Jennifer Nagel, University of Toronto
March 5, 2012

Forthcoming in: *Alvin Goldman and his Critics*,
Hilary Kornblith and Brian McLaughlin, eds.; Blackwell.

Knowledge and Reliability

Great theories can have counter-intuitive consequences. When intuition clashes with theory, sometimes the best course is not to reject the theory but to argue that something is wrong with the intuition. This paper examines the best-known intuitive counterexamples that have been pressed against Alvin Goldman's reliabilist theory of knowledge, and argues that something is wrong with them. Under close scrutiny, the scenarios that internalists take to undercut reliabilism are ambiguous. Interestingly enough, on either way of resolving the ambiguity, these scenarios fail to present us with clear instances of unjustified but reliably formed belief. In what follows, I argue that on the most natural way of reading the internalist cases, the beliefs they invite us to evaluate are not in fact reliably formed: although the cases describe a series of true judgments, these judgments only happen to be true, and are not the products of a process that tends to hit the truth in the way that reliabilism requires. With some effort, it is possible to find a way of construing the cases so that the relevant beliefs actually are reliably formed; however, this way of reading the cases makes it difficult to conduct an intuitive evaluation of the justification of those beliefs, for reasons having to do with certain natural limitations on our ability to track the mental states of others. The intuitive appeal of the internalist cases arises in part from their ability to coax us into a self-conscious frame of mind in which it is difficult to judge less self-conscious belief formation with appropriate accuracy.

Although the main aim of this paper is to argue that our intuitions do not tell against reliabilism, a concluding section discusses the possibility that these intuitions might accord equally well with a more extreme externalist view, Williamson's 'knowledge-first' approach. The paper does not attempt to adjudicate between these programs; however, it observes that the move used against the internalist cases on behalf of reliabilism in the first half of the paper is quite similar to a move that could be used against reliabilism by an advocate of the more radical knowledge-first view. If the illicit intuitive appeal of the internalist cases is traced back to the distorting influence of a certain self-conscious frame of mind, this same self-conscious frame of mind can also be identified as a factor that would push us towards a belief-centered epistemological theory such as reliabilism, and away from the knowledge-first approach. Reliabilism has been examined largely in contrast to internalism, but its strengths and weaknesses arguably come into sharper focus if we compare it with more radical forms of externalism as well.

1. Intuitive counter-examples to reliabilism?

Not just any true belief counts as knowledge. The core idea of Goldman's original reliabilist theory of knowledge is that a true belief attains the status of knowledge just when it is produced by a reliable cognitive mechanism, where reliability is understood to consist in a tendency to produce true beliefs (Goldman 1976). There is something appealing about the idea that there is a modal connection between truth and knowledge: the person who knows does not merely happen to get it right on this occasion, but somehow thinks in a way that has some deeper tendency to hit the truth.

Many worries have been raised about the details of reliabilism, about the question of how exactly the relevant cognitive mechanisms are to be individuated, or what should count as reliability or a tendency to hit the truth. But one of the earliest and best-known lines of attack on reliabilism challenges the position's core commitment: according to internalists, no matter how we

flesh out the details, we will be missing something important about knowledge if we retain reliabilism's singular focus on tending to get things right.

What is missing, according to the internalist, is a special place for the subject's point of view in the generation of knowledge. In support of this view, internalists have developed several well-known cases in which subjects with a strong tendency to be right about something lack accessible evidence of this tendency: they make accurate judgments without being in a position to know or even reasonably judge that they are reliable. Intuitively, these blindly accurate subjects seem to lack knowledge.

Close examination of the exact formulation of a sample of these cases may help us to figure out why we respond to them as we do. The first example is BonJour's psychic case, originally presented in (Bonjour 1980), but more recently clarified and rephrased as follows:

> Suppose then that Amanda is a reliable clairvoyant under certain specific conditions with respect to a particular range of subject matter. Owing perhaps to some sort of causal process that has so far eluded scientific investigators, beliefs about that subject matter now and then occur spontaneously and forcefully to Amanda under those conditions and such beliefs are mostly or even perhaps invariably true. Amanda, however, though she generally accepts the beliefs in question, has never checked empirically to see whether any of them are true, nor has the potentially available empirical evidence for the truth of any of the specific claims and in consequence for her general reliability been supplied to her by others. (BonJour and Sosa 2003, 28)

BonJour concludes that beliefs formed in this manner could not constitute knowledge. Considering an elaboration of the case in which it is explicitly added that the subject lacks any reason to think she is unreliable, BonJour contends that this subject "is still being epistemically irrational and irresponsible in accepting beliefs whose provenance can only be a total mystery to her, whose status is as far as she can tell no different from that of a stray hunch or arbitrary conviction" (2003, 32).

A broadly similar case was developed by Keith Lehrer:

> Suppose a person, whom we shall name Mr. Truetemp, undergoes brain surgery by an experimental surgeon who invents a small device which is both a very accurate thermometer and a computational device capable of generating thoughts. The device, call it a tempucomp, is implanted in Truetemp's head so that the very tip of the device, no larger than the head of a pin, sits unnoticed on his scalp and acts as a sensor to transmit information about the temperature to the computational system of his brain. This device, in turn, sends a message to his brain causing him to think of the temperature recorded by the external sensor. Assume that the tempucomp is very reliable, and so his thoughts are correct temperature thoughts. All told, this is a reliable belief-forming process. Now imagine, finally, that he has no idea that the tempucomp has been inserted in his brain, is only slightly puzzled about why he thinks so obsessively about the temperature, but never checks a thermometer to determine whether these thoughts about the temperature are correct. He accepts them unreflectively, another effect of the tempucomp. Thus, he thinks and accepts that the temperature is 104 degrees. It is. Does he know that it is? Surely not. (Lehrer 1990, 4)

When a temperature thought occurs to Mr. Truetemp, Lehrer continues, "he has no idea why the thought occurred to him or that such thoughts are almost always correct. He does not, consequently, know that the temperature is 104 degrees when the thought occurs to him." (1990, 164)

There is something intuitively compelling about these cases: Amanda and Truetemp are perfectly accurate reporters with respect to some special subject matter, but it is at the very least controversial whether their mysterious accuracy is enough for knowledge. If reliabilism were the right theory, BonJour and Lehrer contend, then these cases should come across as clear examples of knowledge, but they do not.

Goldman himself reports sharing the internalist intuition that Amanda and Truetemp lack knowledge (1992; 1994). He volunteers an error theory to deflect the challenge posed by these cases: according to Goldman, our natural intuitions about knowledge and justification are generated by a two-step process in which particular cases are compared to familiar prototypes of good and bad belief formation. If a case closely matches a good type (whether perceptual, memorial or inferential), we evaluate it positively; if it matches a bad type, or simply fails to match any recognized type, we evaluate it negatively. Goldman further proposes that the underlying prototypes are divided into good and bad types according to their reliability. The cognitive

workings of Amanda and Truetemp are peculiar enough that their judgments fail to match any familiar positive prototype; however, according to Goldman, this result does not directly indicate that these judgments lack the feature that fundamentally matters to positive epistemic status. Reliability is what really matters in epistemic evaluation; indeed, it is because Amanda and Truetemp simply lack the familiar signs of reliability that our rough intuitive mechanisms assess them negatively.

Goldman grants to the internalists that Amanda and Mr. Truetemp are actually reliable in their belief formation, albeit reliable in a confusingly unfamiliar way. In his view these cases come across negatively because intuition only tests for reliability indirectly, via comparison to its familiar manifestations. This is an interesting suggestion, but it leaves certain questions unanswered. At least in his (1994), Goldman does not contend that Mr. Truetemp's thinking comes close to matching some 'vice' prototype of bad belief formation; his claim is rather that Mr. Truetemp's thinking simply fails to match what Goldman takes to be the closest positive prototypes, prototypes of perception. It fails to match in virtue of the stipulated absence of any conscious sensory phenomenology. What Goldman does not explain is why exactly we would have a *negative* intuitive response to such cases, as opposed to a failure to respond intuitively one way or the other.[1] Goldman claims, plausibly enough, that closely matching a prototypically 'vicious' mode of belief formation (like wishful thinking) would result in the intuitive sense that a belief is unjustified, but it is not obvious that sheer failure to match a positive prototype should have a similar effect.

One might have thought that we would simply have no clear intuitive response to belief-forming processes which were neither close to virtue-prototypes nor close to vice-prototypes;

[1] Goldman has elsewhere suggested that internalist cases involving hypothetical psychics might trigger a negative response in virtue of their association with 'scientifically disreputable' faculties (1992, 1993). This suggestion is not unreasonable, but because it does not generalize very smoothly to the superficially similar case of Mr. Truetemp, one wonders whether a better explanation of our negative response to these cases could be found.

Goldman does not explain why our default response would be negative, rather than blank, or even positive. Some background pressure towards a positive evaluation could arguably arise from our general psychological inclination to over-attribute knowledge, especially to those who are forming beliefs we know to be true (on this point, see e.g. Nickerson 1999). Indeed Goldman himself is somewhat tentative in his line on these cases: having made an initial claim that failure to match a positive prototype will make it intuitive that a belief is unjustified, Goldman later says more conservatively that failure to match will generate 'a certain measure of reluctance to judge that the belief is justified' (1994, 315).

If we feel not just reluctance to evaluate positively, but something decidedly negative towards the belief formation of Amanda and Mr. Truetemp, perhaps reliabilists need a stronger explanation of why this is so. The most direct way of defending reliabilism against these cases would be to deny that they are counter-examples to the theory in the first place: perhaps the subjects of these cases, as we most naturally understand them, are actually unreliable in their belief formation. It may seem like obtuse resistance to stipulation to say such a thing, but in what follows I'll argue that, despite the involvement of some reliable components in their thinking, each of these subjects is most naturally evaluated as forming beliefs in an unreliable manner. The reason for this is that both cases prompt us to think of the subject's belief-formation as though it were self-conscious, while simultaneously stipulating that these subjects are violating a core condition on reliable self-conscious belief formation. The intuitively unjustified belief in these cases is not reliably formed. The power of the cases is secured in part by certain intuitive difficulties in representing unselfconscious belief formation, but I will argue that the cases fail to show that unselfconscious belief formation is intuitively unjustified; they also fail to show that the core condition of reliable self-conscious belief formation is a condition of either reliable belief formation or justified belief formation in general.

2. Believing and accepting

The Amanda and Truetemp examples share a common feature: each of the cases describes the subject as passing through a stage of *acceptance,* where this stage is distinct from the reliable psychic or thermo-sensor driven process that supplies the propositional content to be judged. In the Truetemp case the division between these stages is particularly clear: the tempucomp generates accurate thoughts about the temperature, and then, although Mr. Truetemp is puzzled by the intrusion of these odd thoughts, 'he accepts them unreflectively, another effect of the tempucomp.' Paying close attention to Lehrer's description of the case, we see that the belief-forming process is not complete when the original temperature thoughts are generated: Lehrer characterizes the device as 'causing [Truetemp] to think of the temperature recorded by the external sensor', but thinking of something is not the same as believing that something is so. Many attitudes other than outright belief are compatible with entertaining a given thought about the temperature. Although Lehrer says of the first stage that it is 'all told … a reliable belief-forming process', it is not until he mentions Mr. Truetemp's unreflective acceptance of the thoughts about the temperature that Lehrer actually characterizes Mr. Truetemp as endorsing those thoughts, attributing to Mr. Truetemp a belief that the temperature is 104 degrees. Here the reliabilist might wonder whether our negative intuitions about the case are being generated by the pivotal second stage of Mr. Truetemp's judgment—his 'unreflective acceptance' of the puzzling intrusive thoughts. If, as will shortly be argued, this kind of unreflective acceptance is naturally seen as an unreliable method of belief formation, then the whole process that terminates in Mr. Truetemp's endorsement of a proposition about the temperature is not a reliable one. The fact that the surgeon's manipulation produces a string of true beliefs in Mr. Truetemp does not ensure that these beliefs are being generated by a reliable process: reliability is a modal notion, and not a diachronic one.

What matters as far as reliabilism is concerned is not simply whether Mr. Truetemp's many temperature beliefs are true, but whether they are generated by the kind of process that tends to produce true beliefs.

BonJour's cases feature a similar emphasis on what is accepted: Amanda does not simply form spontaneous beliefs, but in addition it is stipulated that "she generally accepts the beliefs in question", notwithstanding her lack of accessible evidence for their truth or for her own reliability. We are told something about why Amanda initially forms the beliefs—some odd causal process produces them—but we are not told exactly why she subsequently accepts those beliefs, although we are given reasons to worry that something is seriously amiss with whatever processes are operative at the stage of acceptance. Indeed, for Amanda as for Mr. Truetemp there is some reason to wonder whether the initial (pre-acceptance) stage is more easily seen as a stage in which ideas simply come to mind, rather than a stage in which full-blown beliefs are already formed: when we read that "beliefs about that subject matter now and then occur spontaneously and forcefully to Amanda", it is easy to imagine her as just experiencing some moments of inner speech, or visual imagery, or something along those lines, and not entirely obvious that what she has at this stage are already the beliefs whose legitimacy we are to evaluate.

If subjects like Amanda and Mr. Truetemp do not even seem to *have* outright beliefs until they engage in acceptance or endorsement, then the truth-conduciveness of those processes of acceptance or endorsement will have to matter to the reliabilist evaluation of their beliefs. But even if these subjects are already seen as having beliefs prior to the acceptance stage, if we are cued to evaluate the legitimacy of these subjects' beliefs only after the acceptance stage, then the contribution of the latter stage can still make a difference in our appraisals. On the latter understanding of the cases, there is some potential for confusion about what is meant by expressions like 'the subject's belief': we might use that expression to pick out either Amanda's

initial mysterious impression or her subsequent endorsement of that impression. But any tendency to apply the same phrase as a label for either stage of Amanda's thinking should not be allowed to obscure the differences between them: if different cognitive processes underpin the two stages, then they can receive different epistemic evaluations from the reliabilist. As Goldman has stressed—e.g. in his (1979)—what matters in reliabilist epistemic evaluation is not simply the reliability of the process that originally produced a belief, but rather the reliability of the complete set of processes responsible for sustaining the belief through to the moment of evaluation.

To complicate matters further, there seems to be a larger ambiguity about the role played by acceptance in the internalist examples: a phase of acceptance might be seen as either actual or potential with respect to the formation of the belief to be evaluated, either as a causally necessary part of the formation of that belief, or as a merely hypothetical stage that the subject might pass through without alteration to the key belief's epistemic standing. In the latter way of reading the cases, the acceptance stage is not itself a part of the formation of the intuitively unjustified belief, but one potential diagnostic maneuver we can imagine the subject going through in order to reveal the epistemic faults of the original belief. It will be argued that either way of understanding of the role of acceptance will generate problems; because the problems are somewhat different, however, they will be examined separately.

3. Evaluating the moment of acceptance

We can first examine the consequences of taking the subject's acceptance as a causally necessary part of the formation of the beliefs we are supposed to evaluate. On this way of reading the cases, prior to acceptance the subjects may have some state of mind falling short of belief, or they may have some initial beliefs which are then actually endorsed by the subject in a way we are invited to evaluate. Either way, our focus will be on the subject's epistemic standing following the stage of

endorsement or acceptance, keeping in mind that the epistemic characteristics of this stage might differ from those of the initial impression. According to the reliabilist, Mr. Truetemp could be doing well when he forms his initial thought or belief that the temperature is 104 degrees, but doing badly when he subsequently accepts this thought or belief in the absence of any consciously available reasons.

If the reliabilist wants to argue that Mr. Truetemp's acceptance of his accurate thoughts will not be reliable, the challenge now is to explain what it is about acceptance that would make acceptance in the absence of consciously available reasons unreliable. A closer look at acceptance is in order. Although the odd psychic and thermometer-driven elements in Amanda and Mr. Truetemp's thinking are novel, I think acceptance is something more familiar; indeed its familiarity contributes substantially to the psychological realism of the cases, helping to engage our capacity for intuitive epistemic evaluation.

At least as the term is used in these cases, 'accepting' is not interchangeable with 'believing'. Lehrer himself spells out a positive theory of what he takes to be the character of acceptance, characterizing it as a particular type of purposive attitude formation, demanding the explicit evaluation of a proposition ideally guided by the goal of attaining the truth (2000, 13-4). Belief, on his view, is a weaker condition that is not necessarily attained through purposive thinking. Various aspects of Lehrer's way of drawing the contrast are questionable in ways that will not concern us here, but its core idea is that some judgments involve explicit, controlled evaluation—the kind of evaluation we engage in when we carefully weigh evidence for and against a certain hypothesis—while other judgments are made without this kind of attention. The controlled capacity to accept is something higher than the mere capacity to believe: so, for example, nonhuman animals might be seen as having beliefs, but it would be odd—or cutely anthropomorphic—to describe a nonhuman animal as accepting that something was the case.

Whether or not Lehrer has articulated it in the standard way, there is an easily recognizable psychological distinction in this neighborhood, a distinction which seems to me to play a key role in our understanding of the internalist cases. This is the distinction between the automatic, rapid and effortless form of thought known as System 1 or type-1 processing (henceforth S1), and the controlled, explicit and sequential form of thought known as System 2 or type-2 processing (S2) (for reviews, see Sloman 1996; Stanovich 1999; Evans 2007). There is some debate about the right way to draw the line between these two modes of thought, but it is now widely thought that the involvement of working memory is crucial (De Neys 2006; Evans 2009). Because the contents of working memory are generally available to consciousness (Baddeley 2007), S2 processing will make us conscious of a series of relevant contents en route to supplying an answer; S1 processing will not do this. S1 processing—the type of processing involved in ordinary face recognition, for example—operates through automatic mechanisms closed to introspective access. We see someone's face and immediately recognize her as a friend, but would be hard-pressed to list the qualities of her features in virtue of which we did so. This type of processing is involuntary; the recognition happens whether or not we want it to happen. Meanwhile, S2 processing—the type of processing involved in complex mental arithmetic, for example—runs through the bottleneck of our limited working memory capacity, and so is slower, but operates in a way that is both more controlled and more open to view. In the characteristically S2 task of long division, for example, we go through a sequence of conscious intermediary steps where the output of each intermediary step supplies input to the next (Sloman 1996; Kahneman and Frederick 2005). This kind of thinking is relatively controlled: we can, for example, decide to abandon an S2 task halfway through, say, giving up on the arithmetical exercise and directing our attention elsewhere. The reliability of S2 processing depends on sustaining our focus on the task and running appropriate operations on the consciously available content it presents.

Crucially, there are some problems to which either type of processing may be applied. So, we naturally compute 5 times 11 automatically and effortlessly, without consciousness of any series of stages, but if we are appropriately motivated we can run through an explicit and sequential verification of this answer digit-by-digit. In general, routine, low-stakes, familiar matters are more likely to be settled in S1; unusual, high-stakes and novel issues are more likely to elicit S2 thinking. Of particular relevance to the internalist cases, the task of source monitoring can be handled either by S1 or S2, where source monitoring is the process by means of which we assess the origin of our mental experiences, attributing a given content to a source such as recent or more distant memory, present sensation of one kind or another, or imagination. For example, when asked about a recently past event (what did you have for breakfast today?) the content that comes to mind needs to be verified as arising from the appropriate memory store in order to answer the question (Johnson, Hashtroudi et al. 1993). Under routine, low-stakes conditions we verify the source in S1, relying heuristically on the felt degree of vividness in the content and felt ease of retrieval (or familiarity). Source monitoring is a crucial process in belief formation. The confidence of our everyday judgments depends on it: for example, one's level of confidence in a recalled trivia fact is a function of the fluency with which it is recalled (Kelley and Lindsay 1993). Indeed, across the board, confidence in typical unselfconscious judgments, including perceptual judgments and social attitudes, is generated heuristically on the basis of metacognitive feedback from one's cognitive operations, in a way which is generally responsive to their reliability (Koriat 2011; Koriat and Adiv 2011).

However, when made to feel self-conscious (are you sure that was today's breakfast, and not yesterday's?) or placed under high-stakes conditions (the police are questioning you after a fatal poisoning), one applies more effort to the source-monitoring problem, for example by deliberately reconstructing the temporal context of the remembered event, assessing its various features for plausibility, and so forth. Whatever our mode of thought—whether we are say, remembering,

perceiving or inferring—peculiarities in activated content, or in its fluency of activation, naturally trigger this more effortful S2 thinking (Alter and Oppenheimer 2009; Thompson 2009). This self-conscious source monitoring provides a useful check on the operations of lower-level belief formation, both because it is triggered by signals that something may have gone wrong at the lower level, and because the global broadcast of what is in working memory can draw many different sources of potentially relevant information into solving the problem of figuring out what is really going on (Shanahan and Baars 2005).

The relationship between conscious thinking and reliability is not simple. Other things being equal, more cautious and systematic thinking does enhance reliability (Koriat and Goldsmith 1996; Lerner and Tetlock 1999; Stanovich 1999), but the elevation to S2 is not of course necessary for reliability (so, e.g. face recognition is generally accurate notwithstanding its automaticity). Meanwhile, systematic thinking has certain vulnerabilities of its own, for example when we are hasty or unreasonable in our handling of what is consciously available, either because our limited working memory is distracted by other tasks, or because our controlled cognition is controlled in the wrong way, say, swayed by some biasing motivation (Kunda 1990). There is no simple correlation between the presence or absence of S2 processing and the presence or absence of reliability. However, the contrast between the two types of thinking can still serve to clarify the reliabilist's position on the importance of what is available from the first-person perspective. In evaluating methods of belief formation strictly on the basis of their tendency to produce true beliefs, the reliabilist does not have to rule out any special consideration for first-person availability. Whenever there is a mode of belief formation whose reliable operation depends on the availability of appropriate material to consciousness, the reliabilist has an obvious reason to insist on the availability of appropriate material to consciousness. If the core characteristic of S2 thinking is that it operates on material present in working memory (and therefore consciously available),

then the reliability of those who are engaged in S2 thinking will depend on what is consciously available to them.

This psychological background enables a fresh reliabilist interpretation of the internalist cases. Amanda and Mr. Truetemp find themselves judging—or perhaps merely disposed to judge—certain odd propositions to be the case. Given the oddness of these propositions, and the stipulation that the inclination to endorse them arises from no ordinary faculty whose operations might be evaluated in the ordinary heuristic or automatic fashion, it is natural for these subjects to pass through a stage of explicitly or consciously monitoring their judgments. It is perfectly natural to see these subjects as not only having beliefs, but entering a distinct phase of accepting or maintaining them, just as Lehrer and BonJour's descriptions stipulate. We are familiar with this kind of evaluation as a process of weighing consciously available material. In the internalist cases, however, this phase is described as taking place without the help of appropriate consciously available material. Because explicit evaluations depend for their reliability on what is consciously available, when Amanda and Mr. Truetemp nonetheless endorse their initial impressions, they are thinking in a way that is not generally reliable.

The internalist might protest here that although Amanda and Mr. Truetemp's practices of acceptance are unusual, these practices have been altered in a way that should shield them from reliabilist criticism. Mr. Truetemp is not described as now generally failing to seek reasons for what he accepts, but quite specifically as a blind accepter of the temperature thoughts (this pattern of acceptance was described as 'another effect' of the tempucomp). If these thoughts are simultaneously guaranteed by the tempucomp to be accurate and to be accepted unreflectively, then the surgeon's manipulation does not really disable or compromise the reliability of Truetemp's cognition in any significant way, the internalist might contend. Granting that our ordinary ways of monitoring our thinking may be reliable partly in virtue of how they handle what is available to

consciousness, the internalist could still insist that even greater reliability could be produced by a special psychological compulsion to endorse the deliverances of a paranormally accurate faculty.

Whether this internalist rebuttal can succeed depends in no small measure on the way in which we naturally individuate the relevant belief-forming mechanisms in reading these cases, and how we see the relationship between the accuracy of the first stage and the operation of the second. One way of deflecting the internalist response here would be to point out that the cases leave open the possibility that it is merely coincidental that the subject has both an accurate faculty and a hard-wired tendency to accept its deliverances. If there is no necessary connection between the accuracy of the first faculty and the operation of the second, the overall process is still not of a type that tends to produce true beliefs. That is, we might see the final and decisive effect of the tempucomp as forcing Mr. Truetemp to accept just whatever readings the device initially sends him, regardless of their accuracy, so that if his device were to read overly hot or cold he would be equally stuck with the relevant false beliefs. Given that the tempucomp disables his ordinary conscious critical capacity as far as the temperature is concerned, he is just lucky that the tempucomp is supplying accurate information on that point. Equally, Amanda's acceptance of her paranormally-formed ideas happens to result in a string of true beliefs, but the character of her acceptance is under-described. The case leaves it open that we might for example think of her as being inclined to believe the deliverances of any novel faculty, whether or not this faculty tracks the truth; indeed the impression that she is simply gullible might be underscored by BonJour's emphasis on her epistemic 'recklessness and irresponsibility'. She might be like Goldman's Humperdink, who very capriciously selects what is in fact a good algorithm for solving a class of problems, and then answers every problem in the set accurately: given his initial caprice, his entire belief-forming process is unreliable, despite its production of a string of true beliefs (Goldman 1979). Amanda's 'algorithm' is a sub-personal faculty capriciously planted in her, together with a process that

disarms her psychologically normal self-monitoring; the sub-personal faculty happens to be accurate, but the belief-forming process as a whole is not reliable.[2]

Persistent internalists might at this point concede that our intuitions about the original BonJour and Lehrer cases could be explained by a sense that the subjects in these particular cases are thinking in an unreliable manner, but then attempt to devise modified cases in which the accuracy of the first-stage manipulation is essential to the operation of the second stage. Perhaps the benevolent scientist has designed the tempucomp to compel Mr. Truetemp to accept the temperature thoughts just when the thermometer is reading accurately; some failsafe mechanism will switch it off altogether if the thermometer ever loses accuracy. To make their point against reliabilism, meanwhile, internalists will have to stipulate that the accepting subject still has no consciously available evidence of the accuracy of his paranormal belief-forming process; the process of acceptance will depend perhaps causally on this accuracy, but will not work by making this accuracy available to consciousness.

The reliabilist needs a deeper argument to neutralize these revised cases. One possible strategy would be to resist their tacit assumptions about the intuitive individuation of belief-forming processes. The revised cases assume that the relevant beliefs are formed on the basis of a specially modified process of acceptance which is not governed by consciously accessible content, but not just any process will register as an intuitively imaginable form of acceptance. In the every-day course of our thinking, where our epistemic intuitions have been trained in the first place, the process of acceptance works through the operations it performs on consciously available content. The phenomenal grasp we have on acceptance does in fact capture an essential feature of this way of thinking: as an S2 process, acceptance really does need to operate through manipulations of what

[2] This way of understanding the cases has something in common with John Greco's approach, according to which, "Truetemp accepts the truth because the *tempucomp* is reliable, not because *he* is reliable" (2003, 40). It may well feel natural for us to identify the agent himself with his explicit or personal-level reasoning. I am not convinced that the specific commitments of agent reliabilism are required to make sense of our overall pattern of intuitions, however.

is in working memory. So, if acceptance in the form we will find intuitively recognizable is really going to depend on the accuracy of a lower-level belief-forming mechanism, the accuracy of this lower-level belief-forming mechanism needs to make itself available to consciousness.[3] This is not to say that we would need to have conscious awareness of the fact that the lower-level mechanism is accurate; it would be enough to have the kind of conscious phenomenology that ordinarily serves as a guide to our acceptance of the deliverances of lower-level mechanisms, the kind of phenomenology that is stipulated to be absent in the internalist cases. As long as we are looking at the relevant process as a form of acceptance, the modal facts around its operation depend on what is available to consciousness—our stipulation that the acceptance will cease if the temperature readings fail to be accurate is arbitrary, as far as the nature of acceptance is concerned; we could equally well have stipulated that the acceptance would cease if the temperature readings were ever anything other than five degrees too hot.

Alternatively, there is another reliabilist strategy for managing both these revised cases and the original ones, a more radical strategy which does not insist that the relevant process is actually a form of explicitly reasoned acceptance. This will be explored in the next section.

So far, we have been concerned to adopt a reading of the cases on which an explicit process of acceptance is a vital part of the beliefs we are invited to evaluate. Because the stage of acceptance involves a kind of thinking whose reliability depends on the conscious availability of appropriate materials, the reliabilist can agree with the internalist that justification in these cases demands the conscious availability of appropriate materials. However, the reliabilist does not need to insist that justification always depends on what is consciously accessible to the subject, exactly

[3] As a matter of empirical fact, it seems that consciously available evidence of the accuracy of a belief-forming mechanism actually does have to accumulate before we can reflectively accept its deliverances (on this point see Beebe, 2004). Indeed, there is evidence that our ordinary capacity for self-monitoring is the product of learned associations between various consciously accessible characteristics of our own cognitive processing and reinforcement of the deliverances of that processing (e.g. Unkelbach, 2007). For further discussion of the relationship between metacognition and endorsement, see (Michaelian, forthcoming).

because not all belief formation is a function of operations on consciously available material. When beliefs are formed through S1 processing their reliable formation does not necessarily depend on what passes through consciousness.[4] In this spirit the reliabilist can draw our attention to the fact that we do not ordinarily demand self-conscious reflection on the deliverances of perception, memory or testimony in order to see beliefs formed on these bases as justified; the reliabilist can also draw our attention to the threat of vicious regress lurking in such demands (cf. Bergmann 2006). According to the reliabilist, there is no special reason to favor reliability-as-secured-by-consciously-available-reasoning over reliability secured otherwise (on this point, see Kornblith 2010; Kornblith MS). On a reading of the internalist cases which stresses the psychologically distinctive character of conscious acceptance, the reliabilist can happily agree with BonJour's assessment that a blindly psychic subject would be "epistemically irrational and irresponsible in accepting beliefs whose provenance can only be a total mystery to her", without conceding that this difficulty at the level of acceptance constitutes any problem for the epistemic status of those original beliefs themselves. There is nothing wrong with blind reliability, even when there would be something wrong with its reflective endorsement.

4. Explicit acceptance as a merely hypothetical diagnostic device

There is another way to read the internalist cases, however. Rather than focusing on the psychological reality of the acceptance stage, and having to grant that the reliabilist can agree that what is consciously available matters there, the internalist might suggest instead that explicit

[4] There are of course various other ways of combining reliabilism with respect for evidence. For example, Juan Comesaña has advocated a form of 'evidentialist reliabilism' in which justified beliefs must always be based on evidence, where the appropriate type of reasoning from evidence to belief is reliable (Comesana 2010). Comesaña supports the demand for evidence by appeal to the intuitive pull of the internalist cases, rather than seeing evidence as something whose value is secured exactly by its contribution to reliability. However, if we can account for the internalist cases without positing a need for evidential input across the board, then it would be more economical for the reliabilist just to allow that evidence is needed in some conditions for the sake of its contribution to reliability.

acceptance need not actually figure as a stage in the formation of the beliefs we are invited to evaluate. On this reading of the cases, Amanda and Mr. Truetemp are already in trouble at the first stage, when they have formed beliefs on some non-evidential basis; considering how poorly they would fare on reflection is strictly a hypothetical device, a possible diagnostic measure that might be undertaken to reveal the unacceptability of the first-stage beliefs themselves. On this reading the internalist concedes that we don't expect everyone to reflect self-consciously on what they believe at every moment in order to count them as justified; what matters is just the truth of some subjunctive conditional of the form *if she were to reflect on this belief, she would have appropriate grounds to endorse it reflectively.* Whether or not the subject actually does reflect on or consciously endorse her belief, she can count as justified only if this conditional holds true.

In support of this reading of the cases, BonJour might direct our attention to his character Norman, who is not initially described as passing through any stage of acceptance:

> Norman, under certain conditions that usually obtain, is a completely reliable clairvoyant with respect to certain kinds of subject matter. He possesses no evidence or reasons of any kind for or against the general possibility of such a cognitive power, or for or against the thesis that he possesses it. One day Norman comes to believe that the President is in New York City, though he has no evidence either for or against this belief. In fact the belief is true and results from his clairvoyant power, under circumstances in which it is completely reliable. (Bonjour 1980, 62)

Immediately after presenting the scenario, BonJour raises the question of whether Norman's belief constitutes knowledge. He does not immediately answer this question, but proceeds to offer two distinct ways of elaborating the case, both of which he takes to support a negative answer. First, he imagines a version of the scenario in which Norman further believes without evidence that he has a clairvoyant power like the one he in fact possesses, and that this higher-order belief of Norman's 'contributes to his acceptance of his original belief about the President's whereabouts in the sense that were Norman to become convinced that he did not have this power, he would also cease to accept the belief about the President.' BonJour contends that the higher-order belief would be

'obviously irrational' because it is neither evidentially supported nor (BonJour stipulates) even reliably formed; he then concludes that the charge of irrationality must also be brought against 'the belief about the President which *ex hypothesi* depends on it.'

The character of this dependence is somewhat unclear. BonJour is not claiming that the original belief about the President was generated by some reasoning that included reliance on the problematic higher-order belief: indeed, in an endnote BonJour stresses that the original belief is non-inferential, and that the higher-order belief 'is not in any useful sense Norman's reason for accepting that specific belief.' (1980, 72) It is not entirely clear how to reconcile this last remark with the claim that the higher-order belief 'contributes to' Norman's acceptance of the belief about the President, in the sense that its absence would spell the end of that acceptance. Perhaps Norman's irrational belief about his powers somehow impedes him from giving his belief about the President's whereabouts serious rational scrutiny; rather than positively supporting his lower-level belief, the higher-order belief may simply stop him from entering a more self-critical frame of mind in which the lower-level belief would come to seem problematic to him. In any event, the main thrust of the argument here is then that the lower-level belief should be seen as unjustified because its continued maintenance depends on the possession of an unjustified belief. However, Norman's initial formation of his belief about the President is not itself directly criticized.

BonJour then considers a more radical version of the scenario, in which Norman does not even believe that he has clairvoyance. Again BonJour imagines the possibility of Norman engaging in some self-conscious moment of reflection:

> But if this specification is added to the case, it now becomes more than a little puzzling to understand what Norman thinks is going on. From his standpoint, there is apparently no way in which he *could* know the President's whereabouts. Why then does he continue to maintain the belief that the President is in New York City? Why is not the mere fact that there is no way, as far as he knows or believes, for him to have obtained this information a sufficient reason for classifying this belief as an unfounded hunch and ceasing to accept it?

> And if Norman does not do this, is he not thereby being epistemically irrational and irresponsible? (1980, 62-3)

In answering his own rhetorical question, BonJour says that "Norman's acceptance of the belief about the President's whereabouts is epistemically irrational and irresponsible, and thereby unjustified." (1980, 63) When we attend to what Norman would do under reflection, BonJour stresses that there would be a problem if Norman were to "continue to maintain" his original belief: it is emphasized that it would be very bad of him to do such a thing. What is striking in these passages is that the hypothetical phase of reflection is still not quite presented as testing the legitimacy of Norman's original belief. The original belief is never made the focus of criticism: BonJour invites us to imagine Norman reflecting, and elicits the intuition that Norman would be going wrong if he continued to maintain the belief on reflection, but he does not overtly draw the conclusion that this shows Norman's original belief about the location of the President to have already been epistemically unjustified at the time of its formation. So Norman is never described as 'irresponsible' for having formed his belief about the President in the first place, for example, although it is stressed that it would be irresponsible of him to maintain it under reflection. We can intuitively sense that the belief about the President would seem unwarranted from the standpoint of a reflective Norman—indeed from that perspective he might come to classify it (incorrectly) as an 'unfounded hunch'—but we are never explicitly directed to go back and re-consider Norman's original unselfconscious state of mind and find fault with it. The intuitive cases do not give us reasons to accept the internalists' subjunctive conditional about justification; rather, they presuppose its truth.

There is a natural explanation why Norman, Amanda and Mr. Truetemp cannot be faulted for the original formation of their preternaturally accurate beliefs: these beliefs are not the product of controlled cognition, and the charge of irresponsibility can only stick where there is some

possible degree of control. When we are thinking systematically, directing attention to the contents of working memory, we can be distracted or fail to control our attention appropriately; one can be irresponsible in generalizing hastily from a few data points, for example, or in doing a complex arithmetical calculation carelessly, with somewhat divided attention. But the kind of thinking that produces answers automatically cannot itself be irresponsible; because it does not depend on cognitive effort we do not have the same sense that supplying additional effort would have made it better. If the notion of epistemic justification is tied to the regulation of epistemic effort, then it has no clear application to involuntary cognition as such: it is not as though by trying harder Norman, Amanda or Mr. Truetemp could cease to form their paranormal beliefs. There is nothing epistemically objectionable about having a hunch, even when there would be something epistemically objectionable about endorsing it on reflection.

The internalist cases are supposed to make the point that beliefs cannot be justified in the absence of consciously accessible reasons; however, they invite us to focus on the epistemic state of subjects who are thinking reflectively, and indeed actively searching for reasons in support of what they believe. It is easy for us to do this because the judgments being made by these subjects involve exactly the sorts of unexpected propositions that would naturally make us stop and double-check ourselves. In this self-critical frame of mind, the absence of consciously available reasons would indeed be a problem; what has not been shown is that consciously available reasons are always required whether or not one is presently engaged in self-criticism. It has furthermore not been shown that the self-critical frame of mind is itself required for the production of justified beliefs.

BonJour does offer the following thoughts about the self-critical frame of mind: "Part of one's epistemic duty is to reflect critically upon one's beliefs, and such critical reflection precludes believing things to which one has, to one's knowledge, no reliable means of epistemic access" (Bonjour 1980, 63). Reasonably enough, BonJour does not here attempt to argue that critical

reflection must be one's constant mode of thought: it is perfectly compatible with its being one of our duties that we only reflect critically on our beliefs from time to time. But even intermittent reflection is a state of mind about which the reliabilist might have mixed feelings. The thought that reflection will preclude our believing things to which we have no known reliable means of access might be a selling point for reflection, if one were antecedently committed to the view that knowledge of one's means of epistemic access is a requirement for epistemic justification. But those who are not yet sold on internalism might be less pleased by the thought that reflection can stop one from forming beliefs by means of processes of (as yet) unknown accuracy. For a reliabilist, the benefits of being barred in this manner from having certain unreliably formed beliefs would have to be weighed against the costs of being barred from having reliably formed beliefs as long as the reliability of their formation remained unknown. The self-critical frame of mind brings dangers as well as rewards, and the reliabilist can hold that there are some circumstances in which it would be good to avoid it. Going back to the first version of the Norman scenario, in which his unjustified higher-order belief stops Norman from being self-critical, the reliabilist could part company with the internalist and characterize this failure to reflect as a happy accident. From the reliabilist perspective, Norman's poorly-founded thought about his powers fortunately enough enabled him to make use of his paranormal faculty to gain accurate beliefs about the President's whereabouts. BonJour's cases do succeed in eliciting intuitions that a self-critical frame of mind would result in Norman's abandoning his belief; BonJour has not however given an independent argument to support the claim that a self-critical frame of mind always yields a better epistemic outcome.

This section has argued that there is something strangely off-target in BonJour's attempts to use a hypothetical process of acceptance as a diagnostic of the justification of unreflectively formed beliefs. Our conclusions about what Norman should do if he were being self-conscious are never directly related to the propriety of his unselfconscious beliefs. The interesting question is why it is not immediately evident to us that our imaginative exercises with the internalist scenarios end up

answering a question somewhat different from the question to which we were originally promised an answer. One possible reason is that reading the elaborated cases puts us in a state of mind in which it is naturally difficult to appreciate or even register the state of mind we originally wanted to investigate. In general, we have difficulty representing the state of mind of more naïve agents; the fundamental bias of mental state ascription is a bias towards egocentrism, in which privileged concerns tend to be projected onto others, without our realizing it (for reviews, see Royzman, Cassidy et al. 2003; Birch and Bloom 2004; Apperly 2011). Once we are worried about how a reflective Norman would search his mind in vain for higher-order support for his belief about the President, it will be hard for us to evaluate Norman's actual unreflective belief formation accurately, even if we try. We would have a natural tendency to evaluate Norman as though he were also engaged in self-scrutiny, despite our awareness of the explicit stipulation that he is not (for more detailed argument on this point, see Nagel 2010; Nagel 2012). If these cases still elicit the intuitive sense that Norman has gone wrong from the start, these intuitions need to be handled with caution: given that they are triggered by manipulations that generally lead to inaccurate mental state representations, their evidential value is dubious. As long as they trigger a self-conscious frame of mind in the reader, internalists can exploit a vulnerability of our natural capacity to evaluate the mental states of others: they can make it seem that all belief formation should pass the tests appropriate to self-conscious belief formation. Whether or not those tests really are decisive is a question that is not well answered by the intuitions produced in this manner.

5. Reliabilism versus another rival

So far the argument has been strictly defensive: the aim has been to establish that certain celebrated internalist cases provide no clear reason to reject reliabilism. This defense partially accommodates the internalist point that there is a special place for the subject's point of view in the generation of knowledge: as long as subjects are thinking in a manner whose reliability depends on

the availability of appropriate material in consciousness—and much of our thought does work this way—then their beliefs should be supportable by consciously accessible reasons. The elegance of the reliabilist position is that it can explain the value of consciously accessible reasons in terms of their contribution to reliability, in the broader context of an epistemological theory that insists uniformly on the value of reliability in all belief formation.

Looking at the broader context, one useful contribution of reliabilist epistemology has been in the range of cases it has offered up for intuitive assessment. Breaking away from a tradition of self-conscious evaluation of difficult cases, Goldman has drawn to our attention cases in which we would not ordinarily be self-conscious—say, remembering a well-known trivia fact such as Lincoln's birthdate (in Goldman 1967); he has also paid special attention to our third-person assessments of others making routine perceptual judgments (e.g. in Goldman 1976). If there is a risk of misrepresenting others' thinking when we enter an especially self-conscious frame of mind, then by focusing on these mundane cases of unselfconscious thought we can give intuition a firmer footing across a wider range of circumstances. Our natural tendency to evaluate these cases positively does seem to put pressure on views according to which positive evaluation should always demand some more elaborate or deliberate kind of reasoning. One reason why it is important for the reliabilist to account for the internalist cases is that reliabilism itself leans on intuition: the theory gains much of its plausibility from its agreement with intuitive responses to a wide range of particular cases.

Someone impressed by the fit between reliabilism and epistemic intuition might still wonder exactly how much support intuition can provide for reliabilism. Even if reliabilism can be fully defended against the cases advanced by Lehrer and BonJour, there could be other, more effective, intuitive counter-examples waiting in the wings. A further worry concerns the inherent limitations of the intuitive method: perhaps the absence of solid intuitive counterexamples to

reliabilism could be explained by something other than the truth of reliabilism. It is this latter worry that I want to explore, briefly, in this concluding section.

The original aim of reliabilism was to 'specify in non-epistemic terms when a belief is justified', or more broadly, when a belief has positive epistemic status (Goldman 1979, 90). In contrast to theories which analyze knowledge in terms of 'good reasons' or 'the right to be sure', reliabilism looks to furnish 'an account of knowing that focuses on more primitive and pervasive aspects of cognitive life' (Goldman 1976, 791). The starting point for reliabilism—the primitive 'non-epistemic term' on which the theory is built—is *belief*, and the natural condition which distinguishes the beliefs which are justified or amount to knowledge is their having been formed by a process that tends to yield true beliefs. As is well-known, we face a difficult problem in explaining how the relevant processes are individuated. If an account of knowledge is to be constructed from strictly non-epistemic materials, it needs to be able to identify the process responsible for a given belief without already using our understanding of the nature of knowledge to do so.[5] There is a live question about whether our intuitive individuation of the processes responsible for belief formation is actually 'non-epistemic'. It would be good news for reliabilism if it were: reliabilism would be capturing something very significant about the generation of our epistemic intuitions if we did in fact naturally ascribe knowledge by recognizing an agent as having some belief and then evaluating the process that gave rise to it as a reliable one, on the basis of purely non-epistemic features of that process and the agent's environment. But this is not the only way our intuitions might work, and one possible limitation on the support they could provide to reliabilism can be explored by looking at another model of intuitive assessment.

This rival model is associated with a program in epistemology which has an interesting resemblance to reliabilism: Williamson's 'knowledge-first' program. The core condition of

[5] For a particularly clear statement of this challenge to reliabilism, see (Brewer 1999, ch.4).

reliabilism is a modal one: knowledge requires thinking in a way that tends to yield true belief. The core commitment of Williamson's view also has a truth-centered modal element: knowledge is distinguished as the most general factive state of mind, where factive states of mind are 'states whose essence includes a matching between mind and world' (Williamson 2000, 40). Again, knowing is not simply happening to hit the truth; the person who knows is in an essentially (as opposed to just accidentally) correct state of mind. One key difference between these rival views concerns the relationship between knowledge and belief: for the reliabilist, the state of belief is our starting point in epistemology, and knowledge is seen as a special type of belief—the type that not only *is* but also *tends to be* right. For Williamson, knowledge is the starting point in epistemology: the person who believes is doing something which in some sense approximates or aspires to the condition of knowing, and might fall short of that condition in any number of ways. Knowledge does also entail belief in his view, but the knower is not simply a believer who meets some set of further conditions themselves specifiable in non-epistemic terms.

For an advocate of the knowledge-first view, it is not surprising that we have difficulties producing robust intuitive counter-examples to reliabilism. If we try to build a theory of knowledge by starting with belief, we are starting with a state of mind that is not essentially factive: after all, we can believe propositions that are not true, or whose truth is not essential to our believing them. From this field of states that are roughly knowledge-like but might or might not tend to be right, reliabilism selects as worthy of positive epistemic evaluation just the ones that tend to be right. From a knowledge-first perspective, sorting beliefs by their tendency to be right would be the best possible step in the direction of recapturing the distinctive character of knowledge (its essential correctness). What the advocate of the knowledge-first view will expect, however, is that the reliabilist will not find a natural or illuminating non-epistemic way to individuate the processes that tend to be right. On this view, our individuation of the relevant processes is made possible by our understanding of knowledge, and not vice-versa. If we naturally come up with ways of

individuating belief-forming process types that save reliabilism from counter-examples, this is because our intuitive identification of ways of thinking are shaped in the first place by our sense of what it is to know something.

When we intuitively recognize someone as knowing something, it is not obvious that we start by recognizing that person as having a belief—a representational state that might or might not be correct—then identify the process responsible for it, and then assess the reliability of this process. Our default understanding of others might presume that they have states of mind that essentially reflect various features of our shared environment, and there may be a restricted range of circumstances in which we naturally see others as having states of mind that are potentially out of line with reality. Whether our starting point is knowledge or belief seems to me to be an extremely difficult question in the theory of mental state ascription, and not something to be tackled here (for some discussion, see Nagel 2012). For present purposes, it will be enough to focus on one factor that could be clouding our view of this question, not least because this factor was an active ingredient in the internalist cases discussed above.

The cautionary lesson of the internalist cases was that the self-conscious first-person perspective can distort our understanding of an agent's state of mind. One risk of the self-conscious frame of mind is that it can make self-conscious thinking seem like the only good path, and one of the attractions of reliabilism as a program in epistemology was its invitation to step back to the unselfconscious third-person perspective. But another possible risk of the self-conscious perspective is that it may make it seem to us that our starting point in epistemology must always be belief, rather than knowledge. From the inside, when we reflect on our commitment to any particular proposition, we can typically raise the question of whether it might be wrong, and come to see our mental state as potentially out of line with reality. The problem may be worsened by a natural tendency to see ourselves as having privileged, transparent access to our own mental

states.[6] But if it appears to us in some circumstances that no state of mind could be essentially right, and that all our mental states at their most essential level must have the potential to be wrong, this appearance might be an understandable product of the circumstances, rather than a clear guide to how things are.

Goldman observed in the introduction to his 2006 book *Simulating Minds* that there has been little contact between epistemology and the theory of mental state ascription, and reported that he had no immediate plans to bring these research projects together (Goldman 2006, 10). In his more recent work, most notably his Romanell Lecture of 2010, Goldman has taken up the project of showing how empirical work on mental state attribution can be useful to epistemology. In particular, he has argued that to the extent that epistemologists derive evidential support for their theories from intuitions about mental states, they can benefit from learning about natural limitations on our intuitive capacities to recognize mental states (Goldman 2010). But it is also possible that work in epistemology can be useful to our understanding of mindreading. Perhaps the strategy that enabled Goldman to advance reliabilism in epistemology could be applied to a core problem in mindreading, the problem of the relationship between belief and knowledge attribution. It is presently an open question whether such efforts could succeed; it is also an open question whether such an application of the externalist strategy could ultimately lead us to reject reliabilism in favor of a more radical alternative.[7]

[6] For a detailed argument that we do not have such access, but do have the illusion of having it, see (Carruthers 2011).

[7] Thanks to Sergio Tenenbaum for comments and discussion, and thanks to the Social Sciences and Humanities Research Council of Canada for funding my research.

References:

Alter, A. and D. Oppenheimer (2009). "Uniting the tribes of fluency to form a metacognitive nation." Personality and Social Psychology Review **13**(3): 219.

Apperly, I. (2011). Mindreaders: The Cognitive Basis of "Theory of Mind". Hove and New York, Psychology Press.

Baddeley, A. D. (2007). Working memory, thought, and action. New York, Oxford University Press.

Beebe, J. R. (2004). "Reliabilism, Truetemp and New Perceptual Faculties." Synthese: An International Journal for Epistemology, Methodology and Philosophy of Science **140**(3): 307-329.

Bergmann, M. A. (2006). Justification without awareness: a defense of epistemic externalism. New York, Oxford University Press.

Birch, S. and P. Bloom (2004). "Understanding children's and adults' limitations in mental state reasoning." Trends in cognitive sciences **8**(6): 255-260.

Bonjour, L. (1980). "Externalist Theories of Empirical Knowledge." Midwest Studies in Philosophy **5**: 53-74.

BonJour, L. and E. Sosa (2003). Epistemic Justification: Internalism vs. Externalism, Foundations vs. Virtues. Malden, MA, Blackwell.

Brewer, B. (1999). Perception and reason. Oxford, Oxford University Press.

Carruthers, P. (2011). The Opacity of Mind: An Integrative Theory of Self-Knowledge. New York, Oxford University Press.

Comesana, J. (2010). "Evidentialist reliabilism." Nous **44**(4): 571-600.

De Neys, W. (2006). "Automatic–heuristic and executive–analytic processing during reasoning: Chronometric and dual-task considerations." The Quarterly Journal of Experimental Psychology **59**(6): 1070-1100.

Evans, J. (2007). "Dual-processing accounts of reasoning, judgment, and social cognition." Annual Review of Psychology **59**: 255-278.

Evans, J. S. B. T. (2009). How many dual-process theories do we need? One, two, or many? In Two Minds: Dual Process and Beyond. J. Evans and K. Frankish. Oxford, Oxford University Press**:** 33-54.

Goldman, A. (1967). "A Causal Theory of Knowing." The Journal of Philosophy **64**(12): 357-372.

Goldman, A. (1976). "Discrimination and Perceptual Knowledge." The Journal of Philosophy **73**(20): 771-791.

Goldman, A. (2006). Simulating minds: The philosophy, psychology, and neuroscience of mindreading. New York, Oxford University Press.

Goldman, A. (2010). "Philosophical Naturalism and Intuitional Methodology: the Romanell Lecture 2010." Proceedings and Addresses of the American Philosophical Association **84**(2): 115-150.

Goldman, A. I. (1979). What is Justified Belief? Justification and Knowledge. G. S. Pappas. Dordrecht, D. Riedel**:** 1-23.

Goldman, A. I. (1992). Liaisons: Philosophy meets the cognitive and social sciences. Cambridge, MA, MIT Press.

Goldman, A. I. (1994). "Naturalistic epistemology and reliabilism." Midwest Studies in Philosophy **19**(1): 301-320.

Greco, J. (2003). Why Not Reliabilism? The Epistemology of Keith Lehrer. E. Olsson. Dordrecht, Springer**:** 31-41.

Johnson, M. K., S. Hashtroudi, et al. (1993). "Source monitoring." Psychological bulletin **114**(1): 3-28.

Kahneman, D. and S. Frederick (2005). A model of heuristic judgment. The Cambridge handbook of thinking and reasoning. K. J. Holyoak. Cambridge, Cambridge University Press**:** 267-293.

Kelley, C. and S. Lindsay (1993). "Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions." Journal of Memory and Language **32**: 1-24.

Koriat, A. (2011). "Subjective Confidence in Perceptual Judgments: A Test of the Self-Consistency Model." Journal of Experimental Psychology: General **140**(1): 117-139.

Koriat, A. and S. Adiv (2011). "The construction of attitudinal judgments: Evidence from attitude certainty and response latency." Social Cognition **29**(5): 577-611.

Koriat, A. and M. Goldsmith (1996). "Monitoring and control processes in the strategic regulation of memory accuracy." Psychological review **103**(3): 490-517.

Kornblith, H. (2010). "What reflective endorsement cannot do." Philosophy and Phenomenological Research **80**(1): 1-19.

Kornblith, H. (MS). On Reflection.

Kunda, Z. (1990). "The case for motivated reasoning." Psychological Bulletin **108**(3): 480-498.

Lehrer, K. (1990). Theory of Knowledge. Boulder, Westview Press.
Lerner, J. S. and P. E. Tetlock (1999). "Accounting for the Effects of Accountability." Psychological Bulletin **125**(2): 255-275.
Michaelian, K. (forthcoming). "Metacognition and Endorsement." Mind and Language.
Nagel, J. (2010). "Knowledge ascriptions and the psychological consequences of thinking about error." Philosophical Quarterly **60**(239): 286-306.
Nagel, J. (2012). "Knowledge as a Mental State." Oxford Studies in Epistemology **4**.
Nagel, J. (2012). Mindreading in Gettier cases and skeptical pressure cases. Knowledge Ascription. J. Brown and M. Gerken. Oxford, Oxford University Press.
Nickerson, R. S. (1999). "How we know--and sometimes misjudge--what others know: Imputing one's own knowledge to others." Psychological Bulletin **125**(6): 737-759.
Royzman, E. B., K. W. Cassidy, et al. (2003). "" I know, you know": Epistemic egocentrism in children and adults." Review of General Psychology **7**(1): 38-65.
Shanahan, M. and B. Baars (2005). "Applying global workspace theory to the frame problem." Cognition **98**(2): 157-176.
Sloman, S. (1996). "The empirical case for two systems of reasoning." Psychological Bulletin **119**(1): 3-22.
Stanovich, K. (1999). Who is rational?: Studies of individual differences in reasoning. Mahwah, NJ, Lawrence Erlbaum.
Thompson, V. A. (2009). Dual process theories: A metacognitive perspective. In two minds: Dual processes and beyond. J. Evans and K. Frankish. Oxford, Oxford University Press**:** 171-195.
Unkelbach, C. (2007). "Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth." Journal of Experimental Psychology: Learning, Memory, and Cognition **33**: 219-230.
Williamson, T. (2000). Knowledge and its Limits. New York, Oxford University Press.