Erkenntnis (2022) 87:2219–2242
https://doi.org/10.1007/s10670-020-00299-0

ORIGINAL RESEARCH

The Cognitive Philosophy of Reflection

Andreas Stephens[1] · **Trond A. Tjøstheim**[1]

Received: 31 January 2020 / Accepted: 21 July 2020 / Published online: 12 September 2020

Abstract
Hilary Kornblith argues that many traditional philosophical accounts involve problematic views of reflection (understood as second-order mental states). According to Kornblith, reflection does not add reliability, which makes it unfit to underlie a separate form of knowledge. We show that a broader understanding of reflection, encompassing Type 2 processes, working memory, and episodic long-term memory, can provide philosophy with elucidating input that a restricted view misses. We further argue that reflection in fact often does add reliability, through generalizability, flexibility, and creativity that is helpful in newly encountered situations, even if the restricted sense of both reflection and knowledge is accepted. And so, a division of knowledge into one reflexive (animal) form and one reflective form remains a plausible, and possibly fruitful, option.

1 Introduction

Throughout the history of Western philosophy, *reflection* has been considered an especially important human ability. Its role has long been prominent and can still be found at the center of theories by contemporary scholars such as, for example, BonJour (1985, 1998), Chisholm (1989), and Sosa (2007, 2009). Accordingly, a lot of effort has been invested in the inquiry of its role for thinking, knowledge, and justification. Common traditional positions have included that reflection is necessary in order to guarantee that an agent's knowledge is acceptable and certain, that her epistemic duty is fulfilled, that her knowledge is accessible, and that faulty beliefs due to inferential errors are avoided (see, e.g., Pappas 2017; see also Bortolotti 2011).

But in contrast to the above-described positions, Hilary Kornblith in his book *On reflection* (2012) points out that the common interpretation of reflection is problematic since reflection actually cannot provide that which many believe it can. Indeed much relevant research seems to indicate that rather than providing trustworthy

✉ Andreas Stephens
andreas.stephens@gmail.com

[1] Lund University, Box 192, 221 00 Lund, Sweden

knowledge, reflection can be quite unreliable. Numerous psychological studies, seemingly, show how human reflection often fails due to, for example, various biases (see, e.g., Stanovich and West 2000; Kahneman 2011). With this in mind, the importance of reflection, and its role for human thinking, knowledge, and justification, should arguably be deemphasized.

This leaves us at an interesting junction. On the one hand, reflection seems to underlie the very essence of human greatness and is commonly seen as a particularly important phenomenon. On the other hand, empirical evidence seems to support Kornblith's view and suggest that reflection only brings a false sense of certainty.

We recognize that inquiries are affected by the inquirer's stance (approach, commitments), which makes it important to briefly clarify our own. In line with Kornblith (see, e.g., 1993, 2002, 2012), we heed a naturalistic stance where philosophy needs to take relevant scientific results into account whenever such results are available. Accordingly, we accept both ontological and (cooperative) methodological naturalism, where natural phenomena and relevant scientific results are seen as more important than language or intuitions (see, e.g., Papineau 2016; Rysiew 2017; Cellucci 2017). We claim, as does Kornblith, that such a stance can offer philosophy new insights that are crucial for keeping the field relevant as well as for dissolving old problems.

In short, we believe that Kornblith's discussion of reflection is problematic due to its too-narrow understanding of what reflection brings to the table. Given this position, our aim in this article is to investigate reflection more broadly by examining relevant psychological constructs and their neural underpinnings. By stepwise investigating reflection on multiple levels of analysis, a synthesizing understanding of reflection that is biologically plausible can arguably be reached (see, e.g., Hassabis et al. 2017). This allows us to triangulate essential features of the natural phenomenon that Kornblith downplays or ignores (Horst 2016). We will, however, also argue that even if we accept a restricted view of reflection as 'second-order mental states,' as well as Kornblith's insistence on that reliability is the only epistemic value to consider, reflection, in fact, often does offer the subject added reliability. Importantly, this would leave the division of knowledge into a reflexive (animal) form and a reflective form a plausible option.

This article comprises five sections. In Sect. 2, we outline and discuss Kornblith's account of reflection. In Sect. 3, we investigate how reflection can be further elucidated by cognitive psychology, also outlining the neural correlates of reflection. In Sect. 4, we then explore philosophical consequences of the reached position pertaining to reliability and knowledge. Finally, in Sect. 5, we offer some concluding remarks.

2 Kornblith on Reflection

Kornblith (2012) argues that most traditional philosophers have valued reflection too highly due to faulty understandings of what it involves. And this overestimation has, in his view, led them to suggest, or even demand, that reflection is necessary when,

in fact, such a view is wrong. Traditional philosophers, on Kornblith's view, tend to call on reflection when problems are recognized at a first-order level. Second-order reflection is then supposed to provide a solution by removing unreliability. This, however, according to Kornblith, is problematic since neither first-order processes nor second-order reflective scrutiny are entirely reliable. Kornblith argues that his points concerning reflection are generalizable and relevant for discussions of knowledge, reasoning, freedom of the will, and normativity. In this article we will focus on his discussion of knowledge.

Importantly, Kornblith addresses reflection specifically seen as consisting in 'second-order mental states.' He further considers reliability as being the only important criteria for belief acquisition processes (Kornblith 2012, p. 34). Kornblith then attacks the traditional view from two angles. Firstly, he argues that a reliance on reflection leads to an infinite regress and that reflection thus cannot provide the sought after reliability for first-order problems. Secondly, he argues that empirical evidence indeed indicates that the processes involved in reflection often are unreliable. Both these arguments, which will be presented more fully in the following subsections, according to Kornblith shows that reflection fails to be relevant for knowledge.

2.1 Infinite Regress

As a first argument against the traditional view, Kornblith claims that demands for reflection lead to an infinite regress since it continuously would require demands of ever higher-level reflections.[1]

According to Kornblith, knowledge, in its paradigmatic formulation, is commonly held to require justified true belief. And, as pointed out by Kornblith, according to many theoreticians, justification involves reflection on the epistemic status of one's beliefs. It is then only reflection that can guarantee the right epistemic status to one's beliefs. An omission to reflect would result in beliefs that cannot be considered knowledge.

We regard this a reasonable estimate of the common-sense view, although it arguably involves an implicit internalist view of knowledge. Indeed, Kornblith starts his discussion by presenting the famous 'Norman the clairvoyant' case by BonJour (1985). In short, BonJour (an internalist) argues that an agent needs active reflection, that makes her epistemically responsible, for knowledge. This is presented, by BonJour, as an argument against reliabilism (a form of externalism) that views knowledge as involving reliably produced true beliefs, hinging on the external connection between the agent and the world.

Now, Kornblith, who is an outspoken reliabilist (see, e.g., Kornblith 2002) argues that if an agent is to meet BonJour's requirements and reflect on her beliefs, the

[1] This same point plays out somewhat differently depending on which area of philosophy one is paying attention to, but we will, as aforementioned, here focus on knowledge.

reached beliefs would themselves, in turn, need to be justified by higher-order reflection, leading to an infinite regress (Kornblith 2012, pp. 12–13).

If one accepts Kornblith's strict understanding of reflection as second-order mental states and knowledge as being dependent on reliability, this indeed seems to be the forced conclusion.

2.2 Empirical Evidence Against the Reliability of Reflection

As a second argument against the traditional view, Kornblith claims that a wide range of empirical evidence shows that reflection often is unreliable. Reflective scrutiny does then most often not succeed in making us able to more reliably judge our first-order beliefs, but seems to make subjects more confident when in fact this is not motivated (Kornblith 2012, pp. 3, 25). This would indicate that it is not a tenable option to accept the aforementioned infinite regress as an inevitability and claim that having some reflective scrutiny at least is better than having none.

Sidestepping the merely logical matter of things, a large amount of empirical evidence seemingly does support Kornblith's interpretation where reflection is best seen as only bringing a false sense of certainty to the table. In defense of his position Kornblith presents, and interprets, several empirical findings that cohere with his account. Notably, he acknowledges the tentative nature of such findings and theorizing (Kornblith 2012, p. 136). It is also important to point out that Kornblith does *not* claim that reflection is useless, rather he argues that reflection might be useful if a more realistic account of it is accepted.

Kornblith focuses on cognitive psychology and the influential dual process theory. Briefly put, reflection figures distinctly in this framework, which partitions the mental into two forms. The first form (the old mind, System 1, or Type 1) is considered to be intuitive, automatic, non-conscious, and implicit, whereas the second form (the new mind, System 2, or Type 2) is reflective, controlled, conscious, and explicit.[2] On this account, the first form generate fast reflexive responses, which the second form sometimes reflectively inhibits (Tversky and Kahneman 1974, 1983; Sloman 1996; Barrett et al. 2004; Kahneman 2011; Evans 2007, 2008; Samuels 2009; Lizardo et al. 2016; Bago and De Neys 2017).

We consider Kornblith's choice to focus on dual process theory reasonable since that framework is canonical and directly addresses aspects of cognition that are highly relevant for understanding reflection and knowledge, being supported '… by a wide range of converging experimental, psychometric, and neuroscientific methods' (Evans and Stanovich 2013, p. 224). But, we want to point out that many interpretations of dual process theory exist, addressing, for example, types, systems or modes. This said, most interpretations of dual process theory can, arguably, be integrated into a common format which makes it fruitful to explore dual process theory as a, more or less, unified field although this should be done with care (Smith and DeCoster 2000, p. 110; Evans 2003, p. 458). Moreover, it should be mentioned that

[2] Kornblith uses the terminology 'System 1' and 'System 2' whereas, for example, Evans and Stanovich (2013, p. 226) argue against such a usage to the benefit of the 'Type 1' and 'Type 2' nomenclature.

there are researchers critical of dual process theory, where critics have pointed out both faults and alternative interpretations (see, e.g., Gigerenzer and Regier 1996; Keren and Schul 2009; Kruglanski et al. 2003; Osman 2004; Kruglanski and Gigerenzer 2011). The force of these lines of critique, though, hinge on which specific form of dual process theory they attack, and, for example, Evans and Stanovich (2013) in our view convincingly counters a number of the more common ones.

Importantly, if dual process theory, more generally, is not accepted as a provider of valid empirical input, Kornblith's argument would indeed be severely stifled. However, our main point here does not involve questioning dual process theory per se. Rather we claim that Kornblith's interpretation of cognitive psychological theorizing and evidence is problematic since it too narrowly *only* focuses on dual process theory. To remain a plausible option, Kornblith's restricted position needs to be developed in a pluralist direction that investigates the many important roles reflection fills for how a subject (organism) acts in her (its) environment (see, e.g., Shah and Vavova 2014). We will in the following Sect. 3 explore what such an account of reflection involves and how it can offer philosophy elucidating input.

2.3 Reflection as Decoupled from Knowledge

Taken together, Kornblith's arguments, indeed, seem to capture essential problems with the traditional positions that he criticises; it is, it seems, deeply questionable whether reflection can solve the problems often assumed that it can. And since reflection, indeed, does take such a center stage in much philosophical discussion, Kornblith's focus is highly relevant. Kornblith interprets the reached position as indicating that theoreticians ought to abandon any false hopes regarding what reflection can provide (Kornblith 2012, p. 7).

Kornblith discusses how Sosa's (1991; see also 2007; 2009) distinction between 'animal knowledge' and 'reflective knowledge' can offer a way out of the infinite regress. On this account, animal knowledge governs direct responses to one's sensory impacts, whereas reflective knowledge governs a wider understanding of one's responses and how they came about (Sosa 1991, p. 240). Animal knowledge is then more or less what externalist theories focus on, and reflective knowledge is what internalist theories focus on. Kornblith claims that this distinction, indeed, would resolve the issue of an infinite regress. Nonetheless he continues to argue that the reflective knowledge of the bisection does not add anything extra that is superior to 'mere' animal knowledge. Kornblith discusses, and rejects, the possibility that what reflective knowledge adds is increased reliability, which is also what Sosa argues (Kornblith 2012, pp. 16–17; Sosa 1991, p. 240). Since Kornblith considers reliability crucial for knowledge he then rejects a division of knowledge, even though he acknowledges that reflection might fill some other important role(s) (Kornblith 2012, pp. 19–20).

Yet, even if we accept the restricted view of reflection as second-order mental states, and accept that reliability is of sole importance (something we believe indicates a rather strong externalist position), then if it turned out that reflective processes do add to a subject's reliability, this would, on Kornblith's own

account, rebut the infinite regress and make reflection eligible as underlying a distinct form of knowledge.

Kornblith accepts this possibility but emphatically denies that this is the case:

> We have examined a number of alternative motivations, and found that these motivations as well cannot bear the weight of the tempting distinction. It seems that there really is no ground at all for drawing a distinction between unreflective knowledge and something better, knowledge which involves reflection. (Kornblith 2012, p. 40)

We will in Sect. 4 specifically address how reflection *can* add reliability, even if the narrow account of it as only involving second-order mental states is accepted. This can be done by providing the subject with an opportunity to remember previous experiences and internally reflect on them in order to find patterns in them and then adjusting ensuing behaviors in accordance with the found patterns. In doing so the subject gains generalizability, flexibility, and creativity that is helpful in newly encountered situations. Therefore, a division of knowledge into one reflexive (animal) form and one reflective form remains a plausible, and possibly fruitful, option (see, e.g., Perrine 2014; Shah and Vavova 2014; Smithies 2016). So, although Kornblith (2012, pp. 16, 19) discusses how an allowance of two forms of knowledge could be seen as arbitrary and might risk leading to that infinitely many multiple forms must be allowed, we will below present a discussion that instead argues that two forms are biologically plausible.

But before we do this, we will next explore what a biologically plausible broader account of reflection involves and how it can offer philosophy elucidating input.

3 A Broader Understanding of Reflection

In this section, we follow Kornblith in focusing on cognitive psychology but, importantly, strive to stepwise develop a deeper multi-level investigation into reflection and its underlying processes that go beyond Kornblith's sole focus on dual process theory. This account, which also encompasses memory systems and neural correlates, offers a broader understanding of reflection that is not restricted to only involve second-order mental states. It is our belief that this account can provide philosophy with elucidating input that Kornblith's restricted focus misses.

In Fig. 1 we present a schematic illustration of how influential models from three levels of analysis cohere with each other, and how they relate to reflection. Although this is not an exhaustive account, we aim to substantiate this interdisciplinary approximation in the following discussion:

We now move to a description of how reflection is understood in cognitive psychology and find that a broader interpretation than the one Kornblith presents is motivated.

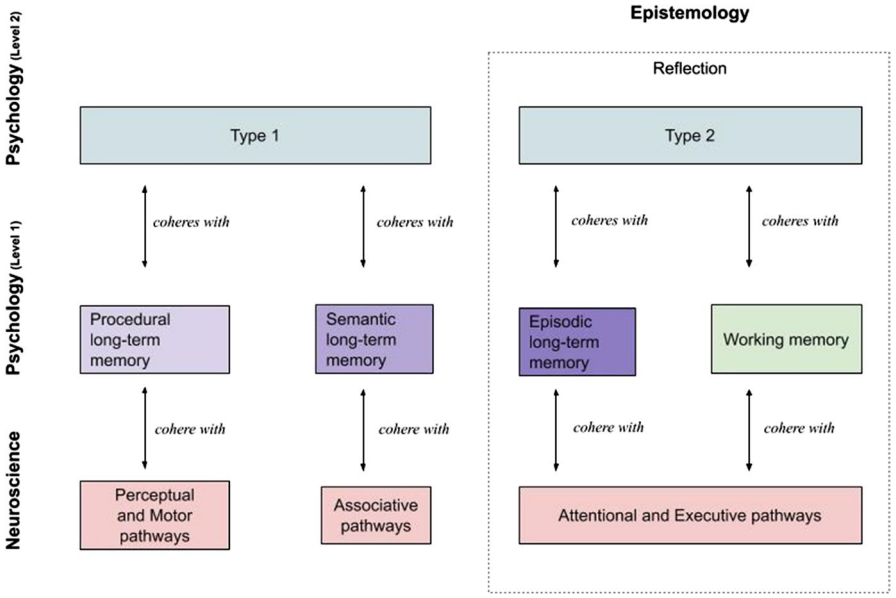

Fig. 1 Schematic illustration of relations between cognitive models, on different levels of analysis, and their relation to reflection. Four perspectives are represented: epistemology (dotted square); psychology level 2 (top row); psychology level 1 (middle row); neuroscience (bottom row). Boxes indicate model categories. Arrows indicate functional relationships

3.1 Reflection in Cognitive Psychology

In the dual process theory-literature, which is Kornblith's specific focus, reflection tends to be explicitly highlighted as an important phenomenon (see, e.g., Carruthers 2009; Mercier and Sperber 2009; Stanovich 2009; Evans and Stanovich 2013). According to dual process theory, reflection is considered to involve many specific functions linked to Type 2 processes (Evans 2008, p. 257). These complex functions encompass, for example, internal linguistics sequences or 'sentences of inner speech' (Frankish 2009, pp. 11–12; see also Carruthers 2009, p. 118), the ability to connect mental images to language, comprehend visual semantics, as well as visual manipulation (visual management) (Frankish 2010, p. 921; Carruthers 2009, p. 112). Moreover, from the perspective of dual process theory, the reflective mind is considered to include decision making, mental simulation, goal-adoption, belief-fixation, the ability for making comparisons, reasoning, metacognition in the form of second-order mental states, as well as hypothetical thinking (Evans and Stanovich 2013). Furthermore, recollection and the binding of information are dependent on reflection. It is crucial for a sense of time and to make out specific events (Yonelinas 2013, p. 2). In addition, Type 2 processes are linked to explicit rule learning (Evans 2008, pp. 257, 261, 267).

Even though human agents might not always be as in control as they believe themselves to be, these functions of reflection are important for their

self-awareness and sense of agency. All these abilities are thus plausible to see as comprising a first outline.

There is a line of critique arguing that cognition is better seen as a continuum of processes than as two distinct ones (see, e.g., Osman 2004). This has some intuitive plausibility, however, by highlighting the difference of various *forms* of dual process theories this issue can, arguably, be circumvented. As Evans and Stanovich (2013, p. 229) point out, there are indeed *modes* of processing ('cognitive styles applied in Type 2 processing') that can vary on a continuum. Specific Type 2 reflections can thus be performed in a variety of different manners. But, what most dual process theories try to point out is that there are two distinct *types* of cognitive processes, where Type 2 processes stand out as being flexible and linked to reflection. And so, '[c]ontinuous variation in both cognitive ability and thinking dispositions can determine the probability that a response primed by Type 1 processing will be expressed—but the continuous variation in this probability in no way invalidates the discrete distinction between Type 1 and Type 2 processing' (Evans and Stanovich 2013, pp. 229–230).

So, even though there are pending issues concerning how we should view reflection from the perspective of cognitive psychology, we consider it initially plausible to link reflection to Type 2 processes. To reiterate, rather than viewing reflection as problematic, dual process theory indicates that it underlies several important cognitive functions such as internal linguistics sequences or 'inner speech,' visual semantic comprehension, visual manipulation, and mental simulation (visual management for short), decision making, goal-adoption, belief-fixation, reasoning, metacognition in the form of second-order mental states, hypothetical thinking, self-awareness, and our sense of agency.

To broaden our understanding of reflection and Type 2 processes we continue by focusing on a second, 'lower,' cognitive psychological level of analysis where the human memory systems are seen as consisting of many interconnected functional processes that encode, store, retrieve, and manage information. On this level, an influential division is made between long-term memory (LTM) and working memory (WM), where LTM can store information over a lifetime whereas WM governs active information handling (see, e.g., Repovš and Baddeley 2006).[3]

LTM is commonly partitioned into an implicit (non-declarative, non-conscious) system and an explicit (declarative, conscious) system. The non-declarative system is thought to govern automatic actions, whereas the declarative system is thought to govern abstracted knowledge about the world and autobiographical remembrance. In Tulving's (see, e.g., 1972, 1985, 2002, 2005) canonical and very influential three-part model of LTM, involving procedural, semantic and episodic memory, procedural memory governs perceptual and motor skills, semantic memory governs conceptual and categorical knowledge, whereas episodic memory governs remembrance of events (Tulving 1985, p. 2). According to Tulving '… procedural memory entails semantic memory as a *specialized* subcategory, and… semantic memory, in turn,

[3] This interpretation follows a development from previous traditional theories and models that placed a more passive short-term memory (STM) in the role now commonly ascribed to an active WM.

entails episodic memory as a specialized subcategory.' (Tulving 1985, pp. 2–3, italics in original).

Regarding WM, various models have been proposed although a very influential multi-component 'standard model' presents it as consisting of four parts: the phonological loop, the visuospatial sketchpad, the central executive, and the episodic buffer (Baddeley and Hitch 1974; Baddeley 2000, 2007; Repovš and Baddeley 2006; D'Esposito and Postle 2015; Chai et al. 2018). In short, the phonological loop controls auditory information, the visuospatial sketchpad controls visual and spatial information, the central executive controls attention and decisions, whereas the episodic buffer binds together information from different domains, working as a link to (episodic) LTM.[4]

Since it is through WM we actively handle information (see, e.g., Miller 1956; Cowan 2001) we argue that it is this system—on this level of analysis—which is primarily involved in Type 2 processes and reflection (Evans 2008). To substantiate this claim we show below how WM coheres with reflection as well as to the various previously mentioned features of Type 2 processes.

The phonological loop includes the articulatory network and the sensorimotor interface (Hickok and Poeppel 2007). It is thought to consist of a phonological store that can hold acoustic information for a couple of seconds, and an articulatory rehearsal process governing subvocalization by which verbal information is kept in memory. Apart from auditory information and speech, information needs to be re-coded through articulatory rehearsal before it can enter the phonological store. Accordingly, the phonological loop connects WM to language, and thus coheres with internal linguistics sequences and inner speech (Repovš and Baddeley 2006, p. 7).

The visuospatial sketchpad consists of two separate subsystems governing visual and spatial information respectively. It is crucially connected to how we perceive the world. Interestingly, we rely on a quite small amount of information from the surrounding world—since it tends to be stable, offering us a continuing 'external memory.' However, this bottom-up information also relies on top-down predictions when being interpreted into meaningful percepts (see, e.g., Friston 2010; Hohwy 2013; but see Firestone and Scholl 2016 for a recent challenge). The visuospatial sketchpad thus coheres with previously mentioned visual management abilities (Repovš and Baddeley 2006, pp. 8, 12).

The central executive is thought to be a form of control system for the other parts of WM (Rottschy et al. 2012, Sect. 1). By controlling attention, it governs how we

[4] There are alternative interpretations that, for example, argue that WM is best viewed as being a *part* of LTM (see, e.g., Ericsson and Kintsch 1995) or as an *emergent* property of numerous combinations of underlying 'possible subsystems' (see, e.g., Postle 2006), where '… working memory may simply be a property that emerges from a nervous system that is capable of representing many different kinds of information, and that is endowed with flexibly deployable attention.' (Postle 2006, p. 29). However, in line with for example Repovš and Baddeley (2006), we regard the empirical findings as providing a strong case for the standard model. Even so, we do acknowledge that it might have to be revised in a more fine-grained direction in light of coming findings, where feasible examples of such revisions might include, not only auditory- and visual-, but more subsystems based on all our different senses in WM.

prioritise, choose, and execute tasks. It is also involved in all information-manipulation (Repovš and Baddeley 2006, p. 14), composing reasoning as well as decision making and planning. But although being a central hub within WM, the central executive nonetheless has a limited degree of attention (see, e.g., Miller 1956; Cowan 2001). This means that the central executive coheres with abilities such as decision making, goal-adoption, belief-fixation, reasoning, metacognition in the form of second-order mental states, and hypothetical thinking.

The episodic buffer works as an interface between WM and LTM systems (Repovš and Baddeley 2006, p. 15). More specifically, it relates information between the central executive and episodic LTM '… forming a limited-capacity system for the ultra-short-term, intermediate storage of incoming sensory information' (Rottschy et al. 2012, Sect. 1). Through a store of limited capacity, it integrates information from the other components of WM into episodes. In doing so the episodic buffer is involved in creating conscious awareness. The episodic buffer binds recollected information, connecting to episodic LTM, which composes explicit rule learning (Strange et al. 2001, p. 1045). This interface thus processes and stores multi-dimensional representations (Rudner and Rönnberg 2008, p. 21). By doing so it helps to create a unitary experience, which is central for our self-awareness, sense of agency, and first-person phenomenological experience:

> Measures of working memory capacity have been shown to be predictive of performance in a wide variety of cognitive tasks… and highly correlated with fluid intelligence… It is the engagement of this system specifically that… has [been] emphasized in the definition of Type 2 processing and which underlies many of its typically observed correlates: that it is slow, sequential, and correlated with measures of general intelligence. [It] has also [been] suggested that Type 2 thinking enables uniquely human facilities, such as hypothetical thinking, mental simulation, and consequential decision making. (Evans and Stanovich 2013, p. 235)

In summary, we have shown that WM governs our internal linguistics sequences and connects to language (the phonological loop), our visual management (the visuospatial sketchpad), our attention, information-manipulation, reasoning, metacognition in the form of second-order mental states, and decision making (the central executive), as well as binds recollected information (the episodic buffer and episodic LTM). In view of the above discussion, we, therefore, claim that Type 2 processes and WM (also relying on episodic LTM) plausibly cohere with reflection.

3.2 Neural Correlates

By exploring the neural underpinnings of reflection, we in this subsection substantiate and ground our understanding of reflection in cognitive neuroscience. We argue that cognitive neuroscience is a suitable level at which to stop for our purposes, as this level provides information about plausible functionality of neural populations. Notably, such information can be effectively mapped to neural network architectures in a computer.

From the neuroscientific perspective, bottom-up perceptive pathways can be dissociated from top-down feedback pathways. The bottom-up pathways are activated by sensory stimuli, tending to align with statistical regularities in the sensorium by various process-signal amplifications (Pozo and Goda 2010). Collectively these processes contribute to the formation of distinctive receptive fields in the sensory cortices. The sensory streams are associated and bound together in association areas, which make up concept-like complexes that are presented to frontal populations involved in executive control (Tanaka 1996; Tsunoda et al. 2001; Caporale and Dan 2008; Magee and Johnston 1997; Ralph et al. 2010).

These frontal networks project back into the sensory pathways, which afford modulation of the perceptive streams via excitation and inhibition. This is the filter of attention, where certain aspects are turned down while others are amplified. Although the particulars of this process are still not fully known, there are indications that such top-down amplification is necessary to realize fine detail from a coarser bottom-up signal (Ahissar and Hochstein 2004).

Focusing on WM, it is closely associated with the processes and pathways of selective attention and executive control (Awh et al. 2006). Information may flow from the exterior world via the senses, or it may come from LTM.

The act of reflecting is, as described above concerning the phonological loop, often associated with internal linguistic sequences—internal monologues (Alderson-Day and Fernyhough 2015). An internal monologue involves both the production of speech as well as its interpretation. The former is realized by the posterior inferior temporal gyrus, premotor cortex, and the anterior insula, making up the articulatory network, along with the sensorimotor interface consisting of the sylvian parietal-temporal area (Hickok and Poeppel 2007). Interpretation, on the other hand, is realized by populations in the posterior middle temporal gyrus and posterior inferotemporal gyrus, making up the lexical interface (Kemmerer 2014). Semantic and grammatical aspects are integrated by the combinatorial network found predominantly in the lateral anterior temporal lobe. Together these pathways mediate understanding of conceptual content of speech. In short, this suggests that the articulatory network (posterior inferior temporal gyrus, premotor cortex, anterior insula), and the sensorimotor interface (sylvian parietal-temporal area) cohere with the phonological loop.

Although there are indications that all sensory modalities are available to WM (vision and audition: Baddeley and Hitch 1974; Baddeley 1986; tactility: Katus et al. 2012; proprioception: Smyth et al. 1988; olfaction: Zelano et al. 2009; somatosensation: Zhou and Fuster 1996), humans, as a species, are to a large degree reliant on vision in order to navigate and interact with the world (D'Ardenne et al. 2012; Brewer et al. 2011; Mason et al. 2007). The visuospatial sketchpad handles the visual and spatial information we encounter, which can be broken down into a number of sub-functions (Repovš and Baddeley 2006). For example, there appears to be a dissociation between purely visual representation, and representation of space as such (Constantinidis and Wang 2004). Spatial WM may be representing space generally, for visual, auditory, or other stimuli, and appears to be mediated by a network involving the dorsolateral prefrontal cortex, superior temporal cortex, posterior parietal cortex, and the lateral intraparietal lobe (Constantinidis and Wang 2004). These

sites are lateral. On the medial side, the anterior cingulate cortex, posterior cingulate and retrosplenial cortices, and the parahippocampal cortex are involved (Constantinidis and Wang 2004). Parietal areas generally mediate integration of sensory streams, while the dorsolateral prefrontal cortex is usually thought to be responsible for maintaining and storing representations (though see Mackey et al. 2016 for a challenge to this in humans). Visual representations in particular also make use of networks in the occipital lobe (see, e.g., Schurgin 2018). These areas thus together cohere with the visuospatial sketchpad.

The most important cortical area for executive function, or cognitive control, appears to be the frontal cortex. A recent review by Badre and Nee (2018) identifies several regions within frontal cortices that mediate central executive control functionality of varying concreteness. In general, more abstract control is found in rostral areas, while concreteness increases caudally, closer to sensory cortices. Thus, the frontal eye fields and the premotor and motor cortices handle concrete sensorymotor control (Badre and Nee 2018). Contextual control is found more rostrally in the dorsal- and ventral anterior (pre) premotor areas, also including the inferior frontal junction area (Badre and Nee 2018). More rostrally still are areas that handle control of context-independent schemas. These include the mid-dorsolateral prefrontal cortex, and the rostrolateral prefrontal cortex (Badre and Nee 2018). In this context, schemas may be thought of as a kind of mental structures that organize classes of percepts and their relationships (Bartlett 1932). These, and other areas such as the frontostriatal circuits, brainstem, and superior parietal cortex cohere with the central executive.

As mentioned, the episodic buffer functions as a mediator between many memory systems, especially between the central executive and episodic LTM (Baddeley et al. 2010). When retrieval is needed for planning and executive control, the episodic buffer helps integrate relevant information (Strange et al. 2001, p. 1045; Rudner and Rönnberg 2008). Although the exact role and underpinnings of the episodic buffer remain unclear, particularly the parietal lobe and the left anterior hippocampus is thought to play a crucial role, in how this temporary storage, with a limited capacity, merge information (Berlingeri et al. 2008; Baddeley et al. 2010). This is enabled by a capacity for multi-dimensional coding, giving the episodic buffer a central role for conscious awareness, as well as for immediate- and episodic recall. Episodic memory is a broad concept, integrating sensory streams along with a sense of space, place, and time, but also a sense of agency. In the brain, this means that diverse and widespread networks are recruited to encode and reconstruct episodes. One of the most important networks is thought to be the hippocampus. Coarsely, it is responsible for spatiotemporal aspects of memory organization, as well as for relations between memories (Eichenbaum 2018). Also involved is the parahippocampal gyrus which more specifically processes aspects of place (Eichenbaum 2018). The ventromedial prefrontal cortex and the angular gyrus process self-referential aspects, and the feeling of agency respectively (Dede and Smith 2018). The middle temporal gyrus is thought to handle semantic aspects of episodes (Dede and Smith 2018). Included in episodic memory networks are neural populations related to attention. The retrosplenial and posterior cingulate cortices are involved in reducing attention and engaging the default network, which can reconstruct episodes. The ventrolateral

prefrontal cortex is also thought to be able to break established attentional patterns to direct attention to other salient events (Corbetta and Shulman 2002; Eriksson et al. 2015). Similar mechanisms to manipulation of chunks may make up the affordance of mental time-travel and mental simulation, which appear to rely on recalling sequences from LTM and somehow parameterizing them. The hippocampus, in particular, appears to be involved with this, but likely in concert with prefrontal populations (Hassabis et al. 2007). Information from LTM route via the default network (Brewer et al. 2011; Mason et al. 2007). Specifically, there are indications that the fusiform gyrus, the inferior temporal and parahippocampal gyri, as well as the left posterior insula, are activated above baseline when gating of LTM is in effect (Brewer et al. 2011).

In this subsection, we have investigated the neural underpinnings of reflection and WM. Although the various parts of WM are interconnected, working in parallel with LTM and numerous other systems, a number of specific brain areas pertaining to selective attention and executive control do stand out. The articulatory network (posterior inferior temporal gyrus, premotor cortex, anterior insula), and the sensorimotor interface (sylvian parietal-temporal area) coheres with the phonological loop. The dorsolateral prefrontal cortex, superior temporal cortex, posterior parietal cortex, lateral intraparietal lobe, anterior cingulate cortex, posterior cingulate, retrosplenial cortices, and the parahippocampal cortex, as well as the occipital lobe, coheres with the visuospatial sketchpad. The frontal and prefrontal cortex, the premotor and motor cortices, also involving frontostriatal circuits, brainstem, and superior parietal cortex coheres with the central executive. The parietal lobe and the (left anterior) hippocampus coheres with the episodic buffer. And, the prefrontal, ventral fronto-temporal, medial temporal, retrosplenial, and posterior cingulate cortices, the parahippocampal, angular, middle temporal, the fusiform, and inferior temporal gyrus, as well as the left posterior insula and the hippocampus coheres with episodic LTM.[5] In short, the processes and pathways of selective attention and executive control cohere with WM and so Type 2 processes and reflection (Awh et al. 2006).

The reached position is thus that reflection involves Type 2 processes, WM and episodic LTM, as well as attentional and executive neural pathways. Reflection can

[5] Research on the cerebellum indicates that it plays a vital role not only in fine motor behaviour, but also in the automation of mental processes. According to Ito (2008), the cerebellum has two principal modi of operation: as a forward model, and as an inverse model. The former implies that the cerebellum can learn to generate and hence simulate sensory signals. The latter means that the cerebellum can learn to control, for example, muscles in the motor system, but may also be interpreted as to involve populations of excitatory and inhibitory neurons that affect contents of WM. Thus, the cerebellum can learn to perform volitional operations in WM automatically. Common examples of this is mental calculation, and certain kinds of planning (Ito 2008). This can be interpreted as the cerebellum being necessary for higher order thought, or being able to automate sequences of thought into building blocks that can be used for more complex problem solving or planning. Further aspects could, for example, include the function of glial cells in signal delay and the function of protein synthesis in regulating density of receptors or neurotransmitter reuptake mechanisms.

thus be differentiated from Type 1 processes, procedural and semantic LTM, as well as perceptual, motor, and associative neural pathways.[6]

We want to point out that even though this partitioning is well-established, highlighting an essential feature of human cognition, both reflexive and reflective processes involve complex intertwined bottom-up and top-down signals that work together. In the following Sect. 4, we will try to elaborate on this interaction.

3.3 Interpreting, Operationalizing and Measuring Reflection

Above, psychological constructs and their neural underpinnings, on multiple levels of analysis, have shown the natural phenomenon reflection to be multifaceted and complex, involving much more than just second-order mental states. This broader understanding of reflection thus provides input that more narrow accounts risk to miss. It is a dual understanding of cognition that emerges, which seemingly ought to influence our view of what a plausible account of knowledge should consist in.

But Kornblith questions the philosophical relevance of psychological findings and theories on the matter of reflection generally. He argues that there is an important difference between how 'reflection' is used in psychology and how it is used in philosophy (Kornblith 2012, pp. 141–142):

> While System 2 is often the source of second-order belief, not all of the beliefs produced by System 2 are second-order, and thus when psychologists speak of System 2 as involved in reflection, their use of that term better accords with everyday usage, which allows that we may reflect on various features of the world around us and not just on features of our mental life, than it does with the technical usage here which ties reflection to second-order states. (Kornblith 2012, p. 140)

Here Kornblith points out that he uses reflection in a technical sense. Accordingly, he accepts that Type 2 processes (System 2) involve other aspects, but considers that the only philosophically relevant aspect is the link to second-order mental states. From a cooperative methodological naturalistic perspective philosophers should look to science for answers rather than make up their own based on intuition, which makes it questionable to restrict scientific input in this manner. And as we have shown above, a broader interpretation is motivated. However, if the traditional view that Kornblith wants to counter demands that reflection is restricted to one of its aspects—second-order mental states—it might be necessary to do so for argument's sake. It is then only the empirical evidence specifically addressing metacognitive second-order mental states that should be considered.

But Kornblith goes further. According to Kornblith, psychological theorists 'mean to say nothing more [with the term reflection] than that the kind of thought characteristic of System 2 is conscious' (Kornblith 2012, p. 141). Reflection

[6] Importantly, semantic memory is connected to both procedural and episodic memory although we will regard it as closer tied to reflexive generalized processes and thus not view it as directly involved in reflection (see, e.g., Binder and Desai 2011; Yee, Chrysikou, and Thompson-Schill 2014).

should then be understood as 'nothing more than' conscious reasoning in System 2 (Type 2 processes)—also involving non-conscious processes from System 1 (Type 1 processes). But we consider this interpretation to be insufficient and problematic. It is one thing to restrict one's focus (to second-order mental states)—against the scientific usage found in cognitive psychology. However, in claiming that cognitive psychologists (or even only dual process theorists) mean nothing more than 'consciousness' when they speak of reflection, we believe Kornblith is in the wrong.

Contrary to Kornblith's interpretation, cognitive psychologists point out how 'the reflective mind' governs our thinking dispositions, having a number of important specific roles, where 'reasoning and decision making sometimes requires both (a) an override of the default intuition and (b) its replacement by effective Type 2, reflective reasoning.' (Evans and Stanovich 2013, p. 236). Rather than indicating 'nothing more' than consciousness, reflection can be seen to encompass many particular states in human cognition, but importantly second-order mental states about one's own thoughts is a focal point where '[c]onclusions accepted for a reason are not intuitive but are, we will say, "reflective"… and the mental act of accepting a reflective conclusion through an examination of the reasons one has to do so is an act of reflection' (Mercier and Sperber 2009, p. 12).

Currently, a common way of operationalizing reflection in the context of cognitive psychology research is by means of the 'cognitive reflection test' (CRT) (see, e.g., Frederick 2005; Campitelli and Labollita 2010; Toplak, West, and Stanovich 2011; Vandekerckhove et al. 2014; Gronchi et al. 2016). The idea of this experimental test is to measure the disposition or ability of a subject to resist the first answer that comes to mind when posed with a set of questions. These questions are deliberately posed in a way to yield different answers if the subject uses quick intuitions, or if they deliberate and reflect. Here is a common example: *A bat and a ball cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?*

The intuitive, quick answer is that the ball costs 10 cents. The correct answer, however, is 5 cents. The original CRT consists only of three questions, including the one posed above and two similar ones, and subjects are given the following instruction: *Below are several problems that vary in difficulty. Try to answer as many as you can*. The measure consists in counting the number of correct answers. Having said that, the test is usually not presented alone, but as part of a larger questionnaire where time and risk preferences are asked for. Perhaps unsurprisingly, studies using the CRT show a correlation between correct answers and reduced temporal discounting (Fredrick 2005). In other words, people that tend to answer correctly tend also to be more patient than those who go with the intuitive answer.

This is all very well, but what does it tell us about the epistemic value of reflection? First of all, it indicates that reflexive, or first-order beliefs may not always be reliable since there is a tendency for the brain to jump to conclusions when effort is involved in making an inference. Second, in the cases pertinent to the CRT, reflection is limited to second-order; i.e., there is no infinite regress. Thirdly, it implies that in many cases truth checking may have to be done with external support, e.g., with pen and paper. The point of this is only that representing symbols in the environment saves on mental energy as it were, since the symbols no longer have to be

kept stable in the mind. This makes it less likely that energy saving processes get activated, which again can yield inaccurate conclusions.

In a sense, this can be interpreted as lending weight to Kornblith's criticism of reflection; it can be unreliable. However, importantly so can reflexive processes. The CRT supports that trains of thought can indeed be unreliable since the brain is prone to be miserly with its resources, and this can lead to inaccurate conclusions. But it appears that at least some of these limitations can be overcome by cognitive offloading onto the external world. Hence the process of second-order thought understood as truth checking intuitions can add reliability and epistemic value.

We have looked to cognitive psychology and gained a multi-level understanding of reflection going beyond second-order mental states, which has enabled a more informed interpretation. While this indicates the advantage of a broader understanding of reflection, we will in the next section grant the more restrictive view of reflection and knowledge. It will however be shown that even on such an account, a division of knowledge into a reflexive and a reflective form remains a plausible option.

4 The Plausibility of Two Forms of Knowledge

As shown in the previous section, reflection fills many important roles, but most crucially for our discussion we will in this section discuss how it adds reliability—even restrictively understood as 'second-order mental states,' which from a scientific perspective involves a view of reflection as consisting purley of metacognition. In accordance with Kornblith's own argument, a division of knowledge into one reflexive (animal) form and one reflective form thus remains a plausible option.

4.1 Reflection can Add Reliability

Reflection in fact does add reliability since a pure reliance on reflexive processes would in many cases be costly because observations risk being too context-specific (see, e.g., Smithies 2016). To test each encounter purely on the merits of current observational stimuli could even lead to disaster. The ability to run multiple test-scenarios, amounting to second-order mental states about previous trials, in one's head has great survival benefits. Agents can use reflection to generalize and abstract away non-essential information thereby gaining an overarching understanding and knowledge. A sole focus on reflexive processes thus risks to only allow specific context-dependent knowledge of specific cases. Reflection, seen as second-order mental processes (metacognition), adds generalizability, flexibility, and creativity that is helpful in newly encountered situations, and this, in turn, adds reliability (see, e.g., Olsson 2017a).

The bottom-up pathways that originate in sensory neurons can automatically associate with each other and with behaviour. By being exposed to a variety of stimuli, they can generalize in their own way and do limited extrapolations based on similarities, and on trial and error. These pathways have evolved to support survival and procreation, and are hence usually able to do an admirable job if

left to their own devices. The limitation of the bottom-up pathways is in their context-specificity. If there is no outward similarity for the senses to latch on to, no behaviour will match. This can result in arbitrary and inappropriate behaviour, fearful behaviour and withdrawal from the situation, or anxiety and no behaviour at all. This is where top-down pathways, second-order mental states, and reflective behaviour comes in. Away from the situation, in a calm and safe place, sensory sequences can be recalled and be played back. Different alternative behaviours can be simulated and evaluated, amounting to thinking about one's thinking or second-order mentals tates, so as to hopefully cope better with similar situations in the future.

The top-down pathways, governing second-order mental states, can inhibit particularities in the sensory streams and hence discover common patterns in them. Particularities of instances of a category are often represented by higher frequency information, while commonalities tend to be represented by lower frequency information (Wiskott and Sejnowski 2002). In general, however, instance particularity is not limited to high frequencies, and full generalization requires an ability to inhibit any kind of property representation, be it shape, sound, or smell. Inhibition carries a burden of effort though (Dixon and Christoff 2012), and humans have learned to use external representations such as drawings to aid in abstract pattern identification and to reduce cognitive load (Risko and Gilbert 2016).

Reflection also affords the extraction of patterns from one context, and the re-concretization of those patterns into different contexts, using imagination to fill in required and appropriate detail. This can save a tremendous amount of energy that would otherwise be needed to arrive at the same behaviour in each specific context via trial and error. To be sure, large differences between the constructed scenario and the actual one may occur. And to an extent, the success of such an enterprise depends on the quality of the second-order models that are employed. That is, how well an agent understands the contexts in question. If both source- and target contexts are understood, re-concretization has a good chance of being successful, otherwise, the probability remains low. Even if the projected behaviour fails, a plan can still be made to gather information in the given context such that correct behaviour can be learned.

Crucially, during the reflective phase, information from cultural sources can be integrated to change behaviour. Human beings can communicate and exchange experience and knowledge, and through writing and reading that experience can be communicated across larger distances and over longer time spans. By means of writing, knowledge about the world can also accumulate over time affording later generations better cognitive methods and tools than previous ones. Such information integration is not possible purely by bottom-up experience of concrete situations, even if direct situational information is more accurate than that generated by means of reflection.

So, reflection, even if solely understood as second-order mental states (or metacognition), *can* add reliability through added flexibility and generalizability for the agent. In the next section, we will go into more depth about the contrast between reflective and reflexive knowledge from the perspective of feedback control.

4.2 Reflective and Reflexive Knowledge

Since it has been shown that reflection can add reliability, Kornblith's account can be evaluated anew. He agrees that if this is the case, the infinite regress (from Sect. 2.1) can be avoided. And this would leave the option of dividing knowledge into two forms, one reflexive (animal) and one reflective. In this subsection we elaborate on this possibility.

Even though the body (including the central nervous system with the brain) forms essentially a unified system under feedback control, it is nevertheless governed by distinct reflexive and reflective pathways (Pezzulo and Cisek 2016; see also, e.g., Friston 2009, 2010; Hohwy 2013). Top-down pathways continuously predict activity of bottom-up sensory pathways, while prediction errors make their way upwards in the hierarchy until they can be adjusted for by activating effectors. Here 'effector' is used as a broad term for processes that bind together and affect other processes, including, for example, low-level hormonal upregulation, reflexive motor actions initiated by spinal cord networks, as well as behaviour guided by high-level plans such as walking to a store to buy food, or even applying to college to get an education. So, albeit that human cognition and knowledge involve several complex intertwined capabilities, they are plausibly partitioned into a reflexive and a reflective form.[7]

Reflection can be interpreted as willful manipulations of WM content using such metaphorical effectors. This process can be applied to question and check the validity of spontaneous intuitions. Take the example from the CRT mentioned above, where the question is what the price of the baseball is given that both the bat and ball cost $1.10, and the bat costs $1 more than the ball. The spontaneous first-order thought is that the ball costs 10 cents. What reflection can do is to check more thoroughly if this is indeed the case. By laboriously setting up an algebraic equation and doing the math step by step, the original intuition can be scrutinized. In this case it was wrong; the mathematics yield the answer 5 cents. As long as this second-order process is trusted, as is usually the case with arithmetics, there is no need for further verification.[8]

Summing up, we claim that Kornblith is correct when he points out that traditional philosophical investigations often do not do justice to the natural phenomenon of reflection. Indeed, folk-psychological notions of reflection ought not to be allowed to take precedence or override scientifically grounded understandings of the natural phenomenon. But the reached conclusion is that philosophy needs to accept a pluralistic account of reflection and knowledge that acknowledges both reflexive and reflective processes that each provide specific information relevant for knowledge (see, e.g., Plotkin 1993; Alston 2005; Olsson 2017b). Moreover, Kornblith's

[7] This also holds true, to various degrees, for all mammals, and many other organisms (see, e.g., Allen and Fortin 2013; Carruthers 2013).

[8] Interestingly, the scientific process can be seen as an example of a kind of infinite regress, since there is seldom a 100% sure probability of experimental validity, and 100% validity can never in practice be reached. But experimental results can converge, which means that further experimentation becomes less urgent. Hence the regress, and the reflection, can be halted.

own interpretation of reflection is problematic, even given his own demarcations and demands. Importantly, there is a link between reflection and reliability making two forms of knowledge a plausible option—one reflexive (animal knowledge) and one reflective.

5 Concluding Remarks

We have shown that a better understanding of reflection is possible by looking at how it actually works. We have therefore moved away from a traditional stance focusing on language, concepts, certainty, and truth. Instead, we have adopted a naturalistic stance, in line with Kornblith, focusing on natural phenomena, scientific results, and plausibility. In accordance with this stance, we have explored how reflection coheres with the psychological constructs Type 2, WM, and episodic LTM, as well as to attentional and executive neural pathways. Importantly, reflection has been shown to fill a number of important functions: our inner dialogues, visual management, attention, information-manipulation, reasoning, decision making, metacognition, sense of agency, self-awareness, first-person phenomenology, remembrance, and awareness, motivating a pluralist account.

But we have also argued that this, more fine-grained, understanding of reflection, also acknowledging the influence and role of reflexive processes, does tie reflection to reliability by providing generalizability, flexibility, and creativity that is helpful in newly encountered situations. This indicates that the possibility to divide knowledge into a reflexive form and a reflective form is a plausible option, contrary to Kornblith's view.

Acknowledgements We want to thank Peter Gärdenfors, Martin L. Jönsson, Christian Balkenius, Maximilian Roszko, Asger Kirkeby-Hinrup, Erik J. Olsson, and Ingar Brinck for sharing their vast knowledge concerning this topic. We also want to thank participants at the Research Seminar in Theoretical Philosophy and the PhD Seminar in Philosophy at Lund University, and our anonymous reviewers for comments.

Funding Open access funding provided by Lund University. The authors gratefully acknowledges support from Makarna Ingeniör Lars Henrik Fornanders fond and Stiftelsen Elisabeth Rausings minnesfond: forskning.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 8*(10), 457–464.

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin, 141*(5), 931–965.
Allen, T. A., & Fortin, N. J. (2013). The evolution of episodic memory. *Proceedings of the National Academy of Sciences of the United States of America, 110*(Supplement 2), 10379–10386.
Alston, W. P. (2005). *Beyond "justification": Dimensions of epistemic evaluation*. Ithaca, New York: Cornell University Press.
Awh, E., Vogel, E. K., & Oh, S. H. (2006). Interactions between attention and working memory. *Neuroscience, 139*(1), 201–208.
Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.
Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Science, 4*(11), 417–423.
Baddeley, A. D. (2007). *Working memory, thought and action*. Oxford: Oxford University Press.
Baddeley, A. D., Allen, R. J., & Hitch, G. (2010). Investigating the episodic buffer. *Psychologica Belgica, 50*(3), 223–243.
Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York: Academic Press.
Badre, D., & Nee, D. E. (2018). Frontal cortex and the hierarchical control of behavior. *Trends in Cognitive Sciences, 22*(2), 170–188.
Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition, 158,* 90–109.
Barrett, L. F., Tugade, M. M., & Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin, 130*(4), 553–573.
Bartlett, F. C. (1932). *Remembering: An experimental and social study*. Cambridge: Cambridge University Press.
Berlingeri, M., Bottini, G., Basilico, S., Silani, G., Zanardi, G., Sberna, M., et al. (2008). Anatomy of the episodic buffer: A voxel-based morphometry study in patients with dementia. *Behavioural Neurology, 19*(1–2), 29–34.
Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences, 15*(11), 527–536.
BonJour, L. (1985). *The structure of empirical knowledge*. Cambridge, MA: Harvard University Press.
BonJour, L. (1998). *In defense of pure reason: A rationalist account of a priori justification*. London: Cambridge University Press.
Bortolotti, L. (2011). Does reflection lead to wise choices? *Philosophical Explorations, 14*(3), 297–313.
Brewer, J. A., Worhunsky, P. D., Gray, J. R., Tang, Y. Y., Weber, J., & Kober, H. (2011). Meditation experience is associated with differences in default mode network activity and connectivity. *Proceedings of the National Academy of Sciences, 108*(50), 20254–20259.
Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgment and Decision Making, 5*(3), 182–191.
Caporale, N., & Dan, Y. (2008). Spike timing-dependent plasticity: A Hebbian learning rule. *Annual Review of Neuroscience, 31,* 25–46.
Carruthers, P. (2009). An architecture for dual reasoning. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 109–127). Oxford: Oxford University Press.
Carruthers, P. (2013). Evolution of working memory. *Proceedings of the National Academy of Sciences, 110*(Supplement 2), 10371–10378.
Cellucci, C. (2017). *Rethinking knowledge: The heuristic view* (Vol. 4). Dordrecht: Springer.
Chai, W. J., Abd Hamid, A. I., & Abdullah, J. M. (2018). Working memory from the psychological and neurosciences perspectives: A review. *Frontiers in Psychology, 9,* 401.
Chisholm, R. M. (1989/1966). *Theory of knowledge* (3rd Ed.). Englewood Cliffs, NJ: Prentice Hall.
Constantinidis, C., & Wang, X. (2004). A neural circuit basis for spatial working memory. *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry, 10*(6), 553–565.
Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*(3), 201–215.
Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and Brain Sciences, 24*(1), 87–114.
D'Ardenne, K., Eshel, N., Luka, J., Lenartowicz, A., Nystrom, L., & Cohen, J. D. (2012). Role of prefrontal cortex and the midbrain dopamine system in working memory updating. *Proceedings of the National Academy of Sciences of the United States of America, 109*(49), 19900–19909.

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology, 66*(1), 115–142.
Dede, A. J. O., & Smith, C. N. (2018). The functional and structural neuroanatomy of systems consolidation for autobiographical and semantic memory. *Current Topics in Behavioral Neurosciences, 37,* 119–150.
Dixon, M. L., & Christoff, K. (2012). The decision to engage cognitive control is driven by expected reward-value: Neural and behavioral evidence. *PLoS One, 7*(12), e51637.
Eichenbaum, H. (2018). What versus where: Non-spatial aspects of memory representation by the hippocampus. *Current Topics in Behavioral Neurosciences, 37,* 101–117.
Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*(2), 211–245.
Eriksson, J., Vogel, E. K., Lansner, A., Bergström, F., & Nyberg, L. (2015). Neurocognitive architecture of working memory. *Neuron, 88*(1), 33–46.
Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454–459.
Evans, J. S. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove: Psychology Press.
Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59,* 255–278.
Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.
Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *The Behavioral and Brain Sciences, 39*(e229), 1–72.
Frankish, K. (2009). Systems and levels: Dual-system theories and the personal–subpersonal distinction. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 89–107). Oxford: Oxford University Press.
Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass, 5*(10), 914–926.
Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives, 19*(4), 25–42.
Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences, 13*(7), 293–301.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138.
Gigerenzer, G., & Regier, T. (1996). How do we tell an association from a rule? *Comment on Sloman. Psychological Bulletin, 119*(1), 23–26.
Gronchi, G., Righi, S., Parrini, G., Pierguidi, L., and Viggiano, M. P. (2016). Dual process theory of reasoning and recognition memory errors: Individual differences in a memory prose task. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 331–335).
Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron, 95*(2), 245–258.
Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences of the United States of America, 104*(5), 1726–1731.
Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402.
Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.
Horst, S. (2016). *Cognitive pluralism*. Cambridge, MA: MIT Press.
Ito, M. (2008). Control of mental activities by internal models in the cerebellum. *Nature Reviews Neuroscience, 9*(4), 304–313.
Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus and Giroux.
Katus, T., Andersen, S. K., & Müller, M. M. (2012). Nonspatial cueing of tactile STM causes shift of spatial attention. *Journal of Cognitive Neuroscience, 24*(7), 1596–1609.
Kemmerer, D. (2014). *Cognitive neuroscience of language*. New York: Psychology Press.
Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science, 4*(6), 533–550.
Kornblith, H. (1993). *Inductive inference and its natural ground: An essay in naturalistic epistemology*. Cambridge, MA: MIT Press.

Kornblith, H. (2002). *Knowledge and its place in nature*. Oxford: Oxford University Press.
Kornblith, H. (2012). *On reflection*. Oxford: Oxford University Press.
Kruglanski, A. W., Chun, W. Y., Erb, H. P., Pierro, A., Mannett, L., & Spiegel, S. (2003). A parametric unimodel of human judgment: Integrating dual-process frameworks in social cognition from a single-mode perspective. In J. P. Forgas, K. R. Williams, & W. von Hippel (Eds.), *Social judgments: Implicit and explicit processes* (pp. 137–161). New York: Cambridge University Press.
Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberative judgements are based on common principles. *Psychological Review, 118*(1), 97–109.
Lizardo, O., Mowry, R., Sepulvado, B., Stoltz, D. S., Taylor, M. A., Van Ness, J., et al. (2016). What are dual process models? Implications for cultural analysis in sociology. *Sociological Theory, 34*(4), 287–310.
Mackey, W. E., Devinsky, O., Doyle, W. K., Meager, M. R., & Curtis, C. E. (2016). Human dorsolateral prefrontal cortex is not necessary for spatial working memory. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 36*(10), 2847–2856.
Magee, J. C., & Johnston, D. (1997). A synaptically controlled, associative signal for Hebbian plasticity in hippocampal neurons. *Science, 275*(5297), 209–213.
Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007). Wandering minds: The default network and stimulus-independent thought. *Science, 315*(5810), 393–395.
Mercier, H., & Sperber, D. (2009). Intuitive and reflective inferences. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 149–170). Oxford: Oxford University Press.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97.
Olsson, E. J. (2017a). Coherentism. In S. Bernecker & K. Michaelian (Eds.), *The Routledge handbook of philosophy of memory* (pp. 310–322). London: Routledge.
Olsson, E. J. (2017b). Explicationist epistemology and epistemic pluralism. In A. Coliva & N. J. L. L. Pedersen (Eds.), *epistemic pluralism* (pp. 23–46). Cham: Palgrave Macmillan.
Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review, 11*(6), 988–1010.
Papineau, D. (2016). Naturalism. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition). https://plato.stanford.edu/archives/win2016/entries/naturalism/.
Pappas, G. (2017). Internalist vs. externalist conceptions of epistemic justification. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition). https://plato.stanford.edu/archives/fall2017/entries/justep-intext/.
Perrine, T. (2014). Against Kornblith against reflective knowledge. *Logos & Episteme, 5*(3), 351–360.
Pezzulo, G., & Cisek, P. (2016). Navigating the affordance landscape: Feedback control as a process model of behavior and cognition. *Trends in Cognitive Sciences, 20*(6), 414–424.
Plotkin, H. C. (1993). *Darwin machines and the nature of knowledge*. Cambridge, MA: Harvard University Press.
Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience, 139*(1), 23–38.
Pozo, K., & Goda, Y. (2010). Unraveling mechanisms of homeostatic synaptic plasticity. *Neuron, 66*(3), 337–351.
Ralph, M. A., Sage, K., Jones, R. W., & Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proceedings of the National Academy of Sciences of the United States of America, 107*(6), 2717–2722.
Repovš, G., & Baddeley, A. (2006). The multi-component model of working memory: Explorations in experimental cognitive psychology. *Neuroscience, 139*(1), 5–21.
Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688.
Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A. R., Schulz, J. B., et al. (2012). Modelling neural correlates of working memory: A coordinate-based meta-analysis. *Neuroimage, 60*(1), 830–846.
Rudner, M., & Rönnberg, J. (2008). The role of the episodic buffer in working memory for language processing. *Cognitive Processing, 9*(1), 19–28.
Rysiew, P. (2017). Naturalism in epistemology. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition). https://plato.stanford.edu/archives/spr2017/entries/epistemology-naturalized/.

Samuels, R. (2009). The magical number two, plus or minus: Dual-process theory as a theory of cognitive kinds. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 129–146). Oxford: Oxford University Press.

Schurgin, M. W. (2018). Visual memory, the long and the short of it: A review of visual working memory and long-term memory. *Attention, Perception, & Psychophysics, 80*(5), 1035–1056.

Shah, N., & Vavova, K. (2014). Review: On reflection by Hilary Kornblith. *Ethics, 124*(3), 632–636.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*(2), 108–131.

Smithies, D. (2016). Reflection on: On reflection. *Analysis, 76*(1), 55–69.

Smyth, M. M., Pearson, N. A., & Pendleton, L. R. (1988). Movement and working memory: Patterns and positions in space. *Quarterly Journal of Experimental Psychology Section A, 40*(3), 497–514.

Sosa, E. (1991). Knowledge and intellectual virtue. *Knowledge in perspective: Selected essays in epistemology* (pp. 225–244). Cambridge: Cambridge University Press.

Sosa, E. (2007). *A virtue epistemology: Apt belief and reflective knowledge* (Vol. I). New York: Oxford University Press.

Sosa, E. (2009). *Reflective knowledge: Apt belief and reflective knowledge* (Vol. II). New York: Oxford University Press.

Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic and autonomous minds: Is it time for a tri-process theory? In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 55–88). Oxford: Oxford University Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645–665.

Strange, B. A., Henson, R. N. A., Friston, K. J., & Dolan, R. J. (2001). Anterior prefrontal cortex mediates rule learning in humans. *Cerebral Cortex, 11*(11), 1040–1046.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience, 19*(1), 109–139.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition, 39*(7), 1275–1289.

Tsunoda, K., Yamane, Y., Nishizaki, M., & Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience, 4*(8), 832–838.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). New York: Academic Press.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26*(1), 1–12.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53*(1), 1–25.

Tulving, E. (2005). Episodic memory and autonoesis: Uniquely human. In H. Terrance & J. Metcalfe (Eds.), *The missing link in cognition: Origins of self-reflective consciousness* (pp. 3–56). New York: Oxford University Press.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293–315.

Vandekerckhove, M., Bulnes, L. C., & Panksepp, J. (2014). The emergence of primary anoetic consciousness in episodic memory. *Frontiers in Behavioral Neuroscience, 7,* 210.

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation, 14*(4), 715–770.

Yee, E., Chrysikou, E. G., & Thompson-Schill, S. L. (2014). Semantic memory. In K. Ochsner & S. Kosslyn (Eds.), *The Oxford handbook of cognitive neuroscience: Volume 1, core topics* (Vol. 1, pp. 353–374). Oxford: Oxford University Press.

Yonelinas, A. P. (2013). The hippocampus supports high-resolution binding in the service of perception, working memory and long-term memory. *Behavioural Brain Research, 254,* 34–44.

Zelano, C., Montag, J. M., Khan, R., & Sobel, N. (2009). A specialized odor memory buffer in primary olfactory cortex. *PLoS ONE, 4*(3), 829–839.

Zhou, Y., & Fuster, J. (1996). Mnemonic neuronal activity in somatosensory cortex. *Proceedings of the National Academy of Sciences of the United States of America, 93*(19), 10533–10537.