Metacognition and Endorsement

KOURKEN MICHAELIAN

Abstract: Real agents rely, when forming their beliefs, on imperfect informational sources (sources which deliver, even under normal conditions of operation, both accurate and inaccurate information). They therefore face the 'endorsement problem': how can beliefs produced by endorsing information received from imperfect sources be formed in an epistemically acceptable manner? Focussing on the case of episodic memory and drawing on empirical work on metamemory, this article argues that metacognition likely plays a crucial role in explaining how agents solve the endorsement problem.

1. Introduction: The Endorsement Problem

One's precise conception of the endorsement problem will vary with one's general epistemology. Given the broadly virtue-reliabilist epistemology assumed here,[1] the problem has two aspects. First: how can beliefs produced by endorsing information received from imperfect sources be formed in a justification-conferring manner?[2] Second: how can such beliefs be formed in an epistemically virtuous manner? On the account of virtue assumed here, solving the virtue aspect of the problem entails solving the justification aspect but not vice versa; I therefore discuss the justification aspect before turning to the virtue aspect. Taking source monitoring as my central example, I argue first, that though metacognition (the monitoring and control of mental processes) is not always necessary for solving the justification aspect of the problem (if the problem for a particular source is mild enough, the

Thanks for comments to Jérôme Dokic, Conor McHugh, Joëlle Proust, the members of the metacognition reading group at the Institut Jean-Nicod, audiences at the Justification Revisited conference at the Université de Genève, the 2010 Aristotelian Society/Mind joint sessions at University College Dublin, the 2010 Société Française de Psychologie at the Université Charles-de-Gaulle Lille 3, and the University of Edinburgh, and two anonymous referees. My understanding of the epistemology of metacognition has benefited greatly from reading unpublished work by Jennifer Nagel and Chris Lepock. The preparation of this article was supported in part by the Agence Nationale de la Recherche, under the contract ANR-08.BLAN-0205-01.

Address for correspondence: Department of Philosophy, Bilkent University, Ankara 06800, Turkey.
Email: kmichaelian@bilkent.edu.tr

[1] In contrast, e.g., to Sosa's virtue epistemology (Sosa, 2007), this theory (developed in Michaelian, 2011c) is not a standard virtue-reliabilism, since it permits non-virtuous justified belief/knowledge.

[2] As I use the term here, justification can be external.

Mind & Language, Vol. 27, No. 3 June 2012, pp. 284–307.

agent can 'solve' it trivially), it can enable agents to form justified beliefs where this would otherwise be impossible, and, second, that due to the more demanding requirements for virtuous belief formation, it is likely that metacognition is necessary for compensating, in a virtuous manner, for dependence on unreliable sources.

The informational sources on which we depend (memory, inference, etc.) are invariably imperfect; the endorsement problem is thus of central epistemological importance. Nevertheless, it has rarely (with the exception of the epistemology of testimony literature [Gelfert, 2009; Michaelian, 2008, 2010]) been explicitly discussed by epistemologists;[3] this is presumably due to a combination of inattention to the internal structure of belief-producing processes (see Section 2.2) and unawareness of the psychology of metacognition. For example, Goldberg and Henderson, 2006 is one of few recent papers to discuss the role of endorsement in memory in any detail; but even there, Goldberg and Henderson view monitoring exclusively in terms of coherence checking, ignoring work on metamemory.[4]

Before turning to the justification aspect of the endorsement problem, I present two clarifications on informational sources. First: as I use it here, the notion of a source is intended to be flexible enough to apply at different levels, so that while we may, as is standard in epistemology, speak in general terms of perception, memory, inference, and testimony as sources,[5] we may also focus more narrowly on, e.g., conditional inference or (as I do here) episodic memory. While the beliefs produced by a source belong to the agent, the operation of the source need not be under her control: many or most beliefs are produced by automatic, unconscious processes, so that, as far as the agent is concerned, she simply finds herself already believing something; cases in which the agent controls a source (deliberately reasoning to a conclusion, actively searching her memory, etc.) are relatively rare.

Second: while I have characterized imperfect sources as sources that sometimes produce inaccurate representations (analogous to type 2 error), obviously a source might also be imperfect in the sense that it sometimes fails to produce accurate representations that it should produce. This second type of imperfection (analogous

[3] It is anticipated in the search for markers to distinguish memory from imagination (Bernecker, 2008; Byrne, 2010), though this literature does not always differentiate between the problem of distinguishing memory from imagination and that of distinguishing memories stemming from experience from memories stemming from imagination.

[4] This makes their description of the role of endorsement in memory psychologically implausible, as the computational difficulty of checking the coherence of any large set of beliefs means that coherence checking is ill-suited to play the central role here. Sperber has emphasized this point with respect to testimonial beliefs (Sperber, 2001; Sperber *et al.*, 2010), but the point applies equally to beliefs derived from internal sources. Note that, while Sperber takes it that internal sources are reliable enough to make coherence checking superfluous, he does not argue for this point; I take issue with the claim in the specific case of memory below. Note also that, while coherence checking might be too expensive to be worthwhile with respect to a relatively reliable source, a cheaper form of monitoring might still be worthwhile.

[5] Introspection is often included on this list (Michaelian, 2009), which suggests that we should ask whether metacognition itself counts as a source of beliefs; I return to this question in section 2.

to type 1 error) gives rise to a distinct problem, the problem of selecting and directing one's informational sources so that they provide one with needed representations. The problems are interrelated, and I suspect that metacognition is also crucial to explaining how agents solve the 'selection problem', but I will not argue for this here.[6]

2. The Justification Aspect of the Problem

2.1 Reliability

I assume process reliabilism (Goldman, 1979), according to which a belief's degree of justification is determined by the reliability of the process which produces it, where reliability is defined as the tendency to produce a given ratio of true beliefs to total (true plus false) beliefs. There is no clear threshold above which we should say that the beliefs produced by a process are simply justified, but it is presumably the case that unless reliability is significantly better than .5, a level of justification sufficient for knowledge has not been reached; when no confusion will result, I will refer to beliefs that are clearly sufficiently justified for knowledge as being simply justified.

My argument focuses on the reliability of systems (e.g. the episodic memory system). Goldman (1979) argues that the appropriate focus for reliabilism is rather processes, since reliabilism is a theory of justified belief and since a belief can be produced by a process running through multiple systems, each with a distinct level of reliability. This point is compatible with a focus on systems: the justificatory status of a belief is determined by the reliability of the relevant process; a reliable system is one the processes of which are reliable, thus allowing it, when operating on its own, to produce justified beliefs or, when operating in conjunction with other reliable systems, to contribute to the production of justified beliefs.[7] And once we move from the analysis of justification to questions about how agents actually acquire justified/unjustified beliefs, a focus on systems is desirable, for it is the operation of an agent's systems that explains why her beliefs are reliably or

[6] I focus here on retrospective evaluation of object-level processes ('post-evaluation', in Proust's terms), but metacognitive monitoring also includes prospective evaluation ('self-probing'), prediction of 'whether one has the cognitive resources needed for the success of some specific mental task at hand' (Proust, 2008, p. 241); my hunch is that, just as retrospective evaluation helps to explain how agents solve the endorsement problem, prospective evaluation will help to explain how they solve the selection problem. In the case of memory, e.g., an agent's feeling of knowing (Koriat, 1998, 2000) might enable her to determine whether continued efforts to retrieve a given item from memory are likely to be successful, or whether she should rather abandon the attempt (perhaps to consult another source of information) (de Sousa, 2008; Dokic, 2012).

[7] For the sake of simplicity, I focus on beliefs produced by systems operating independently; the implications of the argument for cases in which beliefs are produced by multiple systems operating in conjunction are straightforward.

unreliably formed. Of course, systems are defined in part by the processes that they employ; but a focus on the architecture of systems enables us to move beyond brute claims about the reliability of belief-producing processes to explanations of why the production of a given belief is (un)reliable.

2.2 Two-level Belief-producing Systems

Reliabilists have not always paid attention to the internal structure of belief-producing processes, failing to consider the separate contributions of the various stages of processing in a system to the reliability of the total process by which it produces beliefs. While it is often appropriate to view belief production at this level of abstraction, it is, when the aim is to give more than a shallow explanation of the justificatory status of beliefs, useful to conceive of many belief-producing processes as having a 'two-level' structure. In a two-level system, a first process produces information 'intended' to serve as the content of a belief (in the sense that, in the absence of intervention, the agent will tend to endorse the information), while a second process 'chooses' (in the sense of regulating or governing) between endorsing and rejecting the produced information; if the information is endorsed, a belief having the information as its content is produced, while, if it is rejected, no belief is formed on the basis of the produced information. In a two-level process, the second process functions as a filter or screen on the first.

I will refer to the part of a system in which the information-producing process occurs as its 'information producer' and to the part which determines endorsement/rejection as its 'endorsement mechanism'.[8] The endorsement mechanism can be viewed as implementing an 'endorsement policy' (which it need not represent) consisting of a set of criteria for evaluating produced information together with a rule determining whether a given item of information is to be evaluated as accurate given the extent to which it satisfies these criteria (this is analogous to the variable/criterion-setting distinction in signal detection theory). The reliability of belief production in a two-level system is determined by the interaction of its information producer and its endorsement mechanism, and the operation of the two components can vary independently.

2.2.1 Reliability in Two-level Systems. While the reliability of the information producer (the ratio of accurate representations to total representations) affects the reliability of the total system, the reliability as such of the endorsement mechanism (the ratio of accurate evaluations of representations as accurate/inaccurate to total evaluations of representations as accurate/inaccurate) does not. Since the production of a representation results in the production of a belief only when the representation is evaluated as accurate, it is only the reliability of the endorsement

[8] It is possible that multiple mechanisms cooperate to determine endorsement/rejection; for the sake of simplicity, I will generally ignore this possibility here.

mechanism with respect to evaluations of representations as accurate (the ratio of accurate evaluations of representations as accurate to total evaluations of representations as accurate) that affects the reliability of the total system. In other words, the reliability of the total system is determined by, first, the frequency with which the information producer produces accurate representations and, second, the frequency with which the endorsement mechanism accurately evaluates produced representations as accurate. In what follows, references to the reliability of an endorsement mechanism are to its reliability with respect to evaluations of representations as accurate.

One way of ensuring that a belief-producing system is highly reliable is to equip it with an information producer that is highly reliable, that is, to ensure that the base rate of accurate representations received by the system's endorsement mechanism is high. In a system with a highly reliable information producer, an endorsement mechanism is redundant as far as justification is concerned:[9] if there are few inaccurate representations to filter out, then no additional activity is required to ensure reliable belief production—the endorsement mechanism can simply employ a policy of automatic endorsement. But real agents do not in general have near-perfect informational sources, and thus there is a potential role for endorsement mechanisms to play in ensuring reliable belief production. Thus, assuming that we are dealing with two-level systems, the justification aspect of the endorsement problem can be understood in terms of the interaction between endorsement mechanisms and information producers: how can an endorsement mechanism achieve sufficient reliability to enable a system which includes an unreliable information producer to attain a level of reliability high enough for justification?

2.3 Metacognitive Belief-producing Systems

My suggestion is that metacognition plays a key role here. In any metacognitive system, we can distinguish between an object-level and a meta-level containing a model of the object-level, on the basis of which it intervenes to affect the object-level. The two levels are connected by relations of monitoring and control: in monitoring, information flows from the object-level to the meta-level, shaping the latter's model of the object-level, while, in control, information flows from the meta-level to the object-level, changing the state of the object-level.[10]

Control can be a matter of initiating a new process, continuing a process, or terminating a process. Since the endorsement problem concerns the response of

[9] Though, as I argue in Section 3, it might still be necessary for virtue.

[10] This is the classical conception of metacognition developed by Nelson and Narens (1990, 1994). Some researchers use the term more broadly, to refer to any thinking about thinking. For overviews of metacognition research, see Koriat, 2002 (psychology), Cox, 2005 (computer science), Proust, 2010; Arango-Muñoz, 2011 (philosophy).

agents to already-produced information, I focus here on control operations of the latter two types.

Monitoring includes, in addition to source monitoring processes, processes producing ease of learning judgements, judgements of learning, feelings of knowing, retrospective confidence judgements, and so on. While it is natural to think of control as being based on monitoring (e.g. an agent decides to continue her attempt to retrieve the answer to a question based on her feeling of knowing) Koriat and his colleagues (Koriat, Ma'ayan and Nussinson, 2006; Koriat and Ackerman, 2010) have argued for the existence of control-based monitoring, in which monitoring is effected by means of feedback from control operations (e.g. the feeling of knowing is determined by the amount and ease of access of partial information in failed retrieval attempts [Koriat, 1993]); they do not, however, dispute the importance of monitoring-based control, arguing rather that control-based monitoring and monitoring-based control are complementary processes. Since my central example here is source monitoring, I focus on monitoring-based control.

Metacognition can draw both on automatic (heuristic, unconscious, fast—'system 1') and on systematic (reflective, conscious, slow—'system 2') processing; that is, control operations can be based either on automatic or on systematic monitoring (Stanovich, 1999; Evans, 2008; Frankish, 2010). Even in systems capable of both types of processing, (cheap) automatic processing executed by the relevant system is the default. But (expensive) systematic processing executed at the level of the agent can also occur under certain circumstances: the agent can deliberately initiate systematic processing, or systematic processing might be triggered by the system itself under certain conditions.

A two-level system need not be a metacognitive system: though any endorsement mechanism by definition controls an object-level information-producing process, the mechanism need not monitor that process—endorsement might be determined in some other way (e.g. on the basis of information about environmental conditions). But if endorsement is determined on the basis of the mechanism's monitoring of the information producer, the result is a metacognitive system (see Figure 1): the information producer produces representations; the endorsement mechanism, on the basis of its monitoring of the information producer, either intervenes to prevent a produced representation from being accepted (preventing belief-formation) or permits the acceptance of the representation (permitting belief-formation).[11]

Three points about this structure: First: if suspension of judgement is a propositional attitude on a par with belief/disbelief, the endorsement mechanism should be seen as intervening to either permit formation of a belief (which remains the default) or instead require formation of the attitude of suspended judgement (placing the

[11] There are similarities between my description of metacognitive belief-producing systems and Koriat and Goldsmith's model (Koriat and Goldsmith, 1996).

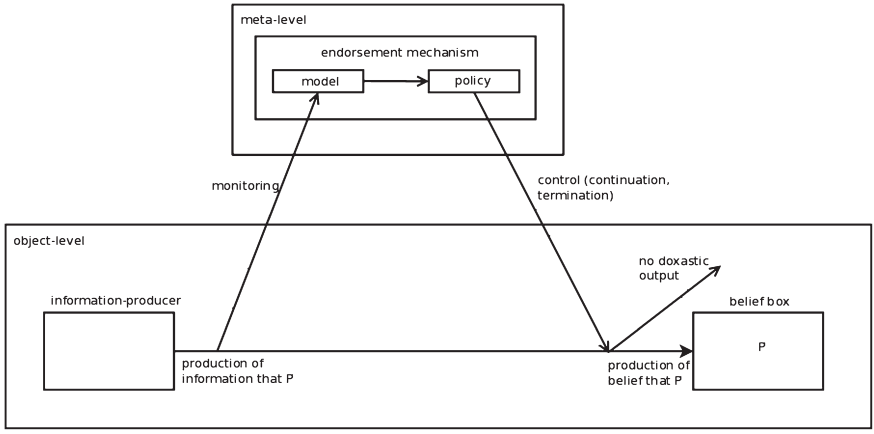

Figure 1 *Structure of a metacognitive belief-producing system.*

produced information in a 'suspended judgement' box rather than discarding it). For the purposes of this article, it does not matter how we conceive of suspension of judgement, since this does not change the effect of the endorsement mechanism on reliability and power: reliability is still the ratio of true beliefs to total beliefs; and power is still the ratio of true beliefs to total beliefs plus failures to form a belief (including suspensions of judgement).

Second: while Figure 1 leaves it open whether belief-production might sometimes bypass the endorsement mechanism, if system 1 monitoring is the default, this suggests that, in practice, the endorsement mechanism is 'always on' in metacognitive systems. Since system 2 monitoring requires more resources, it will be engaged selectively.

Third: Figure 1 does not give a complete picture of the epistemic role of metacognition but only describes its role in solving the endorsement problem. As noted in Section 1, metacognition also plays a role in solving the selection problem. These two roles interact in complex ways: an evaluation of a produced representation as inaccurate can, if self-probing indicates that another process can produce an accurate representation, trigger the production of a representation by another process; e.g. a feeling of disfluency might trigger a shift from system 1 to system 2 processing (Alter *et al.*, 2007; Oppenheimer, 2008; Thompson, 2009, 2010).

2.3.1 Reliability in Metacognitive Systems. Focussing on metamemory suggests looking to communication research for a means of clarifying the epistemic role of metacognition, for memory and testimony are in many ways analogous: though the analogy has its limits, it is sometimes useful to think of retrieved memories as being like testimony received from one's past self; this suggests, in particular, that it might be useful to think of inaccurate memories as being like dishonest

testimony.[12] And the psychology of testimony—specifically, the psychology of deception detection (Michaelian, 2010; Vrij, 2008; Shieber, 2011)—does indeed provide us with a model, the Park-Levine probability model of deception detection accuracy (Park and Levine, 2001; Levine *et al.*, 2006), which can be generalized to provide an account of the reliability of two-level belief-producing processes in general.

Deception detection researchers are interested primarily in deception detection accuracy, the ratio of true judgements by a subject that a speaker is honest/dishonest to total judgements that a speaker is honest/dishonest. Where H abbreviates 'the agent judges that the speaker is honest' and T abbreviates 'the speaker is honest', deception detection accuracy is determined by summing $P(H\&T)$ and $P(\sim H\&\sim T)$, which can be calculated as follows.

$$P(H\&T) = P(H|T) \times P(T) \tag{1}$$

$$P(\sim H\&\sim T) = P(\sim H|\sim T) \times (1 - P(T)) \tag{2}$$

Epistemologists, on the other hand, since they are concerned with the reliability of forming beliefs by accepting received testimony, and since a testimonial belief is formed only when the recipient judges the speaker to be honest (Michaelian, 2010; Fricker, 1995), are interested primarily in 'honesty accuracy' (R), the ratio of true judgements that the speaker is honest to total judgements that the speaker is honest:

$$R = P(H\&T)/(P(H\&T) + P(H\&\sim T)) \tag{3}$$

For which we need:

$$\begin{aligned} P(H\&\sim T) &= P(H|\sim T) \times (1 - P(T)) \\ &= (1 - P(\sim H|\sim T)) \times (1 - P(T)) \end{aligned} \tag{4}$$

[12] One might worry that memory (like other internal sources) is in fact strongly disanalogous to testimony, since, while an agent's internal sources are not built to deceive him, other agents will often have an interest to deceive. This sort of consideration, in fact, leads Sperber and his collaborators to focus their discussion of 'epistemic vigilance' entirely on testimonial information, neglecting vigilance with respect to internal sources (Sperber, 2001; Sperber, *et al.*, 2010). But they do not provide an argument for the assumption that internal sources are sufficiently reliable to render monitoring unnecessary; below, I review work on memory which suggests that, due to the reconstructive character of retrieval and to memory's storage of information deriving from a variety of internal and external sources, remembering (absent effective monitoring) is not highly reliable. Moreover, even if a given system is highly reliable, vigilance might still be appropriate if it can be accomplished cheaply; e.g. perception is presumably more reliable than memory, but work on metaperception suggests that there is a role for vigilance here as well (Levin, 2002; Loussouarn, 2010).

The modified Park-Levine model makes clear that honesty accuracy varies as a function of both the base rate of honesty ($P(T)$) and the agent's sensitivity to honesty/dishonesty—the conditional probabilities of judging that a speaker is honest given that she is honest ($P(H|T)$) and that she is dishonest given that she is dishonest ($P(\sim H|\sim T)$).

The accuracy of testimony is determined not only by the speaker's honesty but by her competence as well (Fricker, 1995; Mascaro and Sperber, 2009), but if we abstract away from competence, then equation (3) gives us the reliability of the process used to form testimonial beliefs. Since any two-level process shares the same basic structure (production of a representation, followed by endorsement/rejection of the representation according to whether it is evaluated as accurate/inaccurate), the same equation can be used to describe the effects of any endorsement mechanism on the reliability of belief production by the relevant system: if T abbreviates 'the information produced is accurate' and H abbreviates 'the information is evaluated as accurate', then equation (3) gives us the reliability of the two-level belief-producing system. As before, reliability (R) is determined by the base rate of accurate representations ($P(T)$) and sensitivity to accuracy/inaccuracy ($P(H|T)$ and $P(\sim H|\sim T)$): the better the information producer, the higher $P(T)$; the more sensitive the endorsement mechanism, the higher $P(H|T)$ and $P(\sim H|\sim T)$.

Given a perfect information producer, there is no role for an endorsement mechanism to play: if $P(T) = 1$, then $R = 1$, whatever the specific endorsement policy employed by the mechanism—at $P(T) = 1$, all policies are equivalent to the policy of automatic endorsement. But as $P(T)$ decreases, the potential role of an endorsement mechanism becomes increasingly important for ensuring a high R.

As Figure 2 illustrates, there is no guarantee that an endorsement mechanism will make a significant contribution to the reliability of belief production; whether the mechanism makes a difference depends on its sensitivity to the (in)accuracy of the informational source. The .75/.35 curve represents the effects of an endorsement mechanism employing a policy such that for it $P(H|T)$ and $P(\sim H|\sim T)$ are .75 and .35, respectively; such a policy is 'truth-biased' in a manner analogous to the policies actually used by agents in the formation of testimonial beliefs (Levine, Park, and McCornack 1999), and the reliability of such a system is nearly equivalent to one employing the policy of automatic endorsement, very nearly mirroring the base rate of accurate information.

The .8/.8 curve, on the other hand, represents the effect of an endorsement mechanism such that for it $P(H|T)$ and $P(\sim H|\sim T)$ are both .8; this is a mechanism which is reasonably sensitive to both accuracy and inaccuracy on the part of the information producer. Such a mechanism improves the reliability of belief production across the board (at whatever base rate of accurate representations). And it allows the endorsement mechanism to compensate for even an absolutely unreliable information producer; e.g. the system can produce true beliefs more than 70% of the time despite relying on an information producer that gets it right only 40% of the time.

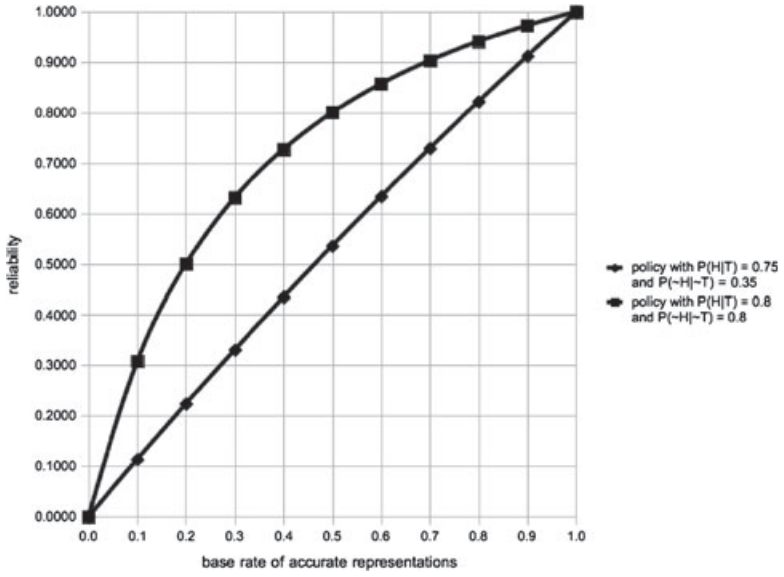

Figure 2 *Effect of different hypothetical endorsement mechanisms on the reliability of belief-producing systems.*

It is unlikely that this sort of improvement in reliability can be achieved by a non-metacognitive mechanism, a mechanism without access to information about the operation of the information producer (though this cannot be ruled out a priori). Given the computational intractability of coherence checking, such a mechanism would in effect have to rely on environmental cues or predict in advance the occasions on which the information producer will make a mistake; it is difficult to see how this might be done.

In contrast, if features of the production of representations (or of the representations themselves) can provide indications of the accuracy of the relevant representations, a metacognitive mechanism can determine on a case-by-case basis whether the information producer has made a mistake. In a metacognitive system, information about the operation of the information producer is passed to the endorsement mechanism and serves as a basis for the decision whether to accept the representation produced on a given occasion; if the endorsement mechanism employs an appropriate policy, it can intervene at the object-level to prevent belief formation when the information producer delivers an inaccurate representation and can refrain from intervening, thus permitting belief formation, when the information producer delivers an accurate representation. Thus a metacognitive mechanism provides a possible solution to the justification aspect of the endorsement problem: a system with an information producer that is sufficiently reliable for justification need not incorporate an endorsement mechanism in order to produce justified beliefs; but in a system with an information producer that is insufficiently reliable

for justification, an endorsement mechanism can in principle enable the attainment of a level of reliability sufficient for justification; this can be accomplished by a metacognitive endorsement mechanism, in particular, as long as it can use the information that it receives about the operation of the information producer to evaluate, with a sufficiently high degree of reliability, the accuracy of the representations that the latter produces.[13]

It might be worried that, if metacognitive endorsement mechanisms are themselves imperfectly reliable, they would need to be monitored by further mechanisms; if these further mechanisms are likewise imperfect, they would need to be monitored by yet further mechanisms; and so on. While metacognitive mechanisms are indeed imperfect,[14] this need not give rise to a regress, for, as we have seen in this section, a metacognitive mechanism need not be perfectly reliable in order to compensate for a significantly unreliable informational source (Figure 2). Thus imperfection in the endorsement mechanism, if it is not too severe, need not give rise to a need for the endorsement mechanism itself to be monitored.[15]

3. The Virtue Aspect of the Problem

The argument of the preceding section is meant to establish that a capacity for metacognition can in principle enable an agent to solve the justification aspect of the endorsement problem. In this section, I argue that, if it can employ both system 1 and system 2 processing, such a capacity can in principle also enable the agent to solve the virtue aspect of the problem.

3.1 Power and Speed

I adopt an account of virtue broadly similar to that associated with virtue-reliabilism (Sosa, 2007), according to which virtues are stable, reliable cognitive faculties or systems. But (in an approach inspired by Lepock, 2011) I modify this account to require properties in addition to reliability for virtue: reliability is only one among

[13] The information producer need not perform a special operation in order to provide an indication of the accuracy of the representation that it produces; the properties which indicate accuracy can rather be byproducts of its normal operation, which can then be interpreted by the endorsement mechanism as indicating accuracy/inaccuracy. For example, retrieval fluency can be treated as an indicator of the accuracy of retrieved information (Benjamin and Bjork, 1996; Hertwig *et al*., 2008).

[14] For example, the validity of fluency as a cue to truth depends on the majority of statements encountered by the agent being true (Reber and Unkelbach, 2010).

[15] Metcalfe (2008) identifies another source of the temptation to suppose that metacognition implies a regress in the tacit assumption that monitoring requires a sophisticated entity to do the monitoring, an entity the cognitive processes of which would then presumably themselves require monitoring—as she points out, the assumption is false, since a metacognitive monitor can be relatively simple.

a number of epistemically important properties of cognitive systems—as Goldman has argued (1992), though these properties are not reflected in the concept of knowledge, power and speed are also epistemically crucial (see also Cummins, Poirier and Roth, 2004). Reliability is a matter of avoiding false beliefs, and thus the reliability of a system is compatible with its not producing very many true beliefs. But a system that produces mostly true beliefs but produces them only rarely will be of little use to the agent. In addition to reliability, then, cognitive virtue requires power, defined as the tendency to produce a given ratio of true beliefs to (true or false) beliefs plus failures to produce a belief. Similarly, a system that is highly reliable and powerful might be slow. But a system that produces true beliefs only when they are no longer required will be of little use to the agent. Cognitive virtue thus also requires an acceptable level of speed.

Though speed tends to contribute to power (since processes in a faster system less often have to be interrupted to divert resources to other tasks), there are often trade-offs between power and reliability, since, as signal detection theory reminds us, additional true beliefs can often be secured at the expense of forming some additional false beliefs. The levels of reliability, power, and speed that are appropriate for a given system are determined in part by its function: in general, a virtuous system is one that performs its function well; appropriate levels are those that enable the system to perform its function well. While the function of the system determines minimum permissible levels, the precise levels that are appropriate at a given time are determined by the agent's context at that time, for context affects the goals of the agent, the consequences of forming beliefs (or failing to form beliefs) on certain topics, and so on: e.g. in a context in which it is crucial to avoid forming false beliefs, a higher level of reliability will be appropriate; when the consequences of forming false beliefs are minor, it might be appropriate to settle for a lower level of reliability in order to increase the power of the system. A virtuous system, in short, flexibly adjusts its levels of reliability, power, and speed according to the agent's current context.

3.2 Power and Speed in Metacognitive Systems

In a system with a perfect information producer, there is no need to negotiate trade-offs between reliability and power, since automatically accepting produced representations maximizes both reliability and power. But in a system with an imperfect information producer, reliability and power can come apart. One way for a system to achieve high reliability despite an imperfect information producer is for its endorsement mechanism to employ a highly risk-averse policy, rejecting a representation whenever it judges that there is even a slight chance that it is inaccurate; a system employing such a policy will typically be less powerful than an otherwise similar system employing a more risk-tolerant policy. By the same token, increasing power tends to decrease reliability to some extent, for the cost of accepting some additional accurate representations is normally also accepting some additional inaccurate representations.

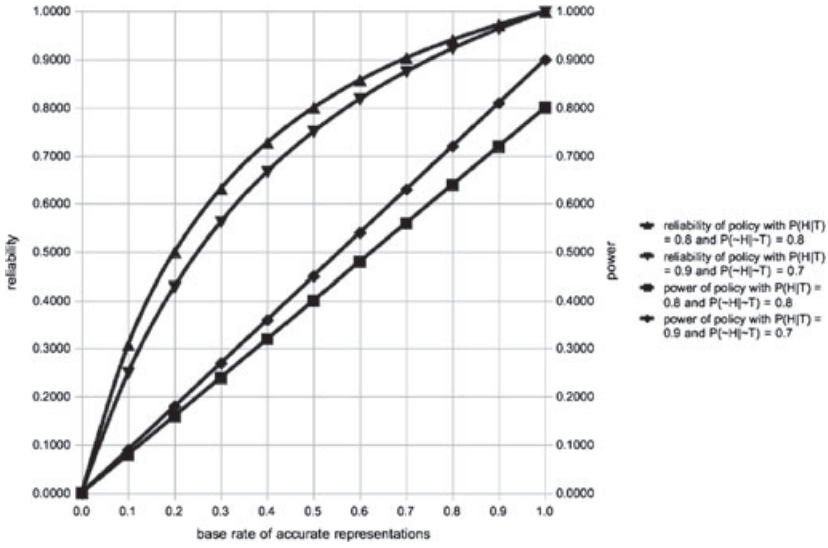

Figure 3 *Effect of more or less risk-averse endorsement policies on power and reliability.*

This trade-off is illustrated in Figure 3. Since a belief is produced only if the endorsement mechanism evaluates a produced representation as accurate, the power of a two-level system is equivalent to $P(H\&T)$. In general, by increasing the 'truth bias' of the endorsement mechanism, it is possible to secure some additional power at the expense of some reliability—adopting a more generous policy increases $P(H|T)$ but tends to decrease $P(\sim H|\sim T)$. By the same token, by decreasing the truth bias of the mechanism, it is possible to secure some additional reliability at the expense of some power.

If the conditional probabilities can vary independently, there need be no trade-off between reliability and power. But it is unlikely that feasible endorsement mechanisms are such that they allow the probabilities to vary independently: an endorsement mechanism that employs a policy other than that of automatic acceptance will tend to decrease the power of the system, for the cost of filtering out some inaccurate representations is filtering out also some accurate representations; similarly, the cost of letting through more accurate representations is normally failing to filter out some inaccurate representations. This gives rise to the virtue aspect of the endorsement problem: given that it is not possible to improve reliability without decreasing power, and vice versa, and given that in different contexts a virtuous agent will prioritize either reliability over power or power over reliability, virtue requires an ability to negotiate trade-offs between reliability and power according to the demands of the current context. To see how this trade-off is negotiated, it is also necessary to take the requirement for speed into account.

In a metacognitive system, endorsement can be determined either by automatic or by systematic monitoring. The policy employed in automatic monitoring is

fixed and relatively simple; this ensures that it is computationally inexpensive and therefore fast. Given that automatic monitoring will be faster than systematic monitoring, a virtuous system will employ automatic monitoring by default, shifting to systematic monitoring sparingly.

The fixed character of the automatic policy means that it is largely insensitive to changes in the agent's context. A well-designed endorsement mechanism will include a policy that is suitable for typical contexts, in the sense that it ensures an appropriate balance of reliability and power in those contexts. But a fixed policy cannot adjust those levels, as the agent moves into atypical contexts, in order to ensure that reliability or power is prioritized, as appropriate. Systematic processing executed at the level of the agent, on the other hand, can allow the temporary adoption of a policy better-suited to an atypical context: by altering the criteria relevant to endorsement or the thresholds that need to be crossed on certain criteria, the agent can adopt more or less risk-averse policies according to the requirements of the current context.[16] In a system capable only of automatic processing, in contrast, reliability and power are fixed by the base rate of accurate representations, so that the balance between them cannot be altered when the agent moves into a new context.

If the metacognitive component of a system contains a set of expectations about the operation of its information producer and a means of comparing the observed operation of the producer to its expected operation, the switch to systematic processing can be initiated by the system itself; this sort of automatic comparison will be useful mainly for indirectly detecting atypical environments and failures of the information producer to operate normally in typical environments. The switch can also be effected by the agent herself, either in response to such discrepancies (if she has expectations about the operation of the information producer) or in response to changes in her interests (Koriat and Goldsmith, 1996).[17]

One might object that it is implausible that metacognition provides the sort of increased flexibility required to solve the virtue aspect of the endorsement problem, for, if it did so, then we should expect it to be much more widespread

[16] A similar but less flexible negotiation of the trade-off between reliability and power can be achieved by employing a much more complicated automatic endorsement policy, but this will involve a significant sacrifice in the speed of automatic processing.

[17] This means that the contribution of metacognition to the achievement of virtue by a system depends on features of the agent beyond the system in question, but the agent herself need not be virtuous in any robust sense but only have enough knowledge of the workings of her own informational sources to enable her temporarily to adopt endorsement policies that ensure appropriate levels of reliability/power in atypical contexts. A related point has been emphasized by Anderson and his colleagues in their work on the 'metacognitive loop' (Anderson and Perlis, 2005; Anderson *et al.*, 2006; Schmill *et al.*, 2008): they argue that a metacognitive 'note-assess-guide' procedure allows an agent to overcome the problem of 'cognitive brittleness', that is, that it allows her to respond appropriately when she encounters circumstances in which her usual behaviour does not produce the expected results, rather than uselessly persisting in that behaviour.

across species.[18] But while we should indeed expect animal belief-forming systems, metacognitive or not, to be reliable, powerful, and fast, since in general true belief is adaptive and false belief is maladaptive (McKay and Dennett, 2009), a system need not have the sort of flexibility required for virtue in order to be adaptive. It is thus not surprising that metacognition appears as an evolutionarily later add-on: as Proust points out, while there is in general selective pressure, due to environmental variability, for behavioural flexibility, metacognition emerges rather because it provides flexibility specifically in cognition (2006).

Though this answers the worry about flexibility, it might seem, given what I have said about the role of metacognition in solving the justification aspect of the endorsement problem, that we should still expect metacognition to be more widespread: if metacognition is required to compensate for unreliable information producers, then it should be present in many species. Reliability is indeed necessary for a system to be adaptive, but I have not claimed that non-metacognitive systems cannot be reliable. Metacognition becomes necessary for reliability only when the information producer becomes significantly unreliable, which tends to happen with significantly increased cognitive sophistication. The emergence of a sophisticated capacity for inference provides one example. Episodic memory, to which I know turn, provides another: given the constructive character of encoding and retrieval, and given storage of information originating in a variety of sources, the episodic system probably could not have evolved if it did not include a capacity for metacognition.

4. Source Monitoring and the Endorsement Problem for Episodic Memory

The argument of Sections 2 and 3 is meant to establish that it is possible in principle that a capacity for metacognition can enable an agent to solve the endorsement problem. In this section, I review work on source monitoring, illustrating how metamemory actually solves the endorsement problem to which the human episodic memory system gives rise.[19]

---

[18] While the question of animal metacognition remains controversial, it is clear that metacognition is evolutionarily very recent, present in humans and perhaps a few other species (Metcalfe 2008; Smith, Shields, and Washburn, 2003; Carruthers, 2008; Proust, 2006).

[19] My focus on metamemory here should not be taken to imply that the endorsement problem arises with respect to memory only: the problem arises for any imperfect informational source, and it will be an interesting problem for any significantly imperfect source. Among the traditionally recognized basic epistemic sources, the endorsement problem for inference is particularly complex, since, when an inference produces a conclusion, the agent must always choose between accepting the conclusion or, instead, rejecting at least one of the premises from which she inferred it (Lepock, 2011; Morton, 2004). Metacognition appears to play a role here as well, with feelings of rightness or confidence affecting the agent's decision to

4.1 The Need for Source Monitoring

Epistemologists tend to view remembering as being a process which takes a belief as input, stores it, and later delivers the same belief as output, but this natural view is mistaken. Memory rather has the sort of two-level structure described in Section 2.2: retrieval from memory first produces a representation, and a belief is formed only when the representation is endorsed—in many cases, the representation is rejected, so that no belief is formed. The source monitoring framework focuses on the decision process that determines endorsement/rejection of retrieved representations.

The framework was developed in part in response to the recognition that memory is thoroughly constructive: encoding of a memory is not a matter of the storage of a copy of a representation stemming from experience but rather involves processes of selection, abstraction, interpretation, and integration of information from various sources; stored memories are susceptible to modification during the period of reconsolidation that follows retrieval; and the representation produced by retrieval is reconstructed not only from stored information but also incorporates information available in the context of retrieval (Tulving, 1982; Schacter and Addis, 2007; Loftus, 2005; Alberini, 2005; Sutton, 2010; Robin, 2010; Lackey, 2005; Michaelian, 2011a, 2011b; Matthen, 2010; Vosgerau, 2010; Shanton, 2011). The constructive character of memory raises the question of how the memory system can 'remain functional and not deteriorate into a pathological quagmire of real and imagined experiences or recombinations of features of real experience' (Mitchell and Johnson, 2000, p. 180). Source monitoring theorists argue that we are able to discriminate the origins of mental experiences by means of attributional judgements processes, evaluative or monitoring processes which take us from phenomenal properties of retrieved information (and, in certain cases, features of its relation to other memories) to a judgement that the information stems from a certain source (and thus is or is not likely to be veridical) (Mitchell and Johnson, 2000, p. 180).

Though the development of the source monitoring framework was motivated in part by the need to explain how memory achieves reliability despite its constructive character, the framework can be motivated even without appealing to constructive memory. As Johnson and her colleagues themselves point out in various places (e.g. (Johnson, Hashtroudi and Lindsay, 1993, p. 3), (Johnson and Raye, 2000, p. 38)), as long as memory stores, along with records originating in reliable sources, a significant proportion of records originating in unreliable sources, and as long as memory does not typically store information about the sources of records (Johnson, 2006), agents face the problem that source monitoring is designed to solve: having retrieved a record, the agent must somehow determine which source it stems from, so that she knows what attitude to adopt towards it.

---

accept the inferred conclusion (Thompson, 2009, 2010; De Neys and Franssens, 2009; De Neys, Cromheeke and Osman, 2011).

4.2 Source Monitoring and Justification

The severity of the justification aspect of the endorsement problem for episodic memory depends on the reliability of the inferences involved in construction/reconstruction and on the reliability of the various sources of information stored by the system (and the the proportion of records originating in each of those sources). While I have argued elsewhere (Michaelian, 2011a) that construction/reconstruction are reliable, they are nevertheless imperfectly reliable, thus decreasing the base rate of accurate representations produced by retrieval. Moreover, it is likely that the episodic memory system stores a high proportion of records originating primarily in unreliable sources, for the system stores not only records originating in experience but also records originating a variety of other sources—dreaming, fantasizing, imagining, intending, planning, experiences of works of fiction, testimony received from untrustworthy interlocutors, etc. Thus, though we cannot determine the base rate of accurate representations produced by episodic retrieval with any real precision, it is likely not sufficiently high for justification.

According to the source monitoring framework, though memory does not normally store information about source, memories typically bear characteristic marks of the sources in which they originate:

> Different types of acquisition processes (e.g., reading, thinking, inferring) and different types of events (e.g., movie, newspaper, dream) tend to produce memorial representations that are characteristically different from each other. For example, memories of imagined events typically have less vivid perceptual, temporal, and spatial information than perceived events and often include information about intentional cognitive operations (e.g., active generation and manipulation of visual images during problem solving). Memories of dreams are often perceptually vivid, typically do not include information about the cognitive operations that created them, and are often inconsistent with knowledge or other memories (Mitchell and Johnson 2000, p. 180)

The presence of these marks means that it is possible to determine the source of a record with some reliability:

> . . . the source of information typically is not something stored as propositional tag along with our memories, beliefs, and knowledge. Rather, we infer source. We use heuristic source monitoring processes to attribute a source to information based on an evaluation of various features of the information. If activated information from the memory being evaluated has qualities that we expect memories from a certain source to have, we attribute the information to that source (Johnson and Raye, 2000, p. 39).

The source monitoring framework suggests that episodic memory solves the endorsement problem (if it does) because retrieved records are endorsed only if they are evaluated as having originated in reliable sources (so that they are likely to be true), and because these evaluations are themselves reliable. Unfortunately,

we cannot be very precise about the reliability of evaluations of source. The source monitoring literature tends to suggest that our evaluations are fairly reliable; if this is right, then source monitoring likely enables an adequate solution to the justification aspect of the endorsement problem.[20]

Since it is difficult or impossible to establish the base rate of accurate representations produced by retrieval with much precision, and since it is also difficult or impossible to establish the reliability with which source monitoring discriminates accurate from inaccurate representations with much precision, I grant that we cannot firmly establish that source monitoring successfully solves the endorsement problem for episodic memory. These are empirical difficulties: it is simply not clear that these questions are subject to direct empirical investigation. Thus any discussion of the epistemic impact of source monitoring will necessarily have a somewhat speculative character.

That the argument is speculative does not, however, mean that it provides no evidence for its conclusion. We know, from countless investigations of various aspects of memory (the work of Loftus and collaborators on suggestibility, to take but one prominent example [Loftus, 1979/1996, 2005]), that the episodic system stores information from a variety of sources; this strongly suggests that the base rate of accurate representations produced by episodic retrieval is too low for justification. We can assume, on evolutionary grounds, that episodic memory belief-formation is highly reliable (and hence reliable enough for justification); in light of the low base rate of accurate representations produced by retrieval, the assumed reliability of memory requires an explanation. The assumption that source monitoring is sufficiently reliable to raise the reliability of memory belief formation to a high level is a plausible candidate explanation. This explanation, moreover, receives additional support to the extent that the source monitoring framework itself is well-supported: the framework incorporates the assumption (though source monitoring researchers put it in slightly different terms) that monitoring is sufficiently reliable to compensate for unreliable retrieval; and the framework overall itself has considerable empirical support.[21] Thus we may conclude (tentatively) that source monitoring solves the justification aspect of the endorsement problem for episodic memory.

4.3 Source Monitoring and Virtue

To ensure sufficient speed (and hence power), a metacognitive belief-producing system should by default employ automatic monitoring, since this is in general faster

[20] Evaluations based on phenomenal charactersitics of retrieved information are bound to be imperfectly reliable—as Mitchell and Johnson note, 'because of variability within ... source types, the distributions of features of memories from different processes and events overlap' (2000, p. 180). In consequence, e.g., agents with vivid mental imagery tend to do worse at distinguishing imagined from perceived information (Johnson, 1997, p. 1735).

[21] I cannot provide a serious review of the evidence for the framework here. See, e.g., Johnson, Hashtroudi, and Lindsay 1993; Johnson and Raye 2000; Horton, Conway, and Cohen 2008 for reviews of some of the evidence.

than systematic monitoring. As Johnson and Raye point out, source monitoring does indeed employ heuristic processing by default: 'We are not always conscious of these processes. Heuristic source attributions take place constantly without notice, and are relatively automatic or effortless' (2000, p. 39). Source monitoring researchers are not always clear about whether heuristic monitoring is executed by the memory system itself or whether it is rather executed by the agent; but given the automaticity of the processing in question (the fact that it is triggered without intervention by the agent and normally concludes without the agent having become aware that it is occurring), it seems likely that it is executed by the memory system itself. But systematic source monitoring is also engaged under certain circumstances: the source monitoring framework 'posits that source monitoring sometimes also entails more systematic processes that are typically slower and more deliberate, involving, for example, retrieving additional information, discovering and noting relations, extended reasoning, and so on' (Mitchell and Johnson, 2000, pp. 180–81).[22]

Given that agents sometimes engage in systematic source monitoring, when do they do so? In particular, do they tend to engage in systematic monitoring when the current context calls for a non-default balance of reliability and power? The literature suggests that the need for a higher-than-usual level of reliability is normally responsible for the initiation of systematic monitoring (Mitchell and Johnson, 2000, p. 181). Discussions of cases in which reliability is sacrificed for additional power are less frequent, but the framework clearly allows for this—there is nothing in the framework that prevents agents from temporarily adopting more relaxed source monitoring criteria. But it is important to note that our capacity to initiate systematic source monitoring when it is called for, and to adjust our policy so as to achieve appropriate levels of reliability and power, depends both on our appreciation of our own current context and on our metacognitive knowledge of the workings of our own memory systems (Johnson and Raye, 2000, p. 39) (Johnson, 1997, p. 1734).[23] Given that systematic source monitoring is sensitive to the agent's current context, and assuming that automatic monitoring achieves appropriate default levels of reliability and power, it seems likely that source monitoring enables agents to solve the virtue aspect of the endorsement problem, though whether it does will of course depend on the specific manner in which the endorsement policy is adjusted in systematic monitoring.

---

[22] See Johnson and Raye, 2000, p. 48 for more detail on the processes employed in systematic source monitoring.

[23] Source monitoring theorists sometimes suggest that automatic (and not only systematic) source monitoring is sensitive to the requirements of context. The idea might be that the policy used in automatic monitoring is not fixed, in the sense that it can treat properties of retrieved records differently according to context, which would require that information about context (including the agent's current goals, etc.) can be passed to the monitoring mechanism. But there must be some limit on the flexibility of automatic monitoring; otherwise, the distinction between it and systematic monitoring disappears.

5. Conclusion: Externalism and Endorsement

While my approach to the endorsement problem is thoroughly externalist (I have neither attempted to determine how internalists should conceive of the problem nor investigated whether the metacognitive solution to the problem satisfies internalist requirements), there is an obvious move that might permit an internalist appropriation of at least the part of my argument concerned with the justification aspect of the problem: the claim would be that my discussion of the metacognitive solution suggests that agents are able to form internalistically justified beliefs when endorsing information received from internal sources because the relevant process is not only reliable but involves certain internal justifiers, namely evaluations, produced by monitoring, of the probable truth of the received information.

This sort of view might in principle be workable, and it is indeed consistent with my account, but, for two reasons, it is not actually supported by my argument. First: the view requires that monitoring produces explicit evaluations of received information, but, while I have written for convenience as if this is the case, nothing in my argument requires that it is; the argument requires only that control is sensitive to certain features of received information, and this sensitivity need not be mediated by representations—metarepresentation is not a prerequisite for metacognition, or at least the claim that it is requires additional argument (Arango-Muñoz, 2011; Proust, 2007). Second: even if metacognition does require metarepresentation, the representations involved will typically be encapsulated and hence unavailable to other cognitive processes; thus they cannot serve as internally accessible justifiers, justifiers that are accessible to the agent herself. While I have written in terms of agents solving the endorsement problem, in general the problem is solved by the design of cognitive systems that typically operate at a sub-personal level.

Department of Philosophy
Bilkent University

References

Alberini, C.M. 2005: Mechanisms of memory stabilization: are consolidation and reconsolidation similar or distinct processes? *Trends in Neurosciences*, 28, 51–6.

Alter, A. L., Oppenheimer, D. M., Epley, N. and Eyre, R.N. 2007: Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, 136, 569–76.

Anderson, M. L., Oates, T., Chong, W. and Perlis, D. 2006: The metacognitive loop I: enhancing reinforcement learning with metacognitive monitoring and control for improved perturbation tolerance. *Journal of Experimental & Theoretical Artificial Intelligence*, 18, 387–411.

Anderson, M. L. and Perlis, D. R. 2005: Logic, self-awareness and self-improvement: the metacognitive loop and the problem of brittleness. *Journal of Logic and Computation*, 15, 21–40.

Arango-Muñoz, S. 2011: Two levels of metacognition. *Philosophia*, 39, 71–82.

Benjamin, A. S. and Bjork, R. A. 1996: Retrieval fluency as a metacognitive index. In L. Reder (ed.), *Implicit Memory and Metacognition*. Mahwah, NJ: Erlbaum.

Bernecker, S. 2008: *The Metaphysics of Memory*. New York: Springer.

Byrne, A. 2010: Recollection, perception, imagination. *Philosophical Studies*, 148, 15–26.

Carruthers, P. 2008: Metacognition in animals: a skeptical look. *Mind & Language*, 23, 58–89.

Cox, M. 2005: Metacognition in computation: a selected research review. *Artificial Intelligence*, 169, 104–41.

Cummins, R., Poirier, P. and Roth, M. 2004: Epistemological strata and the rules of right reason. *Synthese*, 141, 287–331.

de Sousa, R. 2008: Epistemic feelings. In U. Doğuoğlu and D. Kuenzle (eds), *Epistemology and Emotions*. Farnham: Ashgate.

De Neys, W., Cromheeke, S. and Osman, M. 2011: Biased but in doubt: conflict and decision confidence. *PLoS ONE*, 6: e15954+.

De Neys, W. and Franssens, S. 2009: Belief inhibition during thinking: not always winning but at least taking part. *Cognition*, 113: 45–61.

Dokic, J. 2012: Seeds of cognition. Noetic feelings and metacognition. In M. Beran, J. Brandl, J. Perner, and J. Proust (eds.), *Metacognition, Mental Agency and Self-Awareness*. Oxford University Press. Forthcoming.

Evans, J. St.B.T. 2008: Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255–78.

Frankish, K. 2010: Dual-process and dual-system theories of reasoning. *Philosophy Compass*, 5, 914–26.

Fricker, E. 1995: Telling and trusting: reductionism and anti-reductionism in the epistemology of testimony. *Mind*, 104, 393–411.

Gelfert, A. 2009: Indefensible middle ground for local reductionism about testimony. *Ratio*, 22, 170–90.

Goldberg, S. and Henderson, D. 2006: Monitoring and anti-reductionism in the epistemology of testimony. *Philosophy and Phenomenological Research*, 72, 600–17.

Goldman, A. 1992: *Liaisons*. Cambridge, MA: MIT Press.

Goldman, A. 1979: What is justified belief? In G.S. Pappas (ed.), *Justification and Knowledge: New Studies in Epistemology*. Dordrecht: Reidel. Reprinted in Goldman, 1992.

Hertwig, R., Herzog, S.M., Schooler, L.J. and Reimer, T. 2008: Fluency heuristic: a model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 34, 1191–1206.

Horton, C.L., Conway, M.A. and Cohen, G. 2008: Memory for thoughts and dreams. In G. Cohen and M. A. Conway (eds), *Memory in the Real World*. Hove: Psychology Press.

Johnson, M. K. 1997: Source monitoring and memory distortion. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 352, 1733–45.

Johnson, M. K. 2006: Memory and reality. *The American Psychologist*, 61, 760–71.

Johnson, M. K., Hashtroudi, S. and Lindsay, D.S. 1993: Source monitoring. *Psychological Bulletin*, 114, 3–28.

Johnson, M. K. and Raye, C. L. 2000: Cognitive and brain mechanisms of false memories and beliefs. In D.L. Schacter and E. Scarry (eds), *Memory, Brain, and Belief*. Cambridge, MA: Harvard University Press.

Koriat, A. 1993: How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100, 609–639.

Koriat, A. 1998: Metamemory: The feeling of knowing and its vagaries. In M. Sabourin, F. Craik, and M. Robert (eds.), *Advances in Psychological Science, Vol. 2: Biological and Cognitive Aspects*. New York: Psychology Press.

Koriat, A. 2000: The feeling of knowing: some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, 9, 149–71.

Koriat, A. 2002: Metacognition research: an interim report. In T. J. Perfect and B. L. Schwartz (eds), *Applied Metacognition*. Cambridge: Cambridge University Press.

Koriat, A. and Ackerman, R. 2010: Metacognition and mindreading: Judgments of learning for Self and Other during self-paced study. *Consciousness and Cognition*, 19, 251–64.

Koriat, A. and Goldsmith, M. 1996: Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, 103, 490–517.

Koriat, A., Ma'ayan, H. and Nussinson, R. 2006: The intricate relationships between monitoring and control in metacognition: lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology. General*, 135, 36–69.

Lackey, J. 2005: Memory as a generative epistemic source. *Philosophy and Phenomenological Research*, 70, 636–58.

Lepock, C. 2011: Unifying the intellectual virtues. *Philosophy and Phenomenological Research*, 83, 106–28.

Levin, D. T. 2002: Change blindness blindness as visual metacognition. *Journal of Consciousness Studies*, 9, 111–30.

Levine, T. R., Kim, R. K., Park, H. S. and Hughes, M. 2006: Deception detection accuracy is a predictable linear function of message veracity base-rate: a formal test of Park and Levine's probability model. *Communication Monographs*, 73, 243–60.

Levine, T. R., Park, H. S. and McCornack, S. A. 1999: Accuracy in detecting truths and lies: documenting the veracity effect. *Communication Monographs*, 66, 125–44.

Loftus, E. 2005: Planting misinformation in the human mind: a 30-year investigation of the malleability of memory. *Learning & Memory*, 12, 361–66.

Loftus, E. *1979/1996: Eyewitness Testimony*. Cambridge, MA: Harvard University Press.

Loussouarn, A. 2010: *De la métaperception à l'agir perceptif*. PhD dissertation, Institut Jean-Nicod/Ecole des Hautes Etudes en Sciences Sociales, Paris.

Mascaro, O. and Sperber, D. 2009: The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, 112, 367–80.

Matthen, M. 2010: Is memory preservation? *Philosophical Studies*, 148, 3–14.

McKay, R.T. and Dennett; D.C. 2009: The evolution of misbelief. *Behavioral and Brain Sciences*, 32, 493–510.

Metcalfe, J. 2008: Evolution of metacognition. In J. Dunlosky and R.A. Bjork (eds), *Handbook of Metamemory and Memory*. New York: Psychology Press.

Michaelian, K. 2008: Testimony as a natural kind. *Episteme*, 5, 180–202.

Michaelian, K. 2009: Reliabilism and privileged access. *Journal of Philosophical Research*, 34, 69–109.

Michaelian, K. 2010: In defence of gullibility: the epistemology of testimony and the psychology of deception detection. *Synthese*, 176, 399–427.

Michaelian, K. 2011a: Generative memory. *Philosophical Psychology*, 24, 323–42.

Michaelian, K. 2011b: Is memory a natural kind? *Memory Studies*, 4, 170–89.

Michaelian, K. 2011c: The epistemology of forgetting. *Erkenntnis*, 74, 399–424.

Mitchell, K. J. and Johnson, M. K. 2000: Source monitoring: Attributing mental experiences. In E. Tulving and F. I. M. Craik (eds., *Oxford Handbook of Memory*. Oxford: Oxford University Press.

Morton, A. 2004: Epistemic virtues, metavirtues, and computational complexity. *Noûs*, 38, 481–502.

Nelson, T.O. and Narens, L. 1990: Metamemory: A theoretical framework and new findings. In G. Bower (ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory*. New York: Academic Press.

Nelson, T.O. and Narens, L. 1994: Why investigate metacognition? In J. Metcalfe and A. P. Shimamura (eds), *Metacognition*. Cambridge, MA: MIT Press.

Oppenheimer, D.M. 2008: The secret life of fluency. *Trends in Cognitive Sciences*, 12, 237–41.

Park, H.S. and Levine, T. 2001: A probability model of accuracy in deception detection experiments. *Communication Monographs*, 68, 201–10.

Proust, J. 2006: Rationality and metacognition in non-human animals. In S. Hurley and M. Nudds (eds), *Rational Animals*. Oxford: Oxford University Press.

Proust, J. 2007: Metacognition and metarepresentation: is a self-directed theory of mind a precondition for metacognition? *Synthese*, 159, 271–95.

Proust, J. 2008: Epistemic agency and metacognition: An externalist view. *Proceedings of the Aristotelian Society*, 108, 241–68.

Proust, J. 2010: Metacognition. *Philosophy Compass*, 5, 989–98.

Reber, R. and Unkelbach, C. 2010: The epistemic status of processing fluency as source for judgments of truth. *Review of Philosophy and Psychology*, 1, 563–81.

Robin, F. 2010: Imagery and memory illusions. *Phenomenology and the Cognitive Sciences*, 9, 253–62.

Schacter, D.L. and Addis, D. R. 2007: The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362: 773–86.

Schmill, M., Oates, T., Anderson, M.L., Josyula, D., Perlis, D., Wilson, S. and Fults, S. 2008: The role of metacognition in robust AI systems. *Papers from the Workshop on Metareasoning at the Twenty-Third AAAI Conference on Artificial Intelligence*.

Shanton, K. 2011: Memory, knowledge and epistemic competence. *Review of Philosophy and Psychology*, 2, 89–104.

Shieber, J. 2011: Against credibility. *Australasian Journal of Philosophy*. Forthcoming.

Smith, J. D., Shields, W.E. and Washburn, D.A. 2003: The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, 26, 317–339.

Sosa, E. 2007: *A Virtue Epistemology*. Oxford: Clarendon Press.

Sperber, D. 2001: An evolutionary perspective on testimony and argumentation. *Philosophical Topics*, 29, 401–13.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G. and Wilson, D. 2010: Epistemic vigilance. *Mind & Language*, 25, 359–93.

Stanovich, K. E. 1999: *Who is Rational? Studies of Individual Differences in Reasoning*. Mahwah, NJ: Erlbaum.

Sutton, J. 2010: Observer perspective and acentred memory: Some puzzles about point of view in personal memory. *Philosophical Studies*, 148, 27–37.

Thompson, V. A. 2009: Dual process theories: a metacognitive perspective. In J. Evans and K. Frankish (eds), *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press.

Thompson, V. A. 2010: Towards a metacognitive dual process theory of conditional reasoning. In M. Oaksford and N. Chater (eds), *Cognition and Conditionals*. Oxford: Oxford University Press.

Tulving, E. 1982: Synergistic ecphory in recall and recognition. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 36, 130–47.

Vosgerau, G. 2010: Memory and content. *Consciousness and Cognition*, 19, 838–46.

Vrij, A. 2008: *Detecting Lies and Deceit: Pitfalls and Opportunities*, 2nd edn. Oxford: Wiley-Blackwell.