sojourn on Twin Earth, therefore, the brain states that formerly realized Nick's "gin" thoughts (his thoughts with content *C*1) instead realized "tgin" thoughts (thoughts with a different content *C*2); he has now been back on Earth for long enough that those brain states again realize "gin" thoughts. Suppose that, as before, Nick enters the relevant brain state (and so has an occurrent propositional attitude with *C*1). Suppose also that Nick introspects and thereby forms a second-order belief that he has an occurrent propositional attitude

with *C*1. The monitoring device, of course, produces in Nora a second-order belief that Nick has an occurrent propositional attitude with *C*1.

Two questions about this modified case are salient here: first, whether Nick's second-order belief inherits the content of his lower-order thought (that is, whether that content is a proper part of the content of his second-order belief); second, whether Nora's second-order belief inherits the content of Nick's lower-order thought (that is, whether that content is a proper part of the content of her second-order belief). If content is ever inherited, clearly Nick's second-order belief inherits the content of his first-order thought; the answer to the first question, then, is affirmative. Now, if, contra A, appropriate causation is insufficient for content inheritance, the answer to the second question might nevertheless be negative. (Gertler, at any rate, will have to answer it that way.) Presumably, if one invokes content inheritance in the case of slowly switched Oscar, one does so because one thinks that it can do some work there. If so, then we should say that Nora's belief in the modified case inherits the content of Nick's thought, for, if we do not, we leave ourselves vulnerable to the following argument.

(N1) Suppose that content externalism is true.

(N2) Suppose that the modified case of Nick and Nora (described above) occurs.

(N3) Suppose that, after having been switched back to Earth, Nick has an occurrent propositional attitude with the content [that gin is good to drink].

(N4) Suppose that, via introspection (subserved by a properly functioning mechanism), Nick then forms the second-order belief that he has an occurrent propositional attitude with the content [that gin is good to drink].

(N5) Suppose that the monitoring device, as expected, produces in Nora the second-order belief that Nick has an occurrent propositional attitude with the content [that gin is good to drink].

(N6) The possibility that Nick has an occurrent propositional attitude with the content [that tgin is good to drink] is a relevant alternative to the fact that he has an occurrent propositional attitude with the content [that gin is good to drink]. (from N1–N3, together with an account of the relevance of alternatives)

(N7) Nick's second-order belief inherits the content [that gin is good to drink] of his lower-order attitude. (from N1–N4 and C)

(N8) Nick can rule out the relevant alternative that he has an occurrent propositional attitude with the content [that tgin is good to drink]. (from N6, N7, and the success of the content inheritance strategy in dealing with the Oscar case)

(N9) Suppose (as might be the case if (A) is false) that Nora's second-order belief does not inherit the content [that gin is good to drink] of Nick's lower-order attitude.

(N10) Nora cannot rule out the relevant alternative that Nick has an occurrent propositional attitude with the content [that tgin is good to drink]. (from N6, N9, and the assumption that content inheritance is required to deal with the Oscar case)

(N11) Nick's second-order belief that he has an occurrent propositional attitude with the content [that gin is good to drink] can qualify as knowledge. (from N4, N8 and an assumption about relevant alternatives)

(N12) Nora's second-order belief that Nick has an occurrent propositional attitude with the content [that gin is good to drink] cannot qualify as knowledge. (from N5, N10 and an assumption about relevant alternatives)

(N13) Therefore, if content externalism is true, then possibly Nick, but not Nora, can know that Nick has an occurrent propositional attitude with the content [that gin is good to drink]. (from N1, N11, and N12)

The problem with (N13) is that if Nick can know that he has an occurrent propositional attitude with the content [that gin is good to drink], then so can Nora (or so say the intuitions of those I consulted). Even Gertler, after all, says only that Nick's higher-order belief in the original case is better justified than is Nora's, not that Nora does not know that Nick is thinking about gin. But the foregoing argument demonstrates that, if content externalism is true, then, if we take content inheritance to be doing some work in the case of slowly switched Oscar and yet reject (A), we might have to say (contrary to our intuitions) that Nora does not even know that Nick is thinking about gin. If the content inheritance strategy is to succeed, then, we should in fact let it collapse into the pure causal response to the slow switching argument.[63]

## IV. SCEPTICISM

I take (R), according to which introspection is highly reliable with respect to the contents of occurrent propositional attitudes, to be uncontroversially true.[64] The argument (R1)–(R5) given in §II.C above suggests that the truth of (R) is sufficient to ensure that reliabilism can accommodate modest Cartesian access. Things are not, however, as simple as they seem at first: the difficulty of individuating belief-producing processes means that we cannot straightforwardly infer from the fact that introspection is reliable with respect to occurrent propositional attitude contents that beliefs about occurrent propositional attitude contents produced by introspection are produced by a reliable process, and hence, according to process reliabilism, justified.

## A. GENERALITY

The generality problem for process reliabilism is the problem of individuating belief-producing processes (Feldman 1985; Conee and Feldman 1998). A given process-token, of course, occurs only once, so that reliability must be construed as a property of process-types; otherwise, every process would be either perfectly reliable or perfectly unreliable. But how are we to determine which of the many process-types under which a given process-token falls is the type that is relevantly (un)reliable? Consider, e.g., my belief *B*1 that the string of characters 'generality problem' appears on the screen in front of me now. And consider the process *p*1 that produced *B*1. Is *p*1 the process of taking my visual inputs at face value, or the process of taking my visual inputs at face value after gazing at a computer screen all day, or the process of taking my visual inputs at face value after gazing at a computer screen all day on a Saturday, or . . . ? Any given belief-production, of course, tokens indefinitely many process-types, some of which are reliable, others of which are unreliable.

The implications of the generality problem for the compatibility of process reliabilism and modest Cartesianism are straightforward. Suppose that, via introspection, a subject *S* forms a belief *B* that she has an occurrent propositional attitude with content *C*. If (R) is true, then we know that *B* is probably true. But, if we are process reliabilists, then, if the generality problem is bona fide, we will have to say that this is not yet sufficient to ground a claim that *B* is justified, for the introspection process-token which produced *B* tokens indefinitely many process-types, not all of which will be reliable. Even given (R), then, if we are to be in a position to assert that *B* is justified, we need some principled way of picking out the (reliable) process-type that might be described as "introspecting (with a properly functioning mechanism) in an attempt to detect the contents of one's own occurrent propositional attitudes" as the type relevant to the reliability of the process-token instantiated by *S*.

The significance of the differences between process and counterfactual reliabilism (described in §I.B) for the question of the ability of reliabilism to accommodate modest Cartesian access thus emerges: the counterfactual reliabilist, given (R), can say that *B* (probably) counts as knowledge, since, given (R), *S* (most likely) tracks the truth with respect to the proposition that she has an occurrent propositional attitude with *C*; but no such move is available to the process reliabilist.

I assume here that counterfactual reliabilism is immune to the generality problem. (CR), the version of the theory stated in §I.B above, is indeed not afflicted by the problem, but if a version of the theory sufficiently sophisticated to avoid the known counterexamples to (CR) will have to invoke belief-forming methods (as does Nozick's own favored version of the theory), then (as Bonjour 2002, 253 points out) counterfactual reliabilism, too, will face the generality problem. The assumption, however, is harmless in the present context, for even if the sophisticated counterfactual reliabilist owes us a story about the individuation of belief-forming processes, her theory, together with (R), still implies modest Cartesian access, since

(R) tells us that the method of introspection is reliable with respect to occurrent propositional attitude contents.

Alston (1995) has proposed an attractive solution to the generality problem: he suggests that we are, in order to pick out the belief-producing process-type relevant to the reliability of a given process-token, to appeal to the "psychologically realized" function underlying the token, where psychologically realized functions are psychological mechanisms.[65] If Alston's proposal is correct, then what process reliabilists require to legitimate a move from (R) to (R3) (and hence a move from (R) to modest Cartesianism) is evidence about the reliability of the mechanism subserving the process of introspection.

Though I have nothing novel to say in support of Alston's proposal, I proceed in §§IV.B–IV.C as if it is true. For suppose that the proposed solution is incorrect. Then, I say with Adler and Levin, "the generality problem, to the extent that there is one, cuts across the board" (2002, 97). Any given process-token (regardless of whether it is a belief-producing process-token) falls under indefinitely many types, types that will have various mutually incompatible properties. Yet theorizing and explanation are not and should not as a consequence of the recognition of this fact be paralyzed. For clearly, in a wide range of cases, we somehow—never mind that it is immensely difficult to say just how we do this—manage to pick out the relevant types. And, the intrinsic interest of the generality problem aside, this ability is, in general, enough. I therefore suppose that epistemologists may pick out the relevant belief-producing process-types in the normal way, whether or not we have solved the generality problem. And the normal way of individuating belief-producing process-types coincides, I suppose, with Alston's proposal: we are, in the first place, to look to our psychological architecture.[66] The question whether the truth of (R) entitles process reliabilists to modest Cartesianism, then, reduces to that of the reliability of the psychological mechanism(s) subserving subjects' introspection of their own occurrent propositional attitude contents.

## B. ERROR

Kornblith (1998) gives what I take to be a sort of relevant alternatives argument for a "mitigated scepticism" about self-knowledge; though his sceptical argument is not what I wish to discuss here, it serves to highlight certain relevant features of our cognitive lives. Drawing our attention to the prevalence of psychological conditions that cause subjects with them to make false judgments about their own mental lives, but that nevertheless do not reveal themselves to introspection, Kornblith concludes that we generally lack the evidence required to justify our claims to self-knowledge, and that there is therefore considerably less self-knowledge around than is normally supposed. The argument, I suspect, does not rule out the sort of modest Cartesian privilege I am interested in here.[67] But it does draw our attention to the relevance, given reliabilism, of certain empirical psychological results to the truth of modest Cartesianism.

In addition to evidence on the prevalence and effects of psychological disorders, Kornblith appeals to work in psychology typified by that of Nisbett and his colleagues (Nisbett and Wilson 1977; Nisbett and Ross 1980; Wilson 2002). This work pretty clearly demonstrates that retrospective causal reports about mental processes are highly unreliable, that "there may be little or no direct introspective access to higher-order cognitive processes" (Nisbett and Wilson 1977, 231).[68]

Consider, e.g., subjects' introspective reports about the occurrence or non-occurrence of a position effect (Nisbett and Ross 1980, 207; Kornblith 2002, 111–112). (When making a forced choice from amongst an array of identical objects, subjects tend to prefer the rightmost object.) Introspection does not reveal the occurrence of the position effect; in fact, it seems to reveal the non-occurrence of the effect. Or consider subjects' introspective reports about the occurrence or non-occurrence of an anchoring effect (Tversky and Kahneman 1974, 14; Kornblith 2002, 113–114). (When asked to estimate a certain number, subjects' responses are influenced by an obviously irrelevant number that has previously been made salient.) Again, introspective reports here are unreliable: introspection does not reveal the occurrence of the anchoring effect, and instead seems to reveal its non-occurrence. Or consider introspective reports about the existence or non-existence of confirmation biases (Lord, Ross, and Lepper 1979; Nisbett and Ross 1980, 170–172; Kornblith 2002, 117–118). (Subjects evaluating hypotheses they favor tend to look for confirming but not disconfirming evidence and tend to weight confirming evidence more heavily than disconfirming evidence.) Here, too, introspective reports are unreliable: introspection does not reveal the existence of a confirmation bias; instead, it seems to reveal its non-existence. Examples of similar limitations of introspective access can be multiplied, it seems, almost indefinitely.

Now, evidence of this sort suggests that subjects frequently err when reasoning about their own (past) mental states, but it does not suggest that subjects regularly err when attempting to detect their own occurrent mental states. Even if so, however, the evidence might still suggest that the mechanism subserving introspection is unreliable. Nisbett and Wilson (1977) propose, as an explanation of the evidence, that higher-order beliefs about mental processes are theory-mediated: the tendency of subjects systematically to err in certain areas can then be explained by deficiencies in the theories they employ. If introspective reports of current mental states (and, in particular, of occurrent propositional attitude contents) are produced in the same way, then the unreliability of the psychological mechanism subserving introspection, when used in other areas, "infects" its reliability when used to produce beliefs about occurrent propositional attitude contents: given the wide range of cases in which subjects' higher-order beliefs consistently are erroneous, the mechanism would on the whole seem to be highly unreliable. The point would be not that introspection is strictly fallible—we have granted this from the outset—but rather that introspection is in general an unreliable process.

Hence, though there is no evidence suggesting that (R) is false, we may nevertheless appeal to the relevant psychological findings in the manner of the following

argument for the conclusion that reliabilism cannot accommodate even modest Cartesian access.

(P1) If (PR) is true, then (even given the truth of (R)) modest Cartesianism is true only if introspection is a reliable belief-producing process. (from MC, PR and R)

(P2) Introspection is a reliable belief-producing process only if the psychological mechanism subserving it is reliable. (from Alston's solution to the generality problem)

(P3) There is reason to think that the psychological mechanism subserving introspection is in fact unreliable. (from the psychological evidence alluded to above)

(P4) There is reason to think that introspection is an unreliable belief-producing process. (from P2 and P3)

(P5) There is reason to think that, if (PR) is true, then modest Cartesianism is false. (from P1 and P4)

This argument in effect counters the argument (R1)–(R5) given in §II.C above by showing that the truth of (R) is insufficient, given process reliabilism, to ground modest Cartesian access. The compatibility of reliabilism and privileged access is thus again threatened.

## C. THEORIZING, SIMULATING, MONITORING

The argument (P1)–(P5) is only really vulnerable at the point of the move from the relevant psychological results to (P3). The question for the process reliabilist, then, is whether the mechanism at work in these cases is the same as that at work when subjects introspect their own occurrent propositional attitude contents: if there is a unique mechanism at work in cases of both sorts, then it is unreliable, and (R1)–(R5) does not go through; if, on the other hand, there is a separate mechanism responsible for the production of introspective beliefs about one's own occurrent propositional attitudes (or, perhaps, about these and certain other occurrent mental states), then (given the truth of (R)) it is reliable, and (P1)–(P5) can be blocked at the point of the move to (P3).

In this section, I appeal to current theories of mindreading to argue that there is reason to suppose that introspection of occurrent propositional attitudes is subserved by a distinct (and reliable) mechanism (put differently: that there are really multiple processes of introspection). I consider three going theories of mindreading (or "internal" folk psychology [Stich and Ravenscroft 1993]): the simulation theory (Goldman 1993a; Goldman 1993b), the theory theory (Gopnik 1993; Carruthers 1996), and a recent mixed or intermediate theory (Nichols and Stich 2003). My argument does not require me to take a stand on which of these theories is correct; instead, I argue that, if any of these theories is correct about the manner in which subjects detect their own occurrent propositional attitudes, then there is a distinct

(reliable) mechanism subserving introspection. And if this is right, then we should reject (P3).

There are, of course, multiple going versions of each of these theories of mindreading. The theory theory, e.g., comes in both "scientific" and nativist versions (Carruthers 1996, 22–23), though the distinction between scientific theory theory and nativist theory theory safely can be ignored in the present context. Theory theorists (of whatever stripe) tend to begin with a subject's ability to read other minds, suggest an account of this ability, and then explain the subject's ability to read her own mind in the same terms: detection of the mental states of others is supposed to be a process mediated by a theory of mind, a folk psychological theory stored (subpersonally) by the mindreading subject;[69] detection of her own mental states likewise is (somehow) mediated by the subject's theory of mind. Though the distinction between scientific theory theory and nativist theory theory may be ignored here, that between two theory theoretic stories about the role of a subject's theory of mind in detecting her own mental states cannot.

According to the first story, one detects one's own mental states in precisely the (theory-mediated) manner in which one detects the mental states of others: "[t]he *only* information used as evidence for the inference involved in detecting one's own mental state is the information provided by perception (in this case, perception of oneself) and by one's background beliefs (in this case, background beliefs about one's own environment and previously acquired beliefs about one's own mental states)" (Nichols and Stich 2003, 156). On this version of the theory theory, any epistemic advantage one has with respect to detecting one's own mental states (including one's own occurrent propositional attitudes) is to be explained entirely in terms of the disparity between the amount of observational knowledge one typically has of oneself and the amount one typically has of others.

If this version of the theory theory is correct, then process reliabilism likely will be unable to accommodate even the weak variety of privileged access described by modest Cartesianism: subjects will use the same mechanism for detecting their own occurrent propositional attitudes (and other occurrent mental states) as they use for predicting others' mental states and their own past and future mental states. Precisely because it does not permit sufficient additional reliability in the case of subjects' introspection of their own occurrent mental states, however, there is little reason to suppose that this version of the theory theory stands much chance of being correct: it will permit at most the sort of "privileged" access a logical behaviorist (Ryle 1949) might endorse.[70]

A more plausible theory theoretic story about the role of the subject's theory of mind in the process of introspection is one on which subjects have a source of information about their own occurrent mental states unlike their sources of information about the mental states of others and their own past selves. On this version of the theory,[71] while the subject's theory of mind is somehow involved in her detection of her own occurrent mental states, introspection is subserved by a special mechanism which has access to the subject's occurrent thoughts.[72]

This more plausible version of the theory theory (Fig. 4) explains psychological results of the sort described in §IV.B precisely by pointing out that this special mechanism is not involved in generating the relevant higher-order beliefs.[73] The most plausible going version of the theory theory of self-knowledge, then, implies that there is a special mechanism subserving detection of occurrent mental states (including occurrent propositional attitudes). Should this version of the theory theory be correct, (P3) will turn out to be false.

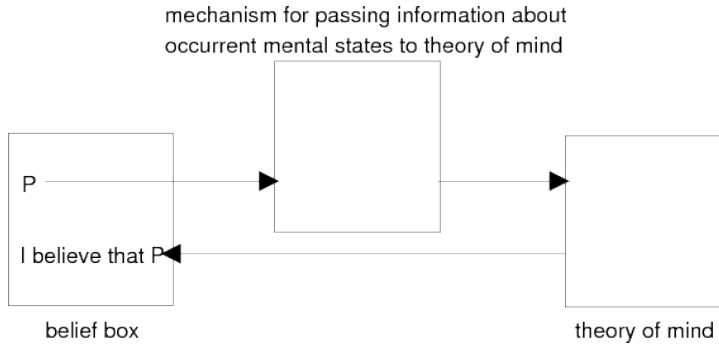


***Figure 4**: The theory theory on reading one's own mind*

In contrast to theory theorists, simulation theorists[74] tend to move from self-knowledge to other-knowledge, and so have in general had less to say about how one reads one's own mind (Carruthers 1996).[75] The crucial thought behind simulationism is that the theory theoretic account of mindreading, on which mindreading is mediated by a rich body of information, is in a sense uneconomical: given the similarity between mindreaders and their targets, perhaps it is possible for the mindreader to use elements of her own cognitive apparatus to simulate the mind of her target—the idea is that certain of the mindreader's cognitive mechanisms are taken "off-line" and receive hypothetical information about the target, with the resulting states used to predict the mental states of the target (Carruthers 1996, 28–29).[76]

Unless they are to say (implausibly) that introspection of occurrent mental states is conducted by means of self-simulation, simulation theorists need to posit some sort of special capacity deployed by subjects in order to detect their own occurent mental states. Goldman (1993b, 1993a), e.g., suggests that subjects might have a capacity for detecting the qualitative properties of those states (including, significantly, the qualitative properties of propositional attitude contents). As Nichols and Stich point out (2003, 196), in order for this proposal to be workable, it will have to posit a special mechanism which detects these properties and outputs appropriate higher-order beliefs (Fig. 5). The simulation theorist can explain the psychological results discussed in §IV.B by appealing to limitations in the self-simulations conducted by subjects to produce judgements about their own past mental states.

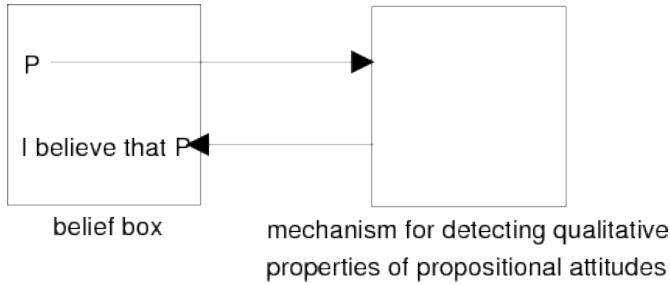


***Figure 5****: The simulation theory on reading one's own mind*

Nichols and Stich (2003, 197–198) also point out that this is a rather implausible story about the production of higher-order beliefs—if indeed propositional attitude contents have qualitative properties, what sort of mechanism could detect them? It is not the ultimate tenability of the proposal that interests me here, but rather its implications vis-à-vis reliabilism and privileged access. Like the more plausible version of the theory theory, this version of the simulation theory posits a special psychological mechanism dedicated to the production of higher-order beliefs about one's own occurrent mental states; this mechanism is supposed to be distinct from the mechanism(s) responsible for production of beliefs about others' mental states and one's own non-occurrent mental states. Given the simulation theory, then, there is reason to posit a special (and reliable) mechanism devoted to the detection of occurrent propositional attitudes. Should simulationism turn out to be correct, then, (P3) will turn out to be false.

Hybrid theories of mindreading intermediate between the theory theory and the simulation theory lately have begun to be developed; here, I consider only that hybrid theory presented in Nichols and Stich (2003).[77] Nichols and Stich posit a special "monitoring mechanism" that "takes the representation *p* in the Belief Box as input and produces the representation *I believe that p* as output. The proposed mechanism (or perhaps a distinct but entirely parallel mechanism) would work in much the same way to produce representations of one's own desires, intentions, and imaginings" (2003, 160–161). This hybrid theory (Fig. 6) incorporates the theory theoretic account of predicting one's own past and future mental states and the mental states of others, and so can handle the evidence described in §IV.B in the same manner as the theory theory.

Clearly, if this monitoring mechanism account of introspection is correct, then introspection of one's own occurrent propositional attitude contents is subserved by a unique (and reliable) mechanism (or set of mechanisms).[78] It follows that if one of the theory theory, the simulation theory, or Nichols and Stich's hybrid theory is correct, then introspection of one's occurrent propositional attitudes is subserved by a unique mechanism. This mechanism (whichever one it is), moreover, is clearly reliable: the challenge posed by the psychological results discussed in the previous section was not to the truth of (R); the challenge, rather, was to show that the errors

subjects make when introspecting, e.g., the causal antecedents of their current mental states are attributable to the workings of some mechanism not centrally involved in introspection of occurrent propositional attitudes. With this challenge met, we should reject (P3), and with it the argument for the view that, the truth of (R) notwithstanding, reliabilism cannot accommodate modest Cartesian access. We should, that is, accept the conclusion (R5) that reliabilism can in fact accommodate such access.

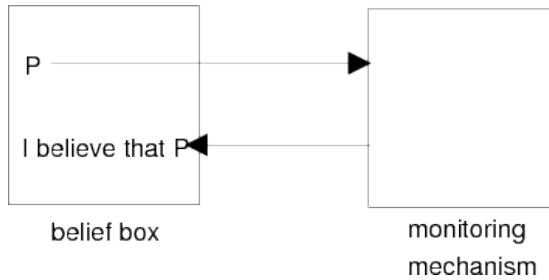


***Figure 6****: The hybrid theory on reading one's own mind*

## V. CONCLUSION

On the basis of the argument of §§II and IV, I conclude that there is reason, at present, to suppose that process reliabilism not only is compatible with but even implies modest Cartesianism. On the basis of §III.A, I conclude that process reliabilism cannot accommodate intermediate Cartesian access. I take myself to have shown, in §§III.B–III.C, however, that intermediate Cartesianism is an untenable position. I therefore conclude that, of those considered herein at least, reliabilism can accommodate the strongest tenable form of access.[79]

## ENDNOTES

1. See Ludlow and Martin 1998 for a brief overview.

2. Of the "Twin Earth" variety motivated by Burge 1979 with the aid of thought experiments related to those used by Putnam 1996 to motivate semantic externalism.

3. §I.A below distinguishes among those of the multiple available theses of privileged access salient in the context of this paper (and Alston 1971 surveys many more); throughout, 'privileged access' (without qualifiers) serves as a generic term for the varieties of access posited by these theses. Note that I do not discuss the "Wittgensteinian" variety of access (Wittgenstein 1958) lately endorsed by Moran (1997, 2001).

The related debate over the joint implications for external world scepticism of content externalism and privileged access is irrelevant to my concerns here. See McKinsey 1991 for the key anti-sceptical argument from the conjunction of content externalism and privileged access. Compare the anti-sceptical argument of Putnam 1981; see Falvey and Owens 1994

on the relationship between the two arguments. Tye and McLaughlin 1998 responds to McKinsey 1991, and Boghossian 1997 contains interesting further discussion.

4. That the question seldom explicitly has been discussed is somewhat puzzling, for reliabilism is among the major going epistemologies (and is the major going externalist epistemology), there are certain affinities between content externalism and epistemological externalism, and self-knowledge is a theme to which analytic epistemology traditionally has accorded considerable significance. Whatever the explanation for the existence of this gap in the literature, the answer to the question should, for reasons given below, be of broad interest. On relations between content externalism and epistemological externalism, see Majors and Sawyer 2005.

5. Though her object of interest ultimately is content externalism, Brown (2000, 2004) does explicitly raise the question of epistemological externalism and privileged access.

6. Exceptions here are Kornblith 1998 and 2002; see §IV.B below.

7. See Wilson 2002 for a recent overview.

8. See Carruthers and Smith 1996 for a brief overview.

9. Given reliabilism, the existence and scope of privileged access is an empirical question—see §IV below—so that a superficial survey of some of the highlights of the empirical literature of the sort undertaken here cannot, even in principle, definitively answer it. (At present, perhaps, not even a thorough review of the literature could accomplish this.)

10. See the discussion of Gertler 2002 in §III below.

11. I do not claim that Descartes endorsed this or any other of the theses discussed herein.

12. Though my argument (like that of Brown 2000) trades on the generality of belief-producing processes and draws attention to the salience of certain differences between counterfactual and process reliabilism to the question of the compatibility of reliabilism and privileged access, her argument, which has to do with the joint implications of reliabilism and the doctrine of object-dependent thoughts (McDowell 1977), covers different ground than does mine.

13. In formulating these, I draw on Gertler 2003a and 2003b; and (especially) on Alston 1971.

14. Though, as noted below, it is something of a stretch to refer to modest Cartesianism as an infallibility thesis.

15. These characterizations of indubitability, incorrigibility, and transparency are more or less those given at Fernández 2003, 354.

16. Alston 1971 describes thirty-four varieties of privileged access (including forms of infallibility).

17. E.g., by Fernández 2003.

18. See Alston 1971.

19. If INF-infallibility is a matter of justification (rather than knowledge), it will only entail a weaker thesis on which if *S* believes that *P*, then probably *P*.

20. As §§IV.A–IV.B below demonstrate, evidence about the general reliability of the process responsible for the production of an INF*-infallible belief is still required before the process reliabilist may conclude that it is INF-infallible.

21. I thus do not consider the extremely weak forms of privileged access that can be granted even by logical behaviorists. See, e.g., Ryle 1949, 169.

22. Results of the sort discussed in Nisbett and Wilson 1977 are relevant here; I return to these in §IV.B.

23. That *S* has introspective access to a state here simply means that she can form a belief about it in the manner described in clause (i).

24. As noted above, given reliabilism, the existence of privileged access becomes an empirical question. Unless there is an appropriate a posteriori necessity lurking in the vicinity, then, reliabilism will also make privileged access a contingent matter.

25. Gertler formulates the view as follows:

> (i) Knowledge of one's own occurrent thoughts can at times achieve a level of certainty higher than that which [other's knowledge of one's thoughts] could achieve; (ii) one can at times know one's own occurrent thoughts without investigating the external world; and (iii) (i) and (ii) are true of rational, reflective creatures in all possible worlds. (2002, 126)

(A similar thesis is mentioned in Gertler 2003b.) It is hard to see just what Gertler has in mind by clause (i) of this formulation, but I take the thought to be more or less that expressed by clause (i) of (IC). (Note that she makes clear that the beliefs mentioned in clause (i) of the formulation result from introspection.)

26. On some views, the belief of *S*'s that she would express by 'I am thinking that *P*' is not type-identical to the belief of *S*"'s that she would express by '*S* is thinking that *P*'; (IC) can be reformulated so as to accommodate such views.

27. As Brown 2000 points out, and as is made clear in §IV.A below.

28. It would be natural to discuss the implications of proper functionalism (Plantinga 1993) for privileged access alongside those of reliabilism, but space does not permit me to do so here.

29. For a brief overview, see Feldman 2003.

30. Though, admittedly, if introspection is reliable with respect to certain mental states but not others, there is some pressure to adopt the view that there are really multiple processes of introspection.

31. With an eye to the second of these consequences, we might introduce a slight variant of modest Cartesianism:

> (MC') Every *S* is such that (i) usually, if *S* believes that she has some occurrent propositional attitude with *C*, her belief was produced by a process of introspection, and the mechanism subserving that process was functioning properly, then she knows that *P*, and (ii) *S* sometimes has introspective access to the contents of her occurrent propositional attitudes.

Process reliabilism is more likely to be able to accommodate (MC) than (MC'), while counterfactual reliabilism is more likely to be able to accommodate (MC') than (MC). In what follows, however, I will in general ignore the distinctions between (MC) and (MC') and between (IC) and its obvious variant (IC').

32. For background, see Putnam 1996 and Burge 1979; for brief statements, see Burge 1988, 650; Boghossian 1989, 5; Falvey and Owens 1994, 107; Gibbons 1996, 287; and McLaughlin and Tye 1998, 349.

33. See Fodor 1987 for an argument in favor of retaining narrow content.

34. I suppose that it is possible to raise the question also of privileged access to narrow content; but the question is of much less interest, for the very existence of narrow content remains controversial.

35. See also Loar 1988.

36. Relevant alternatives and related notions crop up a number of times throughout the paper; I do not define these notions, but rely on our intuitive understandings of them. On proposals as to what is required to rule an alternative out, see Pryor 2001.

37. Compare Warfield 1992.

38. Or not quite—see below.

39. Something like this strategy is discussed by, inter alia, Burge 1988; Heil 1988; Falvey and Owens 1994; Gibbons 1996; McLaughlin and Tye 1998; and Brown 2000 and 2004. It is not always clear whether these authors reasonably can be read as discussing the pure causal strategy, or whether they should, instead, be read as discussing the content inheritance strategy taken up in §III.B below; I therefore make no claim that the pure causal strategy described here is a faithful representation of the view discussed by any of these philosophers.

40. Strictly speaking, the counterfactual reliabilist will have to concede that she can only show that slow switching cases do not prevent her from endorsing (MC') or (IC').

41. It might be objected that, contrary to my claim that the pure causal strategist attempts to show that, given reliabilism, (S6) does not follow from (S1)–(S5), in fact she simply rejects a relevant alternatives-based epistemology in favor of reliabilism. Should this be so, the overall argument of the paper would be unaffected, for my aim in setting out the slow switching argument and the pure causal response to it is simply to get clear on the implications of the assumption about the reliability of introspection made by the pure causal strategist. Nor is it clear that the objection is correct: e.g., if what is required to be in a position to rule an alternative *P* out is simply to be in a position to know that *P* is false, then perhaps the causal strategist explicitly can claim that Oscar is in a position to rule out the possibility that he has an occurrent propositional attitude with the content [that twater is wet].

42. Again, strictly speaking, the argument can only show that counterfactual reliabilism implies (MC').

43. One wonders why Nora does not just program the device to prevent Nick's thinking about gin!

44. Note that (R) does not imply the consequent of (G5).

45. Gertler 2002 deals convincingly with several objections to the argument.

46. Which she finds in Burge 1988; and Heil 1988.

47. Perhaps some other strategy can accommodate intermediate Cartesian access; if so, then those attracted to intermediate Cartesianism will count its inability to accommodate that sort of access against reliabilism. I am aware of no such alternative strategy.

48. It is natural to wonder whether paradox threatens such self-reflexive thoughts; for a response to this sort of worry, see Harman 2006.

49. Note that it would not do to say that higher-order beliefs about one's own occurrent propositional attitude contents inherit (the relevant part of) their content from those lower-order attitudes irrespective of whether the higher-order beliefs result from introspection, for then agents could never err about their own occurrent thought contents, a possibility which we surely want to permit.

50. There is a sense in which reliabilists can, if they like, help themselves to content inheritance, since, if introspection (subserved by a properly functioning mechanism) normally ensures content inheritance, then introspection is reliable (and introspected beliefs normally track the truth). Reliabilists could, that is, given an argument like the pure causal response, but with (C) replacing (R). But note that, given reliable production, content inheritance is in a sense redundant.

51. Note that, even if content inheritance does secure a metaphysical disparity between self- and other-knowledge, the option remains open to the pure causal strategist (that is, to the reliabilist) of arguing that that disparity is epistemically insignificant.

52. The claim is plausible only given that it concerns what we might call "token content inheritance," in which the higher-order thought contains the lower-order thought itself as a part. Sawyer 2002 argues that it is a mistake to read Burge as suggesting that this sort of content inheritance occurs outside of an extremely restricted range of "cogito-like" judgements. But Gertler clearly takes it to occur in a much broader range of cases.

53. See Adams and Aizawa 2001 for a response to Clark and Chalmers 1998; note that they admit that this sort of "transcranial cognition" is a logical and even a nomological possibility.

54. But note that they admit that, in principle, "my mental states [could] be partly constituted by the states of other thinkers" (Clark and Chalmers 1998, 17).

55. They remark also that "the information in the notebook has been consciously endorsed at some point in the past, and indeed is there as a consequence of this endorsement," but hesitate about the relevance of this feature of the case (Clark and Chalmers 1998, 17).

56. Note that the proposal that they constitute a coupled system is consistent with the intuition that the thoughts in question belong, in an important sense, to Nick and not to Nora: they are caused by his environment, his desires, etc., they are bound up with his behavior in certain characteristic ways, and so on.

57. On the distinction between dispositional and occurrent beliefs, see Audi 1994.

58. In fact, even the mechanism is arguably the same in both cases, though instantiated by a different piece of hardware in each; if so, then the only salient difference between the two cases is at the level of hardware.

59. It might be objected that there is after all an asymmetry between Nick and Nora: whereas Nick's higher-order belief that he is having a thought with *C*1 (a belief he would express by 'I am now thinking of gin') contains an essential indexical (is "self-locating" [Perry 1979] or "irreducibly de se" [Lewis 1979]), Nora's higher-order belief that Nick is having a thought with *C*1 (a belief she would express by 'Nick is now thinking of gin') does not. The objection raises two questions: first, whether, given that the asymmetry holds, it provides a principled reason for saying that Nick's higher-order thought, but not Nora's, inherits the content of

Nick's lower-order thought; second, whether the asymmetry really holds (that is, whether Nora's thought really does not contain an essential indexical). As to the first question, there is, on the face of it, no reason to suppose that the occurrence in a higher-order belief of an essential indexical determines whether it inherits the content of a lower-order belief. Gertler (2002, 138) argues that, if we admit a metaphysical disparity between Nick's second-order thought and Nora's second-order thought, we then have a way of grounding a referential difference (viz., the (non-)occurrence of a logically direct reflexive) between them. Perhaps this is right; but what we need here is a reason to think that the referential difference grounds a metaphysical difference. As to the second question, it is no longer even clear that Nora cannot refer to Nick's first-order thought with a logically direct reflexive: if Nick and Nora constitute a coupled system, perhaps she could in fact express her second-order belief that Nick is having a thought with *C*1 by 'I am now thinking of gin.'

60. Or perhaps even to fetishize our (actual) mental hardware.

61. We could easily enough alter the case so that Nora's belief is caused more reliably and supported by better evidence than is Nick's.

62. Why, then, might someone have an intuition consistent with the intermediate Cartesian judgement about the case? I suspect that the source of such an intuition would have to be located in the fact that, around here, introspective beliefs about one's own occurrent propositional attitude contents are normally better justified than are others' beliefs about those contents.

63. When he first arrives back on Earth, Nick continues for some time to think tgin thoughts. Since his relevant higher-order beliefs will at this point also contain the tgin content, the reliability of his introspection process is unaffected. What of the monitoring process used by Nora? At first glance, it would seem that her relevant higher-order beliefs are always gin beliefs, so that Nick's thinking tgin thoughts during this period affects the reliability of the process. But this initial reaction is mistaken: if (A) is right, then the relevant higher-order beliefs of Nora's inherit the tgin content of Nick's lower-order beliefs during the period in question.

64. Nichols and Stich remark—this is already stronger than (R)—that what the theory of self-awareness needs to explain is the fact that "when normal adults believe that *p*, they can quickly and accurately form the belief *I believe that p*; when normal adults desire that *p*, they can quickly and accurately form the belief *I desire that p*; and so on for other basic propositional attitudes like *intend* and *imagine*" (2003, 160).

65. Alston's proposal is challenged in Conee and Feldman 1998 and defended in Adler and Levin 2002, to which Feldman and Conee 2002 is a response.

66. One might worry that this approach requires the "folk" to have too much knowledge of scientific psychology. But the suggestion is not that belief-producing processes are individuated by reference to the architecture of the brain in everyday contexts, but rather that they are individuated that way in theoretical contexts; ordinary individuation of belief-producing processes seems to be a rather rough and ready, approximate affair.

67. Kornblith would likely grant this: he admits that "[m]any cases of self-knowledge will elude [his] argument" (1998, 58).

68. For a dissenting view, see White 1988, where it is argued that there is little confirming or disconfirming evidence for this hypothesis. See also Ericsson and Simon 1993, which

argues that retrospective reports are sometimes more accurate than Nisbett and Wilson 1977 suggests.

69. On the nature of this theory, see Maibom 2003.

70. Note that Frith and Frith 1999, 1694 argues that the brain imaging evidence suggests that a distinct region of the brain is involved in representation of one's own mental states; compare Gallagher and Frith 2003, 80. This view seems to be supported by the fMRI results reported by Gusnard et al. 2001, 4263.

71. Something like which is perhaps endorsed by Carruthers 1996.

72. Nichols and Stich (2003, 159) point out that it is not entirely clear what role is left, on this story, for the subject's theory of mind to play in detecting her own occurrent mental states. I ignore this and other objections to this version of the theory theory here, since my present interest in its account of reading one's own mind is confined to the implications of that account for the question of reliabilism and privileged access.

73. Carruthers writes: "So what one does in such cases, knowing that there must be some explanation for one's action, is to construct an explanation post hoc, either by deploying relevant theoretical knowledge, or by using a simulation strategy. This enables the explanation to be wildly at variance with the facts, just as it can be in our attempted explanation of the behaviour of others" (1996, 27–28).

74. Space does not permit me to take the variety of simulationism developed in Gordon 1986 and 1995 up here.

75. This tendency seems to be a defect of simulationism, but, again, it is one I safely can ignore here.

76. Gallese and Goldman 1998 argues that the activity of mirror neurons in the brains of macaque monkeys—the suggestion is that a similar system exists in humans—supports the simulation theory. (See Gallese et al. 1996; and Umiltà et al. 2001 for background.) Ramnani and Miall 2004, 88 argues against Gallese and Goldman on the basis of an fMRI study the results of which suggest that simulation alone does not explain our ability to predict the actions of others, that is, that a "theory of mind" area of the brain might also be involved.

77. On the basis of an fMRI study, Vogeley et al. 2001 argues, against Gallese and Goldman 1998, that some sort of mixed theory of mindreading is probably correct.

78. It also promises a simple explanation of content inheritance.

79. Thanks to Michael DePaul, Hilary Kornblith, Jonathan Schaffer, and an anonymous reviewer for comments on earlier drafts of this paper.

## BIBLIOGRAPHY

Adams, Fred, and Ken Aizawa. 2001. "The Bounds of Cognition." *Philosophical Psychology* 14: 43–64.

Adler, Jonathan, and Michael Levin. 2002. "Is the Generality Problem too General?" *Philosophy and Phenomenological Research* 65: 87–97.

Alston, William. 1971. "Varieties of Privileged Access." *American Philosophical Quarterly* 8: 223–241.

———. 1995. “How to Think about Reliability.” *Philosophical Topics* 23: 1–29.

Audi, Robert. 1994. “Dispositional Beliefs and Dispositions to Believe.” *Noûs* 28: 419–434.

Boghossian, Paul A. 1989. “Content and Self-Knowledge.” *Philosophical Topics* 17: 5–26.

———. 1997. “What the Externalist Can Know A Priori.” *Proceedings of the Aristotelian Society* 97: 161–175.

Bonjour, Laurence. 2002. “Internalism and Externalism.” In *Oxford Handbook of Epistemology*. Ed. Paul K. Moser. Oxford: Oxford University Press, 234–263.

Brown, Jessica. 2000. “Reliabilism, Knowledge, and Mental Content.” *Proceedings of the Aristotelian Society* 100: 115–135.

———. 2004. *Anti-Individualism and Knowledge*. Cambridge, Mass.: MIT Press.

Burge, Tyler. 1979. “Individualism and the Mental.” *Midwest Studies in Philosophy* 4: 73–122.

———. 1988. “Individualism and Self-Knowledge.” *Journal of Philosophy* 85: 649–663.

Carruthers, Peter. 1996. “Simulation and Self-Knowledge: A Defence of Theory-Theory.” In *Theories of Theories of Mind*. Ed. Peter Carruthers and Peter K. Smith. Cambridge: Cambridge University Press, 22–38.

Carruthers, Peter, and Peter K. Smith. 1996. “Introduction.” In *Theories of Theories of Mind*. Ed. Peter Carruthers and Peter K. Smith. Cambridge: Cambridge University Press, 1–10.

Clark, Andy, and David J. Chalmers. 1998. “The Extended Mind.” *Analysis* 58: 7–19.

Conee, Earl, and Richard Feldman. 1998. “The Generality Problem for Reliabilism.” *Philosophical Studies* 89: 1–29.

Ericsson, Anders K., and Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data*. Cambridge, Mass.: MIT Press.

Falvey, Kevin, and Joseph Owens. 1994. “Externalism, Self-Knowledge, and Skepticism.” *Philosophical Review* 103: 107–137.

Feldman, Richard. 1985. “Reliability and Justification.” *Monist* 68: 159–174.

———. 2003. *Epistemology*. Upper Saddle River, N.J.: Prentice Hall.

Feldman, Richard, and Earl Conee. 2002. “Typing Problems.” *Philosophy and Phenomenological Research* 65: 98–105.

Fernández, Jordi. 2003. “Privileged Access Naturalized.” *Philosophical Quarterly* 53: 352–372.

Fodor, Jerry. 1987. *Psychosemantics*. Cambridge, Mass.: MIT Press.

Frith, Christopher D., and Uta Frith. 1999. “Interacting Minds—A Biological Basis.” *Science* 286: 1692–1695.

Gallagher, Helen L., and Christopher D. Frith. 2003. “Functional Imaging of ‘Theory of Mind.’” *Trends in Cognitive Sciences* 7: 77–83.

Gallese, V., L. Fadiga, L. Fogassi, and G. Rizzolatti. 1996. “Action Recognition in the Premotor Cortex.” *Brain* 119: 593–609.

Gallese, Vittorio, and Alvin Goldman. 1998. “Mirror Neurons and the Simulation Theory of Mind-Reading.” *Trends in Cognitive Sciences* 2: 493–501.

Gertler, Brie. 2002. "The Mechanics of Self-Knowledge." *Philosophical Topics* 28: 125–146.

———. 2003a. "Introduction: Philosophical Issues about Self-Knowledge." In *Privileged Access: Philosophical Theories of Self-Knowledge*. Ed. Brie Gertler. Burlington, Vt.: Ashgate, xi–xxii.

———. 2003b. "Self-Knowledge." In *Stanford Encyclopedia of Philosophy*. Ed. Edward N. Zalta. http://plato.stanford.edu/archives/spr2003/entries/self-knowledge.

Gibbons, John. 1996. "Externalism and Knowledge of Content." *Philosophical Review* 105: 287–310.

Goldman, Alvin. 1979. "What is Justified Belief?" In *Justification and Knowledge: New Studies in Epistemology*. Ed. George S. Pappas. Dordrecht: Reidel, 1–23.

———. 1993a. *Philosophical Applications of Cognitive Science*. Boulder, Colo.: Westview.

———. 1993b. "The Psychology of Folk Psychology." *Behavioral and Brain Sciences* 16: 15–28.

Gopnik, Alison. 1993. "How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality." *Behavioral and Brain Sciences* 16: 1–14.

Gordon, Robert M. 1986. "Folk Psychology as Simulation." *Mind and Language* 1: 158–171.

———. 1995. "Simulation Without Introspection or Inference from Me to You." In *Mental Simulation: Evaluations and Applications*. Ed. Martin Davies and Tony Stone. Oxford: Blackwell, 53–67.

Gusnard, Debra A., Erbil Akbudak, Gordon L. Shulman, and Marcus E. Raichle. 2001. "Medial Prefrontal Cortex and Self-Referential Mental Activity: Relation to a Default Mode of Brain Function." *Proceedings of the National Academy of Sciences of the United States of America* 98: 4259–4264.

Harman, Gilbert. 2006. "Self-reflexive Thoughts." *Philosophical Issues* 16: 334–345.

Heil, John. 1988. "Privileged Access." *Mind* 97: 238–251.

Kornblith, Hilary. 1998. "What Is It Like to be Me?" *Australasian Journal of Philosophy* 76: 48–60.

———. 2002. *Knowledge and its Place in Nature*. Oxford: Clarendon.

Lewis, David. 1979. "Attitudes *de dicto* and *de se*." *Philosophical Review* 88: 513–543.

Loar, Brian. 1988. "Social Content and Psychological Content." In *Contents of Thought*. Ed. Robert Grimm and Daniel Merrill. Tucson, Ariz.: University of Arizona Press, 99–110.

Lord, C., L. Ross, and M. Lepper. 1979. "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence." *Journal of Personality and Social Psychology* 34: 2089–2109.

Ludlow, Peter. 1995. "Externalism, Self-Knowledge, and the Prevalence of Slow-Switching." *Analysis* 55: 45–49.

Ludlow, Peter, and Norah Martin. 1998. "Introduction." In *Externalism and Self-Knowledge*. Ed. Peter Ludlow and Norah Martin. Stanford: Center for the Study of Language and Information, 1–16.

Maibom, Heidi. 2003. "The Mindreader and the Scientist." *Mind and Language* 18: 296–315.

Majors, Brad, and Sarah Sawyer. 2005. "The Epistemological Argument for Content Externalism." *Philosophical Perspectives* 19: 257–280.

McDowell, John. 1977. "On the Sense and Reference of a Proper Name." *Mind* 86: 159–185.

McKinsey, Michael. 1991. "Anti-Individualism and Privileged Access." *Analysis* 51: 9–16.

McLaughlin, Brian, and Michael Tye. 1998. "Is Content Externalism Compatible with Privileged Access?" *Philosophical Review* 107: 349–380.

Moran, Richard. 1997. "Self-Knowledge: Discovery, Resolution, and Undoing." *European Journal of Philosophy* 5: 141–161.

———. 2001. *Authority and Estrangement*. Princeton: Princeton University Press.

Nichols, Shaun, and Stephen Stich. 2003. *Mindreading*. Oxford: Clarendon Press.

Nisbett, Richard, and Lee Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, N.J.: Prentice Hall.

Nisbett, Richard, and Timothy Wilson. 1977. "Telling More Than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84: 231–259.

Nozick, Robert. 1981. *Philosophical Explanations*. Oxford: Oxford University Press.

Perry, John. 1979. "The Problem of the Essential Indexical." *Nous* 13: 3–21.

Plantinga, Alvin. 1993. *Warrant and Proper Function*. Oxford: Oxford University Press.

Pryor, James. 2001. "Highlights of Recent Epistemology." *British Journal for the Philosophy of Science* 52: 95–124.

Putnam, Hilary. 1981. *Reason, Truth, and History*. Cambridge: Cambridge University Press.

———. 1996. "The Meaning of 'Meaning.'" *Minnesota Studies in the Philosophy of Science* 7: 131–193.

Ramnani, Narender, and R. Christopher Miall. 2004. "A System in the Human Brain for Predicting the Actions of Others." *Nature Neuroscience* 7: 85–90.

Ryle, Gilbert. 1949. *The Concept of Mind*. New York: Barnes and Noble.

Sawyer, Sarah. 2002. "In Defence of Burge's Thesis." *Philosophical Studies* 107: 109–128.

Stich, Stephen, and Ian Ravenscroft. 1993. "What *is* Folk Psychology?" Technical Report 5. Piscataway, N.J.: Rutgers University Center for Cognitive Science.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185: 1124–1131.

Tye, Michael, and Brian McLaughlin. 1998. "Externalism, Twin Earth, and Self-Knowledge." In *Knowing Our Own Minds*. Ed. Crispin Wright, Barry C. Smith, and Cynthia Macdonald. Oxford: Clarendon Press, 285–320.

Umiltà, M. A., et al. 2001. "I Know What You Are Doing: A Neurophysiological Study." *Neuron* 31: 155–165.

Vogeley, K., et al. 2001. "Mind Reading: Neural Mechanisms of Theory of Mind and Self-Perspective." *NeuroImage* 14: 170–181.

Warfield, Ted A. 1992. "Privileged Self-Knowledge and Externalism are Compatible." *Analysis* 57: 232–237.

White, Peter A. 1988. "Knowing More About What We Can Tell: 'Introspective Access' and Causal Report Accuracy 10 Years Later." *British Journal of Psychology* 79: 13–43.

Wilson, Timothy D. 2002. *Strangers to Ourselves*. Cambridge: Belknap Press.

Wittgenstein, Ludwig. 1958. *Philosophical Investigations*. Trans. G. E. M. Anscombe. Oxford: Blackwell.