# Extending Metacognition: An Account of How Procedural and Analytic Metacognitive Processes Interact with Extended Cognition.

*A thesis presented in partial fulfilment of the*
*requirements for the degree of*

**Master of Arts**
**in**
**Psychology**

**at Massey University, Manawatū,**
**New Zealand**

**Nicholas Alexander Firth**

**2022**

*"You are not out of the loop; you are the loop."*

Daniel Dennett (2003, p. 242)

## Abstract

This thesis examines the relationship between extended cognition and metacognition by way of three interlocking proposals. First of all, both extended cognition and metacognition should be conceptualised as sub-personal-level explanations that are implemented in the brain and environment and in cultural practices that inform individual skill. Secondly, the procedural metacognition norm of fluency, analytic metacognition, and cognitive skill mutually reinforce and enrich each other when dealing with cognitive obstacles. Finally, my third claim, builds on and refines claims one and two when I examine the involvement of metacognition in relation to expertise; specifically, I focus on the skilled interplay of automaticity and metacognitive control when confronted with cognitive obstacles. To this end, I build on hybrid accounts of skilled cognitive performance to provide a framework that isolates cases of metacognitive extension. This thesis concludes that metacognition, rather than being viewed as wholly internal, can be partially externalised across the environment when the individual exhibits high levels of automaticity and control when using an artefact.

**Acknowledgements**

Thanks to Stephen Hill for your intellectual curiosity, pragmatism, and interdisciplinary ethos.

To Mum and Paul, thank-you for your thoughtfulness, love and encouragement for which I am eternally grateful.

Thanks to Karina for your reflections from the veterinary field on that nebulous space where knowledge converges with action and becomes skill. Most of all, thank-you for your happiness, love and encouragement.

To the non-human animals – the cats, ducks, dogs, and tūis – who have hopped, skittered, slinked, run and flown in and out of my life during the writing of this thesis; thank-you for being such good company and for reminding me that humans aren't the only species to metacognise.

# Contents

## List of Figures



## List of Tables

# 1. Introduction

When we consider metacognition we might have in mind Rodin's sculpture of 'the thinker' with hand cupped around chin, deep in reflection. Detached on his plinth, the cognising appears inward, internal and, if we were to point to its location, presumably somewhere behind the thinker's furrowed brow. But what would it mean to say that 'the thinker's' thoughts are not only self-directed, but embodied, world-involving and sometimes even constituted by the environment? How could this idea be pursued? An initial step might be to move away from the idea that a thinker is engaged in a realm of pure thought that is abstracted from the epistemic feelings that give rise to it. Another step might be to consider that standing apart - isolated from an environment - is indeed a rare form of cognition indeed, and insofar as it is occurs, it is derived developmentally and evolutionarily from the other deeply embedded – and sometimes extended – processes that are rely on perception-action cycles.

The research ties together two strands of research: *extended cognition and 4-E cognitive research*, and *metacognition*. First of all, the extended cognition thesis is the idea that coupling with artefacts outside of the head can count as cognitive processes when integrated with one's internal processes (Clark & Chalmers, 1998; Clark, 2008; Menary, 2010a; Miłkowski et al., 2018). More broadly, cognitive extension is nested within a larger scope of theories oftentimes referred to as 4-E cognition: cognition that is *embodied*, *enactive*, *embedded* and *extended* (Newen et al., 2018) or 'wide' cognition (see section 1.2). The thesis accepts that cognition can under particular circumstances extend, and largely focuses on a pluralistic view on extension (see Miłkowski et al., 2018, that dovetails with several waves of extended cognition Sutton, 2010; Gallagher, 2018) and 4-E cognition more generally (Newen et al., 2018). As for the second strand of research on metacognition, this thesis endorses Proust's (2013) account of *procedural metacognition* and *analytic metacognition*. Metacognition is captured, in its most neutral and inclusive form, by the following definition:

> [T]he set of capacities through which an operating subsystem is evaluated or represented by another subsystem in a context-sensitive way (p. 4).

I propose that Proust's (2013) account of metacognition presents an ideal version of metacognition for extended cognition as it emphasises procedural, affect-based, and experience-based evaluations of one's own cognitive processes, is activity-dependent and

context sensitive, and has an inherently normative structure (see section 1.7). Proust (2013) also makes space for analytic (i.e., conceptual) forms metacognition (see section 1.8). Metacognition in turn is broadly consonant with overlapping themes and commitments in the 4-E cognition tradition (Newen et al., 2018). More specifically the *normative* dimension of metacognition, and its skilful deployment in relation to extended cognition, is under-explored in the literature. The normative structure and inherent reflexivity of metacognition, especially when integrated with skill, acts as a welcome bridge-point between the metacognition and extension.

In **chapter one** I introduce extended cognition (and 4-E) cognition theories, metacognition; and how these connect in the existing literature. In **chapter two** I will examine in closer detail how non-propositional representation and nonconceptual content offer a format that is amenable to metacognition and extended cognition. In **chapter three**, I will argue that extended cognition metacognition exists at the subpersonal explanatory level. **Chapter four** examines normativity as a binding phenomenon that connects cognitive extension and metacognition. **Chapter five** looks at metacognition in relation to a automaticity and control. A brief **conclusion** follows in which I wrap up the main ideas and what future directions of research might look like.

Although the chapters stand alone and explore different ideas along the way, they also overlap and mutually reinforce one another. To that end, the following three claims and overall thematic concerns can be distilled from the thesis.

1. Claim *One*: Procedural metacognition and extended cognition are both subpersonal-*level* explanatory phenomena (chapter three). These exist on the foundation of analogue representations and nonconceptual in (chapter two) that become integrated with norms drawn from cultural practices (chapter four)
2. Claim*: Two*: Fluency and analytic norms mutually reinforce one another in cognitive action (chapters two, four and five).
3. *Claim Three.* Skilled metacognition, when coupled with high levels of control and automaticity, and integrated with an artefact characterise metacognitive extension (chapter five).

### 1.1. Extended Cognition and the 4-E Cognition Family of Views

There is a rich philosophical and scientific history that has influenced and continues to inform the theory of 4-E cognition (embodied, enactive, embedded, and extended) or 'wide cognition' (see Miłkowski et al., 2018). Inspiration has come from a wide-ranging and continent-spanning array of traditions. Detailed analyses of embodiment, perception, action and sociality can be found in the Phenomenological tradition of Edmund Husserl, Maurice Merleau-Ponty (1945), Martin Heidegger (1926) and others (See Gallagher & Zahavi, 2021). 4-E cognition was prefigured in the work of American pragmatists such as Dewey, Pierce, and James (see Gallagher, 2017). Another point of contact is the Logical Behaviourism of Gilbert Ryle[1] (1949) and Psychological Behaviourism of B.F. Skinner (1964), as exemplified by Skinner's line that "the skin is not that important as a boundary." (p. 84).[2] Vygotsky's socially-mediated accounts of cognitive processes (1934, 1978), J. J. Gibson's (1979) ecological psychology, and the philosophical naturalism of Dennett (1969, 1978, 1996) have been influential. The tradition of British cybernetics has led to developments in modelling of and theorising about embodiment and adaptive behaviour (Dewhurst, 2018). Relatedly, work on situated robotics and dynamical systems theory has done much to quantify and refine agent-environment interactions (Chemero, 2009). In addition to this, naïve realism theories in the philosophy of perception posit that external objects and properties play a *constitutive* (rather than strictly *causal*) role in shaping consciousness itself (Fish, 2009); also see (Noë, 2012) on extended perception. Taken together, these historical forerunners and more recent ideas cluster around overlapping commitments toward what can be broadly conceived as an embodied and externalist view of cognition and experience. Moreover, the radicalism implicit in 4-E cognition is well captured by what Hurley (1998) posits as the rejection of the 'sandwich model' of perception, cognition and action. More specifically, the 'sandwich model' assumes that *input* is carried into the system, cognitive processes *'interpret'* the input, and finally the *output* is released in the form

[1] Logical behaviourism, popular in the first half of the twentieth-century, was a way of examining mental states in terms of dispositions to behave in particular ways. It overlaps with psychological behaviorism. See Graham (2019).

[2] Arguably enactivist ideas (and some of the 4-E cognition more generally) bears some resemblance to behaviourism. Alksnis and Reynolds (2021). In many ways, this is to the advantage of 4-E theorists because it deflates criticisms of 4-E critics who have objected that enactivism is too close to behaviourism. It is clear that to be a behaviourist is deemed undesirable and serves as *reductio ad absurdum* (and is not dissimilar to calling someone a *Cartesian* as instant means of defeating an argument). Alksnis and Reynolds emphasise that so much depends on how behaviourism is defined (e.g., with one can endorse Tolman's molar behaviourism instead of more reductive varieties). Nonetheless, this is not the place for a full-throated defence of some versions of behaviourism; but, in short, one way of diffusing the objection that 'enactivism is behaviourist' is to embrace a richer, less reductive form of behaviourism.

of action. It is precisely this 'sandwich' model, with its neatly delineated layers, that 4-E cognition disrupts in various diverging but generally similar ways. I'll briefly look at some of the main forms of 4-E cognition – embodied, enactive, embedded and extended - in turn:

*Embodied* cognition focusses on how the body shapes and partly constitutes thinking processes (Shapiro, 2019). It's worth pointing out that embodied cognition theorists are offering something stronger than the truism that cognition is in some way embodied. Instead, a deeper point is being offered about the intertwinement of embodiment and cognition.[3] Shapiro and Spaulding (2021) note that that embodied cognition concerns how an embodied subject conceptualises the world; in particular, how the morphology of the body shapes the capacity to conceptualise (also see Gallagher 2005 for an account of this is). Lakoff and Johnson's (1980) embodied approach to metaphor and Barsalou's (1999) sensorimotor connection with concept formation have been influential strands in this approach. Another key influence is Thompson, Varela and Rosch (1991) *The Embodied Mind: Cognitive Science and Human Experience.* There is also a phenomenological and political dimension of embodied cognition that is connected with feminism, lived experience, and identity (De Beauvoir, 1949; Young, 2005).

*Enactive* cognition is focused on how embodied cognition is action-oriented (Thompson, 2007) and perceptual, and, as such, cognition is *enacted*.[4] The ability to sense and locomote is intricately and constitutively interwoven with cognition. Cognition is not an essence or substance inside of something. The world is not *pre-given* and cognition and perception do not involve recovering this pre-given world. Instead, cognition – and the world the organism inhabits - is brought forth by action and sense-making; so cognition is a process rather than a substance (Varela et al., 1991). There are various forms of enactivism, the fine distinctions of which are not too relevant here. But, briefly, one can taxonomise them, following Ward et al. (2017), in the following way: There is *Autopoietic enactivism*, which is concerned with the biodynamics of living systems and how these systems self-organise (as put forth by Thompson, 2007); secondly, there is *Sensorimotor enactivism*, which is concerned with contingencies between perception and action in relation to intentional and phenomenal content - this strand

[3] Sometimes the term *embodied cognition* is used as a metonym for a broad range of associated views including extended cognition, enactivism, and alternatives to computationalism drawn from dynamical systems theory, ecological psychology, and connectionism.

[4] One potential criticism is that perception is *not* cognition; however, this is contestable. I follow Allen (2017) in saying "given that areas such as perceptual psychology and machine perception fall fully within the scope of cognitive science, it is not a distinction affects the range of phenomena properly studied by cognitive scientists." (p. 4234).

of enactivism is typified by the philosophers such as Noë (2004); Ward et al. note that this strand is less concerned with an overarching account of the mind. Thirdly, Hutto and Myin (2013) have put forward what they term *Radical Enactive Cognition*. One of the key tenets of this radical enactivist variant is anti-representationalism; Hutto and Myin characterise the other two perspectives as importing too many cognitivist assumptions regarding intentional content. Differences aside, these perspectives have overlapping commitments.

*Embedded Cognition* is the that idea that rather than viewing cognition as extended (see below) it is instead preferrable to view cognition as heavily dependent on the environment. This idea converges with scaffolded cognition;[5] this is the framework that theorises about the way in which environmental resources can amplify cognition (see Sterelny, 2004, 2010). Critics of the extended view (e.g., Adams & Aizawa, 2008; Rupert, 2004) are understandably supportive of an embedded view. Indeed, an embedded approach can be viewed as having a long history in cognitive science (albeit with varying levels of emphasis). Vold and Schlimm (2020) point out that the embedded approach can be found in the work of pioneering cognitive scientist Herbert Simon (1969); "Simon (1969), for example, argues that much of the apparent complexity of cognitive systems is actually external to the agent and residing in the environment. On this view, cognitive systems lean heavily on worldly complexity without internalizing it." (Vold & Schlimm, 2020, p. 3763).

*Extended Cognition* is arguably the more radical of the *E*s as it advances the hypothesis that cognition (and the mind) literally extended into the environment such that parts of the environment realise or constitute cognitive processes (Clark & Chalmers, 1998). It is largely a question about the *location* of cognitive states and processes. Further, the domains of interest that extended theorists have been attracted to range from folk-psychological intuitions about the location of beliefs to questions about the way cognitive mechanisms are demarcated (Pöyhönen, 2014). Extended cognition debates have gone through various iterations (this will be focussed on in greater detail in section 1.2. – 1.6.). Extended cognition forms one of the central threads in this thesis.

[5] The concept of *scaffolding* is often attributed to Vygotsky, although see Shvarts and Bakker (2019) for details on how Vygotsky didn't use the word scaffolding (apart from once in a notebook). It was Jerome Bruner who introduced the term and it has become popular since. Nonetheless, Vygotsky's legacy has, among other concept, been identified with *scaffolding*, and many of the characteristics of Vygotskian theorising can be expressed with this term.

There are related theories that are conceptually adjacent and overlap somewhat to the 4-Es just mentioned. *Distributed cognition* refers to shared processes that take place across a variety of people and artefacts (e.g. Hutchins, 1995a). This can include extended cognition/mind theories; however, distributed cognition is also subtly different, and - as Sutton et al. (2010) point out - it occupies a space between an embedded and an extended view. Hutchins (1995a), who is credited it with developing the distributed cognition framework, is famous for a canonical study of cognitive ethnography onboard a US Navy ship in which he detailed how the cognitive processes and representational states of ship navigation were propagated across interactions between the entire crew and their specialised tools, and, further, how these processes were constrained by socio-historical-cultural practices. In addition, Hutchins (1995b) has examined how instruments are used in the flightdeck of a commercial airliner to the effect that aircraft speed[6] is computed and 'remembered' by a cognitive system composed of humans and technology; thus, cognition can be viewed as distributed, and the primary unit of analysis not that of a single individual but a socio-technical system that has cognitive properties too. Hutchins (2014) has noted that distributed cognition is a *perspective* on cognition rather than a *kind* of cognition, and, more precisely, that the extended cognition hypothesis is nested in an overall distributed cognition perspective. Cussins (2012) diagnoses the difference as follows: "The discussion in Clark and Chalmers [1998] is about *where* we find cognitively useful representations, whereas the ambition of Hutchins (1995) is to change how we understand computation and representation." (p. 17).

Finally, the school of *ecological psychology* is associated with the theories of J. J. Gibson (1979) and is concerned with the way in which organisms perceive and use their environment - and the way the environment in turn affords itself and can be used in various organism-centric ways such that the need for internal computation and representation is obviated (see Baggs & Chemero, 2021; Chemero, 2009). Ecological psychology rejects the 'poverty of stimulus' arguments[7] that state that the environment is not rich enough to supply the relevant stimuli to the organism. Instead, ecological psychology proposes that the world itself *is* sufficiently detailed. In the paradigmatic Gibsonian example of visual perception, the retinal image is not the starting point wherein a snapshot with limited environmental information needs to be

[6] More precisely, the speed memory tasks Hutchins (1995a) focuses on concerns the plane's slower speed shortly after take-off and before landing; during these times, mechanisms called slats and flaps on the plane's wings are configured to give the plane more 'lift' (to prevent stalling). Changing the slats and flaps configuration so as to maintain minimum speed is thus crucial.

[7] These poverty of stimulus arguments were famously deployed in the context of language acquisition. Chomsky famously used them in linguistics to argue for an innate grammar and it was in the context of arguing against what were (then) primarily behaviourist ideas of language acquisition.

supplemented with brain-based representations and inference; instead, the organism uses 'ecological information' *in* the environment - gradient textures and ambient optical arrays - that constitute *affordances*[8] that they can then perceive and act upon (Gibson, 1979).

Taken together, this assemblage of 4-E and 'wide cognition' viewpoints overlap. For example, ecological psychology and the extended mind; or even enactivism and extended views (Gallagher, 2018). Nevertheless, there are tensions between some of the 4-E views. For example, some forms of the extended cognition thesis can allow for representations, and, in theory, are neutral as to the physical substrate in which the cognition is taking place (i.e., the physical substrate is irrelevant as far as its being *realised* is concerned). Stapleton and Thompson (2009) criticise extended views from an enactivist standpoint; and Shapiro (2019) has drawn out the tension between the two views.

The point of emphasis in this thesis is extension; however, given the overlap extension has with adjacent 4-E views (e.g., distributed cognition, enactivism,), some of these views, and their associated terminology, will appear and reappear throughout the thesis. For the moment, however, I will focus in finer detail on *extended cognition*, as this is central to the thesis.

### 1.2. Extended Cognition and Extended Mind

Extended cognition and extended mind has been variously termed *active externalism* (Clark & Chalmers, 1998), *vehicle externalism* (Hurley, 1998), *environmentalism* (Rowlands, 1999). In this thesis I'll use the standard nomenclature of extended cognition and mind (with an emphasis on cognition over mind for reasons that I'll outline shortly).

The extended literature has gone through several phases or 'waves' (Gallagher, 2018; Sprevak, 2019). These waves can of course overlap, but, as Gallagher (2018) points out, expressing them as waves has heuristic value. It is instructive to examine the differences between the original, functionalist[9] ('first wave') conception of the extended cognition and later waves. Indeed, the first-wave mostly recognisable by way of its emphasis on functional isomorphisms between the internal components of cognition and external artefacts; in other words, on how external processes function like inner-processes.

[8] Arguably some of the ideas associated with affordances were anticipated by Henri Bergson (1991 / 1896); for example, Bergson writes: "The objects which surround my body reflect its possible action upon them." (p. 21). It is also well documented that Gibson was strongly inspired by the Phenomenologist Merleau-Ponty (see Baggs and Chemero, 2021).

[9] Piccinini (2010) notes that there is a difference between functionalism and computationalism. Moreover, computationalism and representationalism can be (in theory) separated. See chapter 2 in this thesis for more details on how this relates to metacognition and the extended cognition thesis.

> If, as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process. (1998, p. 8).

The first thought-experiment originally put forward by Clark and Chalmers (1998) is in relation to extended *cognition* processes. The idea ran like this: (a) we are invited to think of someone watching shapes on a screen and this person is asked which shapes will fit particular templates; in effect, the person is rotating a blocks in their head; (b), we are asked to think of them being given the option of either doing it in their head (as in case a), or by pressing a button. And, finally, (c) we can consider someone who has received a neural implant that has the function of rotating the shape. This neural implant operates at the same speed as the button-technology in (b), and the person is given the option of running with the chip or their usual rotation. Clark and Chalmers then ask to what extent we can characterise these three situations as being cognitive. On their view, *all* three cases exhibit functional equivalency and can be deemed cognitive. More precisely, if we are to claim (c), that a brain implant could be genuinely cognitive, then what, aside from a prejudice against external artefacts, would mean we would not classify the second option as cognitive? Indeed, while this thought-experiment is designed to test our intuitions, Clark and Chalmers make clear that it was inspired by research by Kirsh and Maglio (1994) in which physically rotating a shape by 90 degrees took approximately 300 milliseconds (100 milliseconds to rotate, and 200 to select the button); this was much faster than doing so 'in the head,' which took around 1000 milliseconds. Kirsh and Maglio argued that this physical action was not merely instrumental; it amounted to an *epistemic* action insofar as deciding whether the shape would fit happened during this physical action.

The second thought experiment Clark and Chalmers (1998) put forward is concerned with the extended mind, and it has gained canonical status in the literature. It involves a belief state of a proposition; crucially, this state is *dispositional*, so it is accessible, but nor currently considered (i.e., it is not an *occurrent* belief). In the thought experiment we are introduced to two people. Otto has a memory impairment; in the paper his diagnosis is Alzheimer's, although the specifies of the diagnosis are not detailed. Meanwhile, the second person, Inga, has intact memory. The difference between them is that Otto has a notebook in which he stores addresses (such as the address for the Museum of Modern Art in New York); Inga, however, navigates her way around the city with her biological memory. As in the block-rotation example, we are asked to consult our intuitions in relation to the *parity principle* and to bracket biases we might

have toward excluding non-neural artefacts from being classified as cognitive. Indeed, to this end, Clark and Chalmers provide criteria, sometimes known as the 'trust and glue'[10] conditions. Here are the conditions as Clark presented them in 2008:

1. That the resource be reliably available and typically invoked.
2. That any information thus retrieved be more-or-less automatically endorsed. It should not usually be subject to critical scrutiny (e.g., unlike the opinions of other people). It should be deemed about as trustworthy as something retrieved clearly from biological memory.
3. That information contained in the resource should be easily accessible as and when required.
4. That the information in the note notebook has been consciously endorsed at some point in the past and indeed is there as a consequence of this endorsement.* (p. 79)

These conditions are not jointly necessary and sufficient (Clark has characterised them as "rough-and-ready" criteria (See Clark, 2010, p 46); they were designed to ward off various counterexamples. Furthermore, while these conditions were used for the extended mind thought experiment (of dispositional belief states), rather than cognition, the conditions have been liberally applied to many different examples of cognition.

It should come as no surprise that the status of these conditions, and inclusion of other candidate conditions, has been the subject of lively debate.[11] For example, the fourth condition was designed to counter concerns that extension could *over*extend; as Clark puts it in 2008, abandoning the fourth condition potentially "opens the floodgates to what many would regard as an unwelcome explosion of potential dispositional beliefs" (p. 80). After all, we would be loath to say that a library or the contents of the internet - including content one has never encountered - is part of our cognitive system or mind. That is why the past conscious endorsement condition (number four above) was designed to avoid this threat of 'cognitive bloat'[12] (for an early charge of cognitive bloat, see Rupert, 2004). As alternative to the fourth

[10] The term *trust and glue* seems to have made its first appearance in Clark (2010); although of course the parity conditions to which it refers were introduced in Clark and Chalmers (1998).

[11] Sometimes the debates tend to mirror those in epistemology regarding whether 'justified true belief'(JTB) is sufficient for knowledge, and how the JTB conditions can be modified so as secure knowledge in the face of scepticism or luck.

[12] Nonetheless, fourth condition is problematic as it is potentially too restrictive, and would appear to rule as non-cognitive thoughts that have not been endorsed. Even Clark and Chalmers (1998) questioned the necessity of it as it would appear to hold external memories to an even higher standard than biologically-based memories. But Clark in 2008, as noted above, supports it, although with the proviso that it "looks too strong" (p. 80). Meanwhile,

condition (the past-endorsement condition), Gallagher (2018) observes that the best response to the cognitive bloat objection is to focus on the specific way action couples with an artefact; for instance, Gallagher notes that "coupling that involves reciprocal causal relations where outputs are recycled as inputs (Clark, 2008)" (p. 425), and that this puts a constraint on the threat of cognitive bloat.

Aside from the various issues with these conditions, the guiding idea is that unfairly privileging the internal over the external can amount to a form of what Clark calls "bio chauvinism." By this Clark means a form of neuro-centrism that posits the brain as the sole unit of inquiry. Elsewhere, Clark has noted that the parity principle amounts to a "kind of veil of metabolic ignorance" (2005, p. 2; see also: 2007, p. 167; and 2008, p. 114).[13] Smart (2022) prefers to see it as 'equality of opportunity' instead of functional equivalence. For Sprevak (2009), the principle is spelled out in terms of 'fair treatment.' In effect, these different labels converge on the same principle that, although the brain is a necessary organ for cognition, the brain should not be unfairly privileged.

### 1.3. Extended Mind or Cognition?

Carter et al. (2018b) note that Clark (see: 2008) makes no clear difference (in fact makes no distinction) between beliefs and processes, and so he views them as largely interchangeable.[14] The extended mind thesis is "slightly stronger," according to Carter et al. (2018b, p. 3). Drayson (2010) points out that it largely depends on one's broader ideas on what constitutes 'the mental.' Some theorists, for example Gertler (2007), propose that mind is conscious and conscious only, and thus this excludes extended mentality.[15] Indeed, the concept of the *mind* has often been associated with phenomenology (Gallagher & Zahavi, 2021), affect and

Chalmers in 2019 observes that endorsement on retrieval (rather than prior to) may be enough to deal with cognitive bloat. Some theorists looked at ways of replacing it (see Piredda & DiFranceso, 2020) with the criterion of transparency.

[13] The existence of *neurocentrism* is by no means a strawman viewpoint. For example, the neurocentric position is explicitly captured in the following quote: "[T]he brain's capacity to acquire knowledge, to abstract and to construct ideas … is a philosophical burden which neurobiology has to shoulder if it is to understand better the workings of the brain." (Zeki, 1999, p. 2054). It is this style of neurocentrism that, in part, motivates extended views in general (and the parity principle in particular). This view has been criticised by León and Zahavi (2022).

[14] And in more recent times Clark (2020) has seen no difference between desire and belief, however this need not detain us here.

[15] But this definition is rare, and also leaves *extended cognition* untouched. Some versions of the thesis propose that consciousness is extended. This is of course much more radical than claiming that cognition is extended. (See Noë, 2004, 2009; and, more recently Kirchhoff and Kiverstein (2019)). Chalmers (2019) has recently claimed that while he believes that thesis works as a theory of cognition, this is not so for consciousness as the bandwidth required for consciousness is too great. In any case, this is not the concern for the present thesis as I am not arguing for extended consciousness.

sensation (Pöyhönen, 2014), and sometimes, more problematically, with non-derived representation (see Menary, 2010b; see section 3.4. in this thesis). There appears to be no agreement on what constitutes the 'mark of the mental' or 'cognitive' in the extended cognition/mind and 4-E cognition literature (Allan, 2017; Sprevak, 2019). Nonetheless, Allen (2017) observes that the lack of consensus should not trouble philosophers and cognitive scientists in their practices, and by way of illustration Allen invokes the concept *life* in biology as an analogy. "Biology has achieved enormous success without such a definition, and it is far from clear what would be gained by a definitional ruling whether (say) viruses alive or not." (p. 4239). Allen notes that regardless of the 'life status' of a virus, biologists nonetheless continue to research viruses using the same methods applied to cellular life. Allen observes that the same can be said for cognition: cognitive scientists have and will continue to proceed with research in the absence of a definitive mark of the cognitive.[16] Indeed, Allen observes that cognition is an *umbrella concept* that refers to a range of capacities – including memory, problem-solving, and perception – and so it is unsurprising that there has been a lot of contention around how to define it. Allen suggests that real progress concerns studying specific capacities. That is partly why I follow Pöyhönen (2014) in claiming that it is important to keep extended cognition and mind hypotheses distinct; even though they are both umbrella concepts, they at least roughly map on to different explanatory targets (i.e., cognitive processes or dispositional memory of propositions). To the end, and for my purposes, throughout the thesis I will only refer to cognitive extension rather than mind. This consideration of the need to define capacities more precisely in part motivates focusing on metacognitive capacity as a particular form of cognitive extension (see chapter five).

### 1.4. Second-Wave Extended Cognition

The second-wave has typically focussed on *complementarity* (Sutton, 2010) and *cognitive integration* (Menary, 2007). Implicit in the second-wave is a critique of the first-wave. One concern is that functionalism is present in the first-wave accounts of Clark (2008) and Wheeler (2010). In short, the idea of functionalism is that cognition is governed by the *roles* cognition states and processes play in a cognitive system. This is in contrast to considering the *constitution* of these processes and states; in effect, functionalism is at a higher level of

[16] Smart (2022) has a counterview that we do need a view on what constitutes a cognitive status; however, Smart also notes that the need for a status equally affects *internal* accounts of cognition; so, in effect, the absence of a consensus definition of cognition does not uniquely affect externalist accounts of cognition.

abstraction than biological implementation (Fodor, 1975). What matters is the causal relations between cognitive states and processes. (Insofar as extended cognition is functionalist, Sprevak (2009) has offered a critique.) First wave extended cognition can be viewed as exhibiting *multiple realisability;*[17] this is the controversial theory that cognitive and mental states can be realised (i.e., instantiated) by different physical mediums. Thompson and Stapleton (2009) highlight some of the tensions between the enactivist cognition and more functionalist-oriented positions in the extended literature (also see Gallagher, 2017). While it is not the time to adjudicate on this matter, for now it is important to note that these concerns with functionalism are partly what theoretically motivate the case for the *second-wave* iteration of extended cognition. More precisely, Sutton (2010) deems the parity principle to be too generic and as emphasising in too coarse a way supposed functional isomorphisms between biological and external components. On Sutton's view, parity principle does not do justice to the sheer heterogeneity of and complementarity between external interactions and artefacts.[18]

Relatedly, another concern is that that the first wave still conceives of the brain as having a kind of priority, and evinces the idea that cognition is primarily contained[19] within the head and only on occasion extends - or 'leaks'[20] – out into the world (a view that for some theorists demonstrates the undesirable persistency of Cartesian ideas). In this way – and Clark's intentions notwithstanding - cognitive extension would appear to keep intact the conventional, cranium-based boundaries of the mind/cognition. Now of course this is not to suggest that the second-wave view denies that there is an important asymmetry between the brain and non-neural processes; rather, the point is that one can highlight this asymmetry without implying that cognition literally extends in the manner of a *substance* reaching out into the world.

[17] Boone and Piccinini (2016) provide a classic example: a carburettor's role of mixing fuel and air in an internal combustion engine. The carburettor will perform this role regardless of what material constitutes or 'realises' it; the material could consist of cast iron, plastic, zinc, or aluminium; and it could be in a car, motorbike or lawnmower. Functionalist accounts of cognition and the mind have typically made recourse to such analogies (See chapter three of this thesis for more on this).

[18] However, see (Wheeler, 2011), who observes that parity does not necessarily entail identical inner / outer components.

[19] Ideas of containment appear to have an deeply engrained, folk-psychological appeal; they also have a deep history. Knappett et al. (2010) examine and reflect on the archaeological evidence of (non)ceramic pots, vessels, and figurines from Mesolithic and Neolithic periods in relation to embodied and metaphoric notions of 'containment;' these early artefacts, with their boundaries, interiors and exteriors, were not only useful products of early cultural life. Knappett et al also speculate that "containment may have a more basic and to a large extent neglected role in the shaping of human intelligence." (p. 590).

[20] Menary (2010b) points out that some of the rhetoric in favour of extension can be infelicitous. For example, Clark doesn't help his case when he notes that "Cognition leaks out into body and world" (2008, p. xxviii). Nevertheless, Menary cautions us to ignore the rhetorical excesses and focus on the (on his view, persuasive) arguments and evidence for extension

Sutton et al. (2010) have on occasion pitched the second-wave as having significant overlap with Sterelny's (2010) scaffolding hypothesis, and, more specifically, occupying conceptual space between an embedded and distributed view (that may commit the theorist to an extended view too). Additionally, Sutton et al. (2010) state that the parity-style 'trust and glue' criteria should be viewed dimensionally, as a matter of degree, rather than categorically. Elsewhere, second-wave extension theorists invoke what is known as the mutual manipulability (MM) criterion (Craver, 2007) drawn from theoretical research on the nature of mechanisms. Indeed it was Menary (2006, 2007), who drew upon Rowland's (1999) focus on manipulation of artefacts, and Carl Craver's work, who initially applied it to extended theorising. Here is how Craver (2007) stated the idea of mutual manipulability:

> "a component is relevant to the behaviour of a mechanism as a whole when one can wiggle the behaviour of the whole by wiggling the behaviour of the component *and* one can wiggle the behaviour of the component by wiggling the behaviour as a whole. The two are related as part to whole and they are *mutually manipulable*." (p. 153).

This notion is present in Menary's conception; for example when remembering with the notebook, Menary (2006) proposes: "X is the manipulation of the notebook reciprocally coupled to Y – the brain processes – which together constitute Z, the process of remembering." (p. 334). Arguably, this mutual manipulability perspective was anticipated by Clark and Chalmers's (1998) focus on the two-way interactions of coupled systems; additionally, another precursor worth mentioning is Clark's (1997) focus on what he called 'continuous reciprocal causation,' namely, the idea that components in a cognitive-embodied system operate in continuous, mutual, looping cycles of bidirectional influence. Nonetheless, Menary (2006) further specified and fleshed out how this two-way interaction can be conceptualised.

### 1.5. Third-Wave Extended Cognition.

The notion of the a third wave was originally suggested and articulated by Sutton (2010):

> If there is to be a distinct wave of EM [Extended Mind], it might be a *deterritorialized cognitive* science which deals with the propagation of deformed and reformatted representations, and which dissolves individuals into peculiar loci of coordination and coalescence among multiple structured media

> … Without assuming distinct inner and outer realms of engrams and exograms, the natural and the artificial, each with its own proprietary characteristics, this third wave would analyze these boundaries as hard-won and fragile developmental and cultural achievements, always open to renegotiation. (p. 213, italics added).

Kirchhoff and Kiverstein (2019) draw from this four themes: i). the third wave opts for *dynamic singularities and no fixed properties* (the inner and outer have no unique, fixed properties; instead, the properties are dynamic and temporal); ii). *flexible and open-ended boundaries* (the boundaries are porous); iii). *distributed assembly* (cultural practices figure in cognition); and finally, iv). *diachronic constitution* (cognition is not synchronic; see 1.6. of this thesis). Kirchhoff and Kiverstein (2019) point out that there is a tension between the second wave's emphasis on cognitive transformation *and* the complementary role of biological and external elements. They see that cognitive transformation is more dynamic whereas complementarity is more fixed; however, this tension is not unreconcilable, and indeed it is this consideration with dynamism that partly motivates Kirchhoff and Kiverstein's (2019) third-wave theory where predictive processing, cultural practices (including complementary artefacts), and extended cognition are brought together.

Briefly, predictive processing theories stipulate that the brain continually makes unconscious, high-level predictions about the world; these predictions are then updated in light of sensory data. The "goal" is to minimise *prediction error*. When there is a mismatch, prediction errors ascend and these are either ignored or lead to an updating of the priors (sometimes by way of the organism actively changing the environment, as proposed by active inference versions of the account; see Clark 2015). Relatedly, *precision weighting* involves 'confidence' being placed on either high-level (top-down) or sensory (bottom-up) priors (depending on circumstance). Some theorists have applied an enactivist-oriented lens to predictive processing. For example, Gallagher and Allen (2018) refer to prediction as anticipation.[21]. But there are tensions between enactivism and predictive processing (see Thompson et al 2020)[22] and, more generally, with predictive processing accounts which aspire to account for the whole of cognition (Williams, 2020). In any case, some accounts of cognitive

[21] Gallagher (2017) is at pains to note that he is not merely a linguistic dispute; although my personal view is that it verges close to sounding like a verbal – rather than substantive – dispute.
[22] Free-energy principle, which has been connected with autopoiesis and enactivsm. For a critiques see Di Paolo et al. (2022).

extension have made use of predictive processing accounts (Gallagher, 2017; Kirchhoff & Kiverstein, 2019); and Clark (2015) views the precision weighting component as being minimally metacognitive, and with relevance for extended cognition. My view on this is that, *even if* Clark is correct in viewing precision weighting as a form of metacognition, it is nonetheless too narrow a conception of what is an otherwise richly embodied, embedded, and extended phenomenon. Metacognition should not be reduced in that way (see sections 4.5. in this thesis).

### 1.6. Objections to of the Extended Cognition Thesis

Clark (2019) describes the series of objections to and defences of cognitive extension as a 'cottage-industry' (see Clark, 2008; Menary, 2010; Shapiro, 2010). These arguments are often in terms of the parity-principle and 'trust and glue' conditions. Shapiro (2019) is right in calling the arguments 'dialectical.'[23] There have been numerous critiques of the extended cognition thesis; however, it is beyond the scope of this project to go into them. Sprevak (2019) has usefully highlighted some of the main problems associated with the thesis: first of all, issues with *cognitive bloat* (issued I touched upon in section  and footnote 12); second of all, the idea that *embedded cognition* offers a better explanation; thirdly, the coupling-constitution fallacy; and finally, *the mark of the cognitive* objection (see section 3.4 in this thesis).

The coupling-constitution objection is worth briefly looking at as it is arguably the most difficult of the charges (Gallagher, 2017). Adams and Aizawa (2001, 2008), have argued that the thesis falls prey to the *coupling-constitution fallacy*. This is the concern that an inference is made from the causal role of coupling to constitution. The idea is that while a resource maybe coupled with a person and play a causal role in cognitive activity, it nonetheless does not constitute cognition. Indeed, this line of argument presupposes that there is a distinct difference between causality and constitution, and that extended cognition cannot possess both at the same time. This being so, the critics suggest that an embedded view would be preferable (Adams & Aizawa, 2008; Rupert, 2009).

Kirchhoff (2015) argues that a diachronic view is preferable for extended cognition[24]; furthermore, if we think of constitution as being *temporal* rather than location-based, then this

[23] There is something useful in this as a pedagogic tool if one is unfamiliar with the extended mind / cognition thesis. I have had experience on at least one occasion where my interlocutor was initially doubtful about the prospects of extension; however, when I ran through the 'trust and glue' conditions, the *fine print* as it were, of availability, accessibility and so on, my interlocutor was more convinced.

[24] Kirchhoff and Kiverstein use the terms *distributed* and *extended* interchangeably.

makes extended cognition more defendable. To this end, Kirchhoff and Kiverstein (2022) attempt to neutralise Adams and Aizawa's charge by opting for an alternative account of constitution which they term *diachronic process constitution*. Kirchhoff (2015) considers two types of constitution are considered. First of all, synchronic constitution.[25] It is characterised by the following aspects:

i). Synchronic (atemporal).
ii). Asymmetric.
iii). Object-based.
iv). Non-causal.

To unpack this a bit, Kirchhoff uses as an example the Michelangelo's David statue (an example developed by Gibbard (1975) in analytic metaphysics). Firstly, by *synchronic* it is meant that there is a one-to-one, spatiotemporal coincidence between the parts and wholes. More specifically, these synchronic relations are atemporal. Time is only relevant here insofar as a 'snapshot' can be taken at a particular, punctuated time-slice. For example, David is the token statue, and a piece of marble is the token marble. At time *t*, David and the piece of marble spatially and materially coincide, and thus can be created (or destroyed) at the same time. As for the second characteristic, *asymmetry*, the key idea is that the piece of stone helps to constitute David the statue; however, David does not constitute the stone (if I were to break a piece off I would only have stone in my hand, not the statue). In this manner, constitution only works in one direction. Thirdly, constitution, on the standard synchronic view, is *object-based*. The idea is that the constitution is not reliant on dynamic unfolding, as we will see shortly; instead, the object endures. Fourthly, constitution, as standardly conceived in analytic metaphysics, is *non-causal* which is distinct from that which is non-causal relata (recall Adams and Aizawa's contention that causation and constitution are separate).

Kirchhoff (2015) does not deny that material constitution, as standardly developed and applied to material objects (chairs, tables, statues), is the correct approach; however, given that cognition is a process rather than a substance, it makes more sense to characterise constitution of cognition as *diachronic.* On this view, cognition is diachronic because it has symmetric relations, is process-oriented, and involves reciprocal causation in a way that is not distinct from constitution. The diachronic view is antithetical to the synchronic view as it involves

[25] Also known as *material constitution*.

dynamic properties. This dynamic view can be brought out in the example of frying oil in a pan (Kelso, 1995). When a pan is heated, there is a disparity between the (cooler, denser) oil at the top of the pan and the (hotter, less dense) oil at the bottom of the pan. Instability occurs when the temperature disparity is great enough for the oil to start 'rolling' in what is termed a series of 'convection rolls.' The denser, cooler oil falls and the less dense, hotter oil rises to the top of the pan; these rolling cycles continue, and the effect is that of a *self-organising system* that sustains itself. Clark (1997) observes that the actions of individual parts in the system cause the rolling, but the overall, emergent pattern also guides the individual parts such that the causation is 'circular' (p. 107) (also see, Thompson, 2007). These dynamic patterns, as seen in the case of convection rolls, are also what Kirchhoff (2015) has in mind when emphasising the diachronic rather than synchronic ontological status of cognitive constitution when considering extended cognition.

I will now put questions of extended cognition aside for a moment and turn to the second strand of the thesis: metacognition. In section 1.8 I will return to the extended cognition literature and examine extended cognition in connection to metacognition.

### 1.7. Metacognition

Metacognition is a concept present in cognitive psychology, animal cognition, psychopathology, developmental psychology, psychopathology, and educational psychology. The history of metacognition can be roughly dated to 1965 when Josef T. Hart inquired into metamemory, specifically how memory could be used reliably.

Proust (2013) observes that the terms *metamemory* and *metacognition* were posited in the 1970s as a result of John H. Flavell's (1979) work. In short, metamemory means 'knowledge and awareness of memory.' Metacognition refers to 'knowledge and cognition about cognitive phenomena.' Here cognitive phenomena is intended to cover what is standardly examined in the remit of cognitive science: attention, memory, problem-solving, social cognition, self-control and self-instruction. This dual manner of conceiving of metacognition - as metacognition and metamemory – can still be seen in the literature; however, there is ambiguity around the use of these terms because theorists use them in different ways.

Metacognition has been described as cognition about one's own cognition (Nelson and Narens, 1990), and sometimes, more informally, as 'thinking about thinking,' although, as Proust (2013) notes, this characterisation does not capture the full extent to which metacognition can operate, as 'thinking about thinking' is metarepresentational and more

intellectualist way of conceptualising the phenomenon. Proust points out that the standard, informal definition of metacognition as being tantamount to 'thinking about thinking' is problematic because "terms such as '*thinking*' and '*about*' tip the scale in favour of a particular [primarily conceptual, metarepresentational] theory of metacognition" (2013, p. 3, italics added). Indeed, Proust emphasises that there are at least two ways of understanding *meta*. Firstly, in terms of the theoretical knowledge *that*, *what* and *when* one has this knowledge. This definition is metarepresentational and is associated with the work of Carruthers (2009, 2011), Gopnik (1993) and Perner (1991), who have a conception of metacognition that is close to social cognition. Secondly, meta can be interpreted as the activity of monitoring and evaluating cognitive sub-systems; this definition of metacognition can be traced to the work of the Josef T. Hart, and it is also Proust's preferred view. On this account, the point of metacognition is "to evaluate one's present cognitive dispositions or outputs, endorse them, and form epistemic and conative [agentic] commitments." (p. 3). The evaluation can involve mere cues (that have the phenomenology of knowing and ability); but, importantly, the evaluations are not necessarily limited by these cues and feelings. Proust notes that these evaluations can include beliefs and knowledge.

On Proust's account procedural metacognition[26] is independent of social cognition (developmentally, functionally, and phylogenetically) and is a natural kind. Higher-level forms of metacognition involving language and concepts do occur; however, these are framed as analytic metacognition, and, although they interact with the procedural system, they are separate. One gloss one could put on it is that procedural metacognition amounts to system one and analytic metacognition to a system two if one is interested in using the language of dual-systems (see Evans & Frankish, 2009; Koriat & Levy-Sadot, 1999). Proust has the following overall definition of metacognition, which I quoted earlier in this thesis and I'll provide again, as it merits repetition:

> [Metacognition] is the set of capacities through which an operating subsystem is evaluated or represented by another subsystem in a context-sensitive way (p. 4).

[26] Vuorre and Metcalfe (2021) note, that in empirical work, *prospective* ratings tend to measure metacognition in terms of judgements of learning, feelings of knowing, and ease of learning. On the other hand, *retrospective* metacognitive judgements of accuracy use confidence as a proxy. Taken together, the relationship between these judgements and accuracy have variously been called *metacognitive resolution*, *sensitivity*, or *relative accuracy*. To complicate matters further, all of the above can be distinguished from overall accuracy (i.e., as averaged over multiple performances), sometimes called *calibration*, *bias*, *absolute accuracy*, or *over/underconfidence*. The much talked about Dunning-Kruger effect (Dunning-Kruger, 1999) is associated with calibration. (Also see Fleming & Lau, 2014).

This definition is the most neutral that can be offered as a way of covering the *attributivist* and the *evaluativist* views on metacognition. I will now look more closely at procedural metacognition.

### *1.7.1. Proust's Account of Procedural Metacognition*

Procedural metacognition[27] is a nonconceptual, non-misrepresentational form of second-order cognitive monitoring and control. It is regulated by epistemic feelings, the most foundational form of which is fluency, which is at once a functional element in how monitoring and control and a metacognitive norm (see chapter four of this thesis on normativity). Fluency is shared with non-human animals and pre-linguistic infants. Fluency is one of the ways in which perception and action can loop together with higher-norms (I will have more to say on this connection in the chapter three on the personal-sub-personal distinction). There are epistemic feelings and mental actions (see 4.1.1 of this thesis). Also worth noting is the fact that procedural metacognition can operate prospectively (i.e., predictively) and retrospectively.

To argue for procedural metacognition Proust (2013) draws on converging evidence from the comparative literature, developmental psychology, and neuroscience. (Also see: Proust (2019). Fleming (2021) has a more recent mini-review. For example, studies in opt-out paradigms and information-seeking behaviour) to argue that metacognition exists in primates.

### *1.7.2. Analytic metacognition and how it interacts with procedural metacognition.*

Unlike (nonconceptual) procedural metacognition, analytic metacognition is conceptual, linguistic and is best viewed as a cognitive and cultural achievement. It is connected to the language abilities of the agent and is enmeshed in conceptualisations. Indeed, *analytic metacognition* is closer to the metarepresentational definition of metacognition as linguistically and conceptually mediated, and with ideas connected to the phenomenon of explicitly 'thinking about thought.' Like procedural metacognition, analytic metacognition can operate prospectively (predictively) and retrospectively. Moreover, there is a thread that connects procedural and analytic metacognition. These are bound up with metacognitive norms such as *accuracy*, *consensus*, and *relevance* (see chapter four of this thesis). We can understand how

[27] Sometimes *implicit metacognition* is used in the literature instead of *procedural metacognition.* I will follow Proust in using the latter term.

analytic metacognition is still intricately bound to and entangled with procedural metacognition by considering the following passage:

> [A]nalytic metacognition, although its verdicts may contradict feel-based predictions, would not have been able to develop without an ability to intuitively appraise one's certainty in selecting premises or in forming perceptual and memorial decisions. Second, analytic metacognition would not be able to come to a final decision about one's certainty about a given proposal, if no metacognitive experience was available to stop the analytic regress to higher-order evaluations: *even in its analytic forms, feeling right puts an end to higher-order worries about normative appraisal*. (Proust, 2013, p. 70, italics added).

Proust (2014, p. 385) notes that they can coincide and also pull apart in the following three ways:

1. Procedural metacognition and the analytic metacognition *converge.*
2. Procedural metacognitive fluency is *rejected* on analytic metacognitive grounds (see Koriat & Levy-Sadot, 1999).
3. Procedural metacognitive disfluency is *accepted* on analytic metacognitive grounds.

Proust (2007) notes the difference between procedural and analytic metacognition can be drawn out when considering the possibilities of recursion. Procedural metacognition is able to work at the second-order level. I can *know* that I *know* that P. However, with third-order recursion and beyond, it is difficult to distinguish a second from a third-level (or higher) level of recursion. For example:

1. 'I *know* that I *know* that I *know* that P' is not easily, or at least not discernibly, different from merely 'I *know* that I *know* that P.' And even more so:
2. 'I *know* that I *know* that I *know* that I *know'* that P is indistinguishable from 'I *know* that I *know* that P.'

In essence, first-person recursion is limited if one goes beyond one layer of recursion (I know / believe / feel that I know / believe / feel); the levels, in a sense, collapse. Nonetheless, it is

language itself, and specifically linguistic vehicles with their recursive properties, that make it possible to go beyond second-order thoughts. For instance:

3. "I know that Anna knows that her father knows that her mother knows that her grandmother knows that she is invited for lunch on Sunday." (p. 277).

Although convoluted, example (3) demonstrates that linguistic tokening allows us to appreciate n-order recursion. Proust (2007) notes that the iterative element is embodied at once in the linguistic vehicle and in the way the information is arranged. A public symbol system such as a language is qualitatively different from procedural format of metacognition. A lot of the structure of these metarepresentations emerges from the properties of a natural language and can allow for more complicated revisions of beliefs (see Bermúdez, 2003); and it this linguistic version of metacognition that is captured by analytic metacognition

One important point to emphasise is that metacognition – at the procedural and analytic levels – overlaps with but is nonetheless different from the broader concept of *self-regulation*. Self-regulation involves operating on oneself or the environment, but it need not involve any second-order task monitoring and control. So, more precisely, metacognition involves taking one's own cognitive processing as the object. Again, it merits repetition that this is consistent with Nelson and Narens (1990) view where there is a meta-level and a control-level. The meta-level controls the object level (the cognitive subsystem)

It is additionally worth emphasising that the procedural metacognition is not necessarily unconscious and the analytic form of metacognition is not necessarily conscious. Rather, both procedural and analytic metacognition have unconscious and conscious aspects. The difference is in their representational formats: procedural metacognition has a non-conceptual, affect-based format whereas analytic metacognition has a conceptual format that requires possession of a language. Additionally, the degree of flexibility in these two types of metacognition varies; procedural metacognition has an inflexible structure whereas analytic metacognition is more flexible; however, as Proust (2013) observes, the inflexibility of procedural metacognition does not necessarily imply a lack of control (p. 299). Koriat (2000) notes that metacognitive feelings, associated with procedural metacognition, enjoy a special status within metacognition. On his view, there is a 'cross-over mode' in which the automatic-implicit mode of metacognition and a more explicit mode are brought together. The implicit mode shapes epistemic feelings, and these feelings, in turn, guide action. However, Proust (2013) states that consciousness may not be necessary for flexible control of thoughts, and flexible control is not restricted to a given

representational format (i.e., flexible control can work at an implicit, procedural level or an explicit level).

### *1.7.3. Metacognition and Social Cognition*

In the following section I will briefly examine some of the empirical evidence that has been appealed to when defining the boundaries of metacognition. While Proust (2013) has adopted a broad, inclusive definition of metacognition that does not entail metarepresentation, some theorists, such as Carruthers (2009), have been interested in the connection between metacognition and social cognition (also known in the literature as the ability to *mindread* or *mentalise*). As stated earlier, the thesis I am putting forward is consistent with Proust's (2013) approach, and so in the follow I will criticise the more restrictive view of metacognition that excludes non-conceptual and non-linguistic forms of cognition.

Gopnik (1993) and Carruthers (2009) have proposed that metacognition is mindreading turned inward. On this view metacognition and social cognition theories involve some form of metarepresentation[28] and recursive thinking. Similarly, Fleming (2021) proposes an account of 'second-order metacognition' as entailing the use of the same 'computational machinery' that is used in social cognition (p. 57). Indeed, this idea presupposes a symmetry between understanding of self and other. Ryle (1949), in an influential but controversial line, wrote: "The sorts of things I can find out about myself are the same as the sorts of things I can find out about other people, and the methods of finding them out are much the same." (pp. 155-6). Evidence for this claim has typically been drawn from research on autistic people where problems with self and other-directed propositional-attitude attribution lacks dissociation; in other words, problems with metarepresentation affect understanding of both self and other (Wilkinson, 2020).

Nevertheless, Fleming (2021) notes that there is conceptual space for – and empirical evidence to back up – procedural metacognition as a complement to higher, metarepresentational accounts (the latter of which overlap with mindreading). Fleming

[28] Note that social cognition does not necessarily involve metarepresentation or complicated inferences. Zawidzki (2013) (p. 36) notes that sophisticated mindreading is not needed. See Hutto (2008) and McGeer (2007) on narrativity and folk-psychology; and Gallagher (2017), Gallagher and Zahavi (2021) on phenomenological and 'direct perception' versions of intersubjectivity. Nevertheless, how far these non-metarepresentational accounts can go in terms of explaining social cognition is an ongoing question. Apperly (2011) notes that there is still a place for mindreading alongside social scripts, normative rules and narratives (p. 117); and implicit forms of mindreading need not depend on concept-possession and metarepresentation (as evidenced by pre-linguistic children and non-human animals who pass false-belief tests). (p. 109).

acknowledges that there are two systems, and that nonhuman animals are capable of procedural metacognition. Mindreading and metacognition operate with different processes but also interact and overlap. So, in effect, there are some metacognitive processes that are distinct from social cognition. Although, alternatively, Fleming does wonder whether the systems are "tightly intertwined in a virtuous cycle, with good metacognition facilitating better mindreading and vice versa." (p. 60). To this end, Fleming cites evidence that suggests mindreading ability at age four predicts later self-awareness regardless of language development. Additionally, there is evidence that mindreading – in this case, thinking about the feelings of someone else – interferes with task performance and confidence. Finally, eye-tracking studies have shown that children under the age of three are sensitive to false-beliefs.

There are ongoing disputes about the developmental sequence of and the relationship between metacognition and mindreading. Kim et al. (2020) note that many studies in the vast literature typically fail to disambiguate implicit (i.e., procedural) metacognition and mindreading from more explicit (analytic) metacognition and mindreading. They note that until their 2020 study, the only extant child study that has examined this - Bernhard et al. (2015) - found no connection between implicit metacognition and explicit mindreading in children aged three to five. However, Nicholson et al. (2019) conducted a study on mindreading and metacognition in autistic and neurotypical adult participants; the autistic participants had greater difficulty with the accurate explicit metacognitive judgements, but not implicit metacognition. Further, they found that explicit mindreading and explicit metacognition are related, but not explicit mindreading and implicit metacognition; nor, interestingly, were explicit and implicit metacognition connected. This evidence was taken as supporting a primarily mindreading-based, non-procedural account (a threat to the account of Proust, 2013); however, Kim et al. (2020) found in a study on four year old Japanese and German children that implicit and explicit metacognition are related, and, further, that implicit and explicit metacognition are not related to implicit and explicit mindreading. Thus, on the basis of Kim et al.'s evidence, mindreading is not an overarching framework for metacognition. In so far as *explicit* versions of metacognition share mechanisms with mindreading, it is probably because these metarepresentational systems are more oriented toward folk-psychology and thus amenable to the variability and diversity of cultural processes. While implicit (or procedural) metacognition has a more uniform format across cultures, and indeed across species, *explicit metacognition* is more of a social phenomenon, and is thus more likely to reflect differences in culture (Kim et al., 2018).

Regardless of how the empirical research continues to take shape, it is worth mentioning Wilkinson (2020) when he notes that that the Rylean idea of symmetry between self and other-understanding has been exaggerated. The choice is not between a totally transparent, introspectively immediate self-understanding nor a merely social-cognition-turned-inward, interpretivist stance. Indeed, self-understanding comes with a first-person degree of agency and experience that one's appreciation of the minds of others lacks. As Zahavi (2005, 2014) has pointed out, there is a *necessary asymmetry* between self and other. I will never be acquainted with the experience of another in the same way that I am acquainted with my own self-experience (Zahavi, 2005, 2014). In relation to metacognition, it can be stated that one's experience of metacognition – especially procedural metacognition – is inextricably tied to one's first-person experience.

The key point to emphasise here is that procedural and analytic metacognition are separable, yet interact. Furthermore, while there is conceptual space to examine the metarepresentational aspects of analytic metacognition, there are still basic forms of metacognition (i.e., procedural metacognition) that are independent of social cognition.

## 1.8. Literature Review: Cognitive Extension and Metacognition

I will first of examine literature on extended cognition and procedural metacognition; secondly, I will examine the literature in relation to more analytic (i.e., conceptual) versions of metacognition. Along the way I will draw out what I view as the relevant points of contact in these literatures. Bear in mind that procedural and analytic metacognition, when tightly bound, are not completely distinct; so I am merely organising this section in two-parts as a structuring device.

### 1.8.1. *Procedural metacognitive, action and extension*

Clark (2015) has attempted to reconcile the extended cognition thesis with his preferred active inference account of predictive processing.[29] On his view variable precision weighting between top-down modelling and bottom-up sensory evidence is "essentially meta-cognitive in nature" (p. 3768). While the specifics of active inference (and predictive processing) is a contentious

[29] Proust's preferred feed-forward model (see chapter 6, 2013) arguably has similarities with active inference; and more recently Gallagher and Allen (2018) have interpreted active inference in an enactivist-friendly manner; however, Clark (2015) states that active inference is less passive than traditional feed-forward models, and involves predicting sensory cues using high-level predictions before the sensory cues occur (p. 3766).

area (see Williams, 2020) what is of real interest is the way Clark (2015) sets up the problematic. Most extended cognition examples involve "fluid, unreflective use" of tools as one of the characteristics of their incorporation into our cognitive architecture (p. 3773); however, if the extended beliefs can qualify as knowledge there needs to be some awareness to scrutinise the tools, and if this involves conscious awareness then this conscious scrutiny alienates the agent from the environment. Thus the agent can have knowledge without extension, or extension without knowledge. Clark's solution is to argue that the precision weighting in active inference is "essentially metacognitive nature" and provides unreflective scrutiny that ensures that knowledge *can* extend (p. 3768).[30] Andrada (2021) has criticised Clark's approach because it presupposes a stark opposition between awareness, which Clark equated with alienating scrutiny, and lack of awareness, that Clark equates with a form of automatism. Andrada argues that Clark's concern that awareness is misplaced, and is somewhat orthogonal to cognition. On this point, I agree with Andrada that awareness does not alienate a person from cognitive tools and epistemic practices; nonetheless, I also agree with Clark (2015) that metacognition can serve the role of binding an agent to external tools. Clark is essentially correct when he invokes the importance of metacognition, although Clark's version of metacognition can be strengthened if one invokes Proust's (2013) account of metacognition instead of Clark's (2016) predictive processing model of active inference. In effect, Clark (2015) nicely draws out the conceptual balancing-act that needs to be achieved when theorising about extended cognition and metacognition: if the process of belief formation is too reliable, how can we say that cognitive extension is still taking place? Conversely, if the process is too effortful and deliberative, does this disrupt cognitive extension? Again, while Andrada (2021) has invoked a minimal form of agential self-awareness as a way out of the dilemma, I propose that focus on a richer, more robust account of metacognition than the predictive-processing account Clark offers is a suitable candidate for answering the problematic Clark sets up. If one follows Koriat (2000) and Proust (2013) in offering an account that stresses the procedural, experiential aspects of metacognition, then in fact it is this very effortful, procedural metacognitive control that can integrate the coupling of tool and agent. Procedural metacognition is by its nature cognitively effortful and activity-dependent; and even when analytic-metacognition is being exercised, the engaged, activity-dependent characteristic

[30] Wheeler (2019) observes that Clark may be making unnecessary demands on the role of tools; specifically, he Wheeler asks why the world should present itself transparently when internal cognition does not necessarily involve transparency.

involved in procedural metacognition are still present (Koriat, 1993). (See section 4.1.1. in this thesis on cognitive action, and chapter five).

Kirsh (2004) has made a case for how interactions with environments structure and facilitate metacognition in such a way that sometimes agent-environment boundaries blur. Kirsh puts forward five key tenets:

1. Spatial design reduces the complexity of cueing an agent as to what they should do next, and this in turns assists with the scheduling and performance of tasks. Kirsh cites the example of a short-order cook whose ingredients and kitchen workplace are designed so that the environment is cued for fast, efficient action that reduces cognitive load. Kirsh notes that such environments are cued; the cues, tied with evaluating what to do next, are built into the situation so that abstract matters of planning and evaluating become concrete. This is consonant with Sterelny's (2003) case for how humans epistemically engineer their environments; humans (and non-human animals) don't just adapt and react to their environments, they modify them and create epistemic niches to suit their needs.
2. Kirsh's second tenet is the way that culturally-embedded practices can be exploited metacognitively when in organised environments.
3. Kirsh notes that these cues can be cognitively extended, such that the manipulation of these tools can constitute cognition.
4. Metacognitive extension works at different temporal levels (ranging from milliseconds to minutes).
5. Kirsh emphasises the coordinating role the agent plays in dynamic action whereby all the moving parts across body and world are synchronised varying temporal frequencies. Indeed, this tenet anticipates Sutton's (2010) line that an agent can be viewed as a 'loci of coordination' (p. 213). For Kirsh (2004), metacognition plays a role in this coordination.

Kirsh worries that if we conceive of metacognition as reacting to environmental cues, then is it really metacognition (i.e., second-order cognition) or is it first-order cognition? Indeed, Kirsh's account appeals to many first-order cases where the agent is responding to the moment-by-moment unfolding of available environmental cues, and explicitly. The world on Kirsh's account is arguably doing *too much* work and thus making it all *too easy* for the agent. I believe this is a real problem for Kirsh's account, and that Kirsh's response to this problem is generally

unsatisfactory. Kirsh suggests that we adopt a revisionary account where the research focus is on metacognition is less on planning, monitoring and correcting, and more oriented toward how environments cue and prompt metacognition. But, again, this leaves open whether Kirsh's account captures phenomena that is a first-order (*merely* cognitive) or second-order (metacognitive) in nature. Put differently, metacognition involves one cognitive subsystem evaluating or representing another cognitive subsystem (Proust, 2013). This means that metacognition is not merely responding to something in the environment; rather, metacognition involves responding to one's responses to the environment. As Proust (2007) points out, metacognition is not just world-directed, first-order activity, but involves motivational and informational needs of the agent. Returning again to the problematic Clark sets up in his 2015, it is clear that integration requires a balance between the coupling with artefacts being neither too easy nor too deliberative. (I have more to say on this in section 4.1.1. on cognitive action).

In Clark (2008) observes that metacognition involves fine-tuning between motoric and perceptual processes. "The effect of extended problem-solving practice may often be to install a kind of motor-informational tuning such that repeated calls to epistemic actions become built into the very heart of many of our daily cognitive routines." (p. 75). As an example, Clark cites the way in which people automatically saccade their eyes when confronted with a flash. This, Clark writes, "embodies a kind of hard-wired implicit metacognitive commitment to the effect that we may gain useful, perhaps lifesaving, information by such a rapid saccade." (p. 74-75). While Clark's emphasis on motor skills and epistemic actions is suggestive, it is arguably not a good example of metacognition. It appears to be first-order, reactive form of behaviour rather than anything genuinely second-order in nature. Indeed, as with Clark's (2015) ascription of metacognitive status to aspects of active inference predictive processing, it should be noted that if one's description of metacognition is too minimal, then the risk is that metacognition will be ubiquitous in theoretically unsatisfying ways.[31]

Nonetheless, while extending metacognition needs to involve more than first order character, this does not mean the discussion ends there. Carter et al. (2018b), writing in regard to the 'trust and glue' conditions (see section 1.4 of this thesis), suggest that *fluency is aligned with accessibility*. Relating this to Proust's account of procedural metacognition, which has fluency as the most foundational norm, one can see a point of contact between fluency and

[31] Langland-Hassan (2014) has made similar criticism toward Proust's (2013) definition; namely, the concern that procedural metacognition is too close to first-order cognition. See Goupil and Proust (2022) for a recent discussion on how procedural metacognition is distinct from merely sensorimotor control, but instead evaluates informational states associated with cognitive actions such as remembering and deciding.

procedural metacognition. Now of course there are problems with the 'trust and glue' conditions. Recall that they tend to be aligned with more folk-psychological accounts of extended belief (Pöyhönen, 2014); also, by Clark's (2008) own admission they act as a sort of 'rough and ready' heuristic. Nonetheless, given the centrality of the *accessibility* condition of a cognitive artefact for extended cognition, we find the concept appearing in different ways throughout the literature of metacognition in weaker forms such as coupling (2008), or strong forms such as mutual manipulability (Menary, 2007). Carter et al. (2018b) do not provide a description of their preferred account of fluency; however, Proust's (2013) account is, I believe, a good fit. Its value is further strengthened by Proust's proposal that fluency is also normative.

Arango-Muñoz (2013) proposes that epistemic feelings can guide agents to make a decision as to whether memory tasks should be solved internally or externally. Risko and Gilbert (2016), incorporating some of Arango-Muñoz's insights, review the link between cognitive offloading and metacognition, and suggest that an extended view could provide a useful framework for understanding the triggers (e.g., memory overload) and behavioural consequences of cognitive offloading; more precisely, they detail when, how and why cognitive offloading is likely to occur (especially with prospective memory) when we are not confident about our biological ability to remember. Moreover, they note that a sense of fluency could influence choice of metacognitive strategy. On this point, I propose one could go further by integrating other metacognitive norms with cognitive offloading. Indeed, there is a gap in the literature connecting metacognitive norms and the extended cognition hypothesis (see chapter four of this thesis). Risko and Gilbert make no specific commitments on how an extended view should proceed and leave open the question of how much conscious deliberation is involved. Moreover, some of the findings on offloading they appeal to could be characterised as embedded rather than extended; offloading by itself does not entail the hypothesis of extended cognition without further argumentation.

Proust (2013, chapter 13) cites evidence that children can understand utterances gesturally and verbally, before metarepresentation develops, and it is more cognitively efficient if there is no inference required. Proust views fluency as importantly related to the need for low effort gestures that assist with the production and understanding of communicative intentions. This gestural view is consistent with Clark's (2008) examination of cognitive gestures sometimes serving as the vehicles of cognition, and is consistent with Goldin-Meadow's (2003) extensive research on gesture and language. Proust argues that conversational metacognition (i.e., the metacognition that unfolds between people in interaction) might actually *constitute* a type of procedural self-knowledge "designed to publicly control and monitor conversation moment by

moment" (Proust, 2013. p. 269). While this is more a form of embodied cognition rather than extended cognition, it does indicate that the procedural vehicles of metacognition are not strictly brain-bound.

Finally, skilled-based accounts of control overlap with the literature covered here. For example, Christensen et al. (2016), Logan (1985), Sutton et al. (2011), and Toner et al. (2022). While these accounts are not strictly extended accounts per se, they nonetheless involve embodiment, action, and what may be deemed less intellectualist accounts of skilled control and monitoring. Christensen et al. (2015) observe that there is a dearth of studies on metacognition and agency; more specifically, "metacognitive accounts … neither explicitly address skill nor provide clear expectations for the effect of skill improvement on the SoA [sense of agency.]" (p. 341). Also, while Toner et al. (2022), in their account of the conscious control of athletic performance, do make reference to metacognition, they typically advert to more intellectualist accounts such as Flavell's (1979) account rather than that of Proust (2013 or Koriat (2000).

### 1.8.2. *Analytic Metacognition and extension*

Clark has at times opted for a Vygotsky-inspired view on metacognition that he has labelled *second-order cognitive dynamics* (1998); this is concerned with the 'distinctively human capacity' of using language to "think about thinking" (See Clark, 2008, p. 58). This way of looking at metacognition has more in common with Carruther's (2011) more restrictive, metarepresentational conception; however, it can be accommodated within what Proust (2013) calls analytic metacognition. Clark (1998) has emphasised the role of language as a cognition-enhancing tool. Language is an enhancer insofar as 1). It results in *memory augmentation* 2). involves *environmental simplification* 3). *coordinates*. 4) it *transcends path-dependent learning*. 5). has *control loops*, and, 6). allows for *data-manipulation*. This focus on language reflects a cognitive achievement, and, while language undoubtedly allows for particular forms of propositional and recursive thought, it would be unfortunate if metacognition were restricted to this linguistic format.

Proust (2013, chapter 9), has examined the relationship between content externalism and metacognition. More specifically, even when one's metacognitive capacities are well-tuned, problems can be incurred when the quality of feedback – the very feedback that is being fed into the metacognitive sys8tem – can be detrimental. This chapter is concerned with content, and insofar as it relates to this thesis, is more oriented with the content externalism of Burge

(1979, 1986) and Putnam (1975); nonetheless, it does point to the importance of the informational and epistemic ecosystems in which we live. Elsewhere, Proust (2014) has written on the link between extended cognition (and cognitive integration) and extended knowledge; although her account is primarily focussed on how we should credit epistemic achievements to people who make use of external devices. The conclusions Proust (2014) draws are relatively conservative in an attempt to ward off what could be viewed as extreme, revisionist theories of knowledge. Additionally, there has been a focus in the epistemology literature on whether extended cognitive processes can count as knowledge (Carter et al. 2018a). The definition of extended knowledge at play here is more technical (i.e., it is of more interest to epistemologists concerned with justified true belief and how knowledge might be constituted so as to account for luck conditions, etc.; see Carter et al. (2018a). Arguably, many philosophers of mind, cognitive scientists and psychologists will be operating with a less epistemically demanding form of knowledge and beliefs. In short, *extended epistemology*, while an important area of research, is somewhat orthogonal to the concerns of (extended) cognitive psychology in particular, and (extended) cognitive science more generally.

### 1.9. Conclusion

In this chapter I introduced the extended cognition thesis – with a focus on how this thesis is nested with a broader range of 4-E cognitive theories - and I also introduced procedural and analytic metacognition. This thesis is concerned with how extended cognition can be connected with procedural and analytic metacognition; to this end, it is worth briefly refocusing on the three central claims with which I began this thesis. Claim one stated that procedural metacognition and extended cognition are both subpersonal-level explanatory phenomena. This claim is the primary focus of chapter three. Claim two states that fluency (at the procedural level) and analytic norms are mutually reinforcing in cognitive action. This is the primary focus of chapter four. Finally, claim three states that metacognition can extend; more specifically, in chapter five, I will be looking at the subpersonal-level capacities of control and automaticity in regard to skilled coupling with an artefact when engaged in a metacognitive task. For now, the thesis turns to considerations of analogical content and non-conceptual thought in chapter two. The considerations and arguments in chapter two can be viewed as supporting the later chapters.

## 2. Analog Representations, Nonconceptual Content, and Thought

The focus of this chapter is mainly on how analog representations relate to an extended cognition account of metacognition. Analog representations offer a nonconceptual cognitive format that is appropriate for both extended cognition and metacognition. To begin this chapter I will critique the language of thought hypothesis in section 2.1, and, especially, the idea that is exists at an autonomous level as Fodor (1975) theorised. Criticising Fodor's (1975) view thus provides context for understanding the contrasting, non-conceptual account of analog representations. I will reinforce this contrast by examining analog representations in section 2.2. This section builds on Proust's (2013) account in which a nonconceptual cognitive format underlies procedural cognition; however, I take this account further by considering how analog representations offer a suitable form of representation for extended cognition. In section 2.3 bolster the analog representation focus of this chapter when looking at non-conceptual content in relation to procedural metacognition and extended cognition. In consequence, these discussions provide evidence for and background to claim one of this thesis: the claim that procedural metacognition and extended cognition are both subpersonal-level explanations that are undergirded by analog representations (a claim that will be focused on more explicitly in chapter three). Finally, in section 2.4, I argue that – contrary to Proust's (2015) claims – that cognitive penetration *can* occur in the context of epistemic feelings and procedural metacognition; additionally, I consider the cognitive role of language. Taken together, cognitive penetration and language serve as two ways in which procedural and analytic metacognition bind together; in making a case for this binding, I thus begin to provide evidence for claim two of this thesis: the claim that fluency and analytic norms mutually reinforce each other in cognitive action. Claim two will be further examined in chapter four, when I explicitly examine metacognitive normativity, and chapter five, when I examine metacognition in relation to automaticity and control. For now, the thesis turns to the representational format that supports metacognition and extended cognition.

### 2.1. Critique of Language of Thought

Williams and Colling (2018), in their critique of traditional cognitive science's idea that representations are digital, trace the origins of the this traditional conception to three pivotal ideas. First of all, in the late-nineteenth and early years of the twentieth century advances in formal logic by George Boole, Gottlob Frege, and Bertrand Russell showed how semantic

*relations* can be imitated, within limits, by syntactic operations over symbols, and that such operations are insensitive to the symbols over which they operate. Secondly, the contributions of Turing, and later McCulloch and Pitts (1943), showed how simple mechanisms can function in a syntax-sensitive manner. Thirdly, Fodor, building on the generative grammar of Chomsky's linguistic theories, put forward the influential idea that cognition has a systematic and productive propositional basis that involves composition of syntax and semantics . The emerging consensus here was that of a view of cognition as a digital computer running on the 'hardware' of the brain. It is to this view that the thesis turns.

Fodor's (1975) *language-of-thought* hypothesis[32] is a linguaform-representation view of cognition. Williams (2018) nicely captures its central idea when he notes that Fodor supposed that representation consisted in: "word-sized concepts, sentence-sized intentional states and argument-sized inferences" (p. 153). The language of thought hypothesis builds on the idea that thoughts are, like a natural language, compositional: it consists of meaningful constituents that are composed according to a syntax by which these meaningful constituents are arranged. Thought is systematic and productive; so it is able to generate an infinity of novel thoughts using a finite series of systematic rules. Fodor's proposal was that humans comprehend 'inferential relations' between thoughts either in a syntactic manner by way of "rule-governed transitions" or semantically.

Bermúdez (2003) observes that the language of thought hypothesis is epistemologically problematic in terms of how the semantic *content* of these thoughts is determined (Crane, 1990). Relatedly, it is not possible to discover these 'sentences' in the head. Proponents of the hypothesis do not expect that the sentences will be literally found in the head either; rather they suggest that the sentences are at a higher level of abstraction than the neural 'hardware' in which they are implemented (Fodor, 1975). This points to a conflation, observed by Devitt (1990), whereby there is ambiguity around the concept of syntactic properties: On the one hand, the syntactic properties could be referring to the *morphology* of a letter or word; but on the other hand, syntactic properties could be viewed as referring to the *functional* roles the words play. Of course the language of thought hypothesis is concerned with the latter, functional, interpretation (as noted, these sentences are, of course, not going to be literally found in the brain).[33]

[32] The language of thought hypothesis is an inversion of Frege's view on language, but Bermúdez observes that Frege was not concerned with psychology of thought; instead, Frege was interested in the logical form of thought. Whereas Frege was concerned with public language, Fodor was concerned with an internal language of thought. (Bermúdez, 2003).

[33] It would be a rather startling discovery if this were so, to put it mildly.

Bermúdez (2003) agrees with Crane (1990) that there are two possible ways in which the physical structure of the language of thought *could* be discovered. First of all, by way of their semantic features; or, secondly, by way of their syntactic features. The semantic option, however, is a non-starter as it presupposes what is meant to be explained – namely that we *can* discover the content of the language of thought sentences. The second option does not fare much better as we need to know the semantic features to discover the syntactic features. Bermúdez notes that while it is possible to separate syntax from semantics in a natural language (e.g., English), this is only possible because a comprehension of syntax and semantics is already presupposed. He uses the example of someone trying to read a foreign language of which they have no understanding and no means to decipher the words into the natural language they know. In such a case they would have no way of understanding the syntax without knowing the semantics.[34] But then the question is whether identifying these sentences is similar to the constraints entailed when a person attempts to understand a natural language sentence. To this end, Bermúdez (2003) and Crane (1990) note that there are at least two constraints:

1. A constraint is needed for differentiating between *causally relevant inputs* and *noisy inputs*.
2. A constraint is needed for identifying causal properties that derive from the physical instantiation/realization of a sentence (this is concerned with implementation) as distinct from their functional role.

Meeting these two constraints is difficult. Distinguishing between causally-relevant input and noisy inputs requires semantics to fix the causal role. The second constraint is also problematic: we need to distinguish sentences that arise from the *physical implementation* of a sentence from its *functional role* in a sentence. Again, this is a difficult task as the relevance of these properties cannot simply be 'read off' a brain-scan by way of their physical characteristics; the sentences do not have 'brute' identifiers (e.g. morphology) that can demonstrate a boundary between causally functional sentences and noise. The other option of examining higher-level functions is also a problem because this presupposes a semantic understanding. Thus, the focus needs to

[34] Specifically, as Bermúdez notes, "One very basic reason for this is that the syntax of a natural language involves grammatical categories. It involves certain expressions being substantives, others verbs, and others being adjectives or adverbs. One cannot decide what category a given word falls into without speculating about its semantics." (p. 29). Clearly semantic properties are in some sense presupposed for identification of syntactic properties to occur.

be on functional properties (of how the animal represents, and thus that, by necessity, entails semantic considerations.)

Bermúdez goes on to explain that levels of explanation can occur at multiple levels, and this question is pitched as a matter of epistemology, namely how we can find the correct level analysis with which to identify *language of thought* sentences? How would a researcher go about finding the neural correlate? Bermúdez draws on the example of the Grandmother Cell – a neuron that is hypothesised to selectively recognise a specific person[35] - as an example of the complexities involved in terms of interpreting what the cell's postulated role means. As Churchland and Sejnowski (1992) show, and as Bermúdez notes, the existence of such a grandmother cell could be viewed as consistent with at least two different theories. I). A local theory involving a 'dedicated' neuron; or, alternatively, II). a vector of representation (the latter option is more plausible as the vector, or distributed representations, is viewed as having more computational power). So, when one opts for the latter, plausible view that cognition is distributed[36], then the question is: *how are particular tasks performed*? To answer this one must begin in the world; indeed one needs to start by "working backward from the particular tasks being performed" (p. 30). Put differently, the investigation starts externally, at the semantic level, and often outside the head. This is of course an epistemic matter with important implications for cognitive science. (Bermúdez, 2003).

Indeed, Boone and Piccinini (2016) observe that cognitive science often involves 'looking up:'

> Understanding the capacities of a system often requires looking "up" to situate the system within some higher-level mechanism or environmental context as much as looking "down" to understand how those capacities are implemented by the lower level components, their capacities and organisation. (p. 1520).

Nonetheless, while Bermúdez (2003) points to some of the epistemic issues around the trouble with identifying a language of thought, there are also questions around whether there even is a language of thought in the brain. Boone and Piccinini go further in making a case for why -

[35] See Barwich (2019) on the history of the grandmother cell, its problems, and how a distributed view is preferrable.

[36] Here, distribution refers to cognitive-subsystems, and 'representations' in the brain (as can be seen in the work of Dennett, 1991; Hutchins, 1995b, 2014; Sutton, 1998). Although of course extended/distributed cognition take this further by viewing these subsystems as also, in select circumstances, existing outside of the body's boundaries.

contrary to Fodor's (1975) view - cognitive science is *not* autonomous from neuroscience. They note that the cognitive revolution in the mid twentieth century not only took inspiration from advances in computing, but rather a specific form of computation: digital computation; this in turn led to specific ideas about the supposedly autonomous status of cognitive science. The study of cognitive science was viewed as dealing with digital representations (e.g., a Fodorian language of thought), independent of implementation and acting autonomously. To this end, there are two main responses against the two-level picture: *elimination* and *reduction*. Boone and Piccinini (2016) find both responses insufficient. First of all, *elimination*, as argued by Churchland (1981), poses the idea that concepts from psychology and cognitive science, such as mental representation, can be discarded in favour of neuroscientific findings; this mirrors the way in which old scientific theories involving concepts such ether or phlogiston were replaced by more superior concepts. The idea is that neuroscience can supplant cognitive scientific concepts. Secondly, and relatedly, *reduction* is concerned with the idea that psychological theories should be reduced to neuroscience. Boone and Piccinini (2016) note that the models for this approach are drawn from physics (such as the reduction of Newton's theory of gravitation to Einstein's theory of general relativity). But, this sort of reduction is not appropriate for cognitive neuroscience as psychology and neuroscience do not have the requisite mathematical formalisms. Additionally, the central problem with *reduction* is that too much is left out. Studying molecular events can be useful in some contexts, but molecular events lead to neural events, and, more generally, these neural events lead to circuit and network and eventually systems level events, and all of this leads to body-environment coupling where behaviour is produced. (p. 1514).

So, the question is, if one wants to avoid the 'language of thought' – and yet also avoid the extreme and problematic strategies of elimination or reduction – then what is an alternative position? How can one theorise about cognition in a way that examines at once how cognition is *implemented* and yet also takes seriously the various levels – including many that feature the embodied agent and world itself – such that cognition is not isolated in a strange, abstract realm, detached from the world? This is where analogue cognition presents itself.

### 2.2. Analog Cognition

Opie and O'Brien (2015) observe that analog representation has a long history that ranges from Aristotle, the Scholastics, Descartes, and the British Empiricists, through to Craik (1943), Johnson-Laird, (1983), and in more recent times, Cummins (1996), Churchland (2012), and

O'Brien and Opie (2004) have proposed analog accounts. Nonetheless, the history of analog representations is also an overlooked history. Opie and O'Brien wonder why this is so, and propose at least three possible explanations. First of all, the influence of propositional-attitude psychology has had a strong effect. On this view, cognitive vehicles are deemed to possess propositional contents (as we saw with Fodorian language of thought in the previous section). Secondly, some theorists think that whether content is propositional or not makes no difference to the project of naturalising content (Hutto and Satne, 2015). And, thirdly, the concern is that non-propositional representations are too indeterminate. More specifically, only propositions can 'anchor' propositional references (particulars) and truth-conditions (states of affairs) in a way that non-propositional content do (because non-propositional content only has accuracy and veridicality conditions). However, these concerns are not important if one rejects the classic picture of cognitive science as involving linguiform content. As O'Brien and Opie point out, this problem is dissolved if the theorist assumes that cognition – and not just perception or motor abilities – involve analog content. This analog cognition can take the form of maps, graphs, and diagrams. It differs from symbolic cognition is at least two ways: i). contents are determined by way of structural properties of the vehicles. For example, the brightness of an x-ray image. And, ii). Although complex and highly structured, analog vehicles are not suited to propositional contents.

It is worth pointing out that analog representations do not necessarily mean discrete. For example, an hourglass or clock-hands are arguably discrete, yet can also be analog (Maley, 2011, 2018). So what is an analog representation? Beck (2018) note that being analog amounts to more than representation being continuous; rather, it is analog because it is structurally isomorphic to what it represents (Cummins, 1996; Elzinga, 2021; also see O'Brien and Opie, 2015).[37] In relation to procedural metacognition, the sense of fluency operates over the non-propositional, analog processes; however, Proust (2013) observes that the agent, when in possession of the appropriate conceptual resources, can 'replace' this epistemic evaluation with a propositional-attitude description (p. 301). The analog approach is also suitable grounding for extended cognition; specifically, the idea that external artefacts and processes, regardless of their heterogeneity, represent what they are just as 'inner' cognition does. Menary (2015) proposes that the 'ancient' cognitive system is analog and approximate, but a public

[37] This has much in common with cognitive maps. Tolman (1948), in his classic study on rats, was one of the first to empirically test the idea that the brain has a map that it uses to navigate. There are various ways of conceptualising the idea of a cognitive map (see Bermúdez, 1998; Gładziejewski & Miłkowski, 2017; Rescorla, 2009). Elzinga (2021) suggests that *know-how* appears to require a cognitive map (p. 1749).

symbol system – in Menary's case, he uses the example of mathematics, but this can include any language – is discrete (p. 11).

### *2.2.1. Neural analogue cognition and 'representation.'*

Typically, neural spikes, with their all-or-nothing character, have been taken as supporting a digital account of cognition. This view of digital cognition is favoured by traditional cognitive science (see Boone and Piccinini, 2016; McCulloch and Pitts, 1943). Maley (2018) illustrates this classic conception with the vivid example of a smoke detector alarm. The alarm makes the same noise and pitch regardless of the level of smoke. The alarm has an on/off threshold; in other words, it's a binary signal. If the neuron receives sufficient input (if the voltage reaches the required threshold), it will fire (there will be an action potential where a charge propagates down the axon). Nonetheless, there are problems with this standard account.

First of all, even on the assumption that neurons fire in all-or-nothing manner, the timing and frequency of these spikes play a crucial role. This is unlike a computer where the binary code (0 or 1) is discrete and fixed.[38] With a computer, the precise timing doesn't matter. But with neurons the timing between action potentials - and the number of action potentials in a given time-frame - has functional significance. But there is additionally a third problem that Maley (2018) draws out. While the received wisdom is that the voltage level of the pre-synaptic neuron does not have an effect on the post-synaptic neuron, this is not entirely true. In some neurons, the pre-synaptic voltage level *before* the threshold makes a difference to the post-synaptic neuron. The closer the original neuron is to the threshold, the more of an effect there is on the post-synaptic neuron. Conversely, the larger the spike, the smaller the effect.

### *2.2.2. A brief note on Representation*

Metacognition and extended cognition sometimes use the terminology 'cognitive representation'. *Representation* is a contested notion in cognitive science. Nonetheless, there is a history of 4-E cognition rejecting theories of representation. For example, Brooks (1991) theorised that robotics does not need to entail representation. Meanwhile, Chemero (2009) endorses an approach based on dynamical systems theory that doesn't appeal to mental representations. Thompson (2007) notes that representation can be conceived of in terms of *re-*

[38] In addition to this, there are non-neuron cells such as glia, and gaseous neuromodulators such as NO2, that play non-digital roles in the brain (Stephen Hill, personal communication, July 2022).

presentation; this is a presentational account that makes use of perception-motor abilities without recourse to cognitivist concepts of mental representation (Gallagher, 2017; Hutto and Myin, 2017; Thompson et al 1991).

On the other hand, some of the 4-E cognition theories have presented representational accounts (Clark, 1997, 2008; Hutchins, 1995b; Rowlands, 2006; Sutton, 2010; Wheeler, 2005). I will not be adjudicating in this thesis on the relative merits and weaknesses of these accounts in any detail as this goes beyond the scope of this thesis. Sutton (2015), for instance, maintains that representation is orthogonal to extension.

One overarching problem is that there is no agreed on definition among theorists as to how representations should be defined (Ramsey, 2007; Rowlands, 2017). Drayson (2018) highlights how metaphysical, psychological, and epistemic representation[39] can come apart. Finally, Williams (2018), notes that anti-representationalist theorists are typically against specific forms of representation (for example, the idea that there is a richly reproduced representation in one's head). In any case, for the purposes of this thesis, the key point is that the mechanisms are non-propositional, imagistic and make use of perceptual-motor systems (Baggio, 2021; Thomas, 1999). Whether or not these systems licence the word representation is beyond the scope of this thesis, and any use I make of the word representation should be recognised as provisional.

### *2.2.3. Vehicle externalism in extended cognition and metacognition*

Vehicles are often appealed in both the extended and the metacognition theorising. Indeed both metacognition and cognitive extension are deemed to involve vehicles. One way of looking at mental representation is through *vehicles* and *contents*. (Dennett, 1991; Hurley, 1998). This approach has been endorsed as fitting into an extended cognition framework (see Clark 2005). The idea is that the cognitive *vehicle* carries the *content* of a thought or perception. If one endorses an extended cognition account, this can take place outside of the brain (hence the name active externalism or vehicle externalism, as extended theories have alternatively been framed). This approach is mainly relevant in connection with the first-wave of extended

[39] Rorty (1979), in his *Philosophy and the Mirror of Nature,* is strongly against epistemic representations, yet makes space for psychological representation. "Fodor's picture [popular at the time] of the mind as a system of inner representations has nothing to do with the image of the Mirror of Nature I have been criticizing." (p. 246).

accounts (i.e., functionalist accounts, although the vehicle conception can be quite widespread and makes appearances throughout the literature).[40]

Now, given that that the vehicle-content distinction is in some respects a mainstay of cognitivism, should there be reason for concern? The idea, central to extended cognition, is that cognition can be realised outside of the head. This invocation of vehicles and contents is potentially problematic as it would appear to import classic *cognitivist* assumptions into this account of extended metacognition. Shea (2018) characterises vehicles as "individuable physical particulars that bear contents and whose causal interactions explain behaviour … realism about vehicles is a core part of mental representation." (p. 15). But Shea does differentiate vehicles from syntactic types. Similarly, O'Brien and Opie (2015) note that vehicles are different from symbols (which contain a combinatorial syntax) in two ways. Firstly, vehicles are determined by the local, structural properties of the vehicles; for example, an x-ray's brightness is structurally, analogously represents the varying gradients of bone or tissue density (also see Maley, 2011, 2021). Secondly, analog representations are not appropriately structured to represent propositional content.

Proust (2013) observes that information the cognitive vehicle carries is associated with procedural metacognition, whereas metarepresentation is more concerned with content (p. 57). Similarly, Koriat (2007 writes: "feelings of knowing rely on contentless mnemonic cues that pertain to the quality of processing, in particular, the fluency with which information is encoded and retrieved." (p. 19-29). This focus on vehicles overlaps with the extended cognition literature (see Clark, 2005, 2008). Indeed, it is the focus on vehicles in the extended cognition literature that differentiates extended cognition from content externalism (of the Putnam and Burge (1979, 1986) variety).

I argue that there is only reason for concern if we interpret vehicles in the stricter sense that Fodor (1975) proposed (i.e., as being associated with the language-of-thought hypothesis). On a more traditional, cognitivist view, vehicles can be syntactically typed. For example, 'syntactic type' refers to how the content is individuated so that each time the content is used, it will be the same. For example, the vehicle carrying the word BARN, with its intrinsic structural properties, is distinct from its syntactic typing (which, in Swedish means child and in English, of course, refers to a farm building) (Shea, 2018, p. 39). Moreover, on Shea's view external factors can fix the content; and the same vehicle can be used for different syntactic types.

[40] Noë (2004) appears to have no problem with the distinction - although it is worth noting that Noë's conception of vehicles is broad, and he makes liberal use of the word vehicle. This doesn't commit him to a language-of-thought view of representational vehicles, or indeed a representational view at all. (p. 221).

Nevertheless, this syntactic view of vehicles is different from the view of analog vehicles I am considering in this thesis. If, alternatively, we just think of vehicles as being part of the realisers of cognition, then this is a much less demanding criterion. O'Brien and Opie (2015)

Thompson (2007) notes that if one intends to continue making use of representation, then:

> Representational "vehicles" (the structures and processes that embody meaning) are temporally extended patterns of activity that can crisscross the brain-body-world boundaries, and the meanings or contents they embody are brought forth or enacted in the context of the system's structural coupling with its environment. (p. 59)

The take-home message is that we can interpret vehicles in a minimal way that is compatible with extended cognition and metacognition, but does not make commitments to Fodorian syntactic vehicles in a language of thought. There are analog representations that serve this role in a way that is more coherent with the proposals in this thesis.

### *2.2.4. Control systems*

Koriat (2007) observes that a focus on metacognition comes out of a rejection of stimulus-response behaviourist view. A person is not "a mere medium through which information flows." (p. 292). Instead, people exhibit considerable agency when regulating their own cognition. Forward models have been popular as a way of formalising metacognitive models (Nelson & Narens, 1990), and, according to some theorists, these models exist in many forms of skilled activity (see Christensen et al. 2016; also see, Blakemore et al., 2001; Frith, 2012; Wolpert & Kawato, 1998). One interpretation is that the cerebellum,[41] with its grid-like parallel fibres, receives a copy of the motor command, then uses this copy to form a prediction about the consequences of the action; if there is a mismatch between prediction and sensory consequence, fine adjustments are made. Mostly these models invoke Conant and Ashby's (1970) proposed 'good regulator' theorem.

[41] Allen and Friston (2018) have criticised the idea that the cerebellum has a unique role in comparator models. In any case, regardless of how and where the comparator models are implemented in the brain, it is clear that feedback and feedforward mechanisms are relevant to metacognition.

> The theorem has the interesting corollary that the living brain, so far as it is to be successful and efficient as a *regulator* for survival must proceed, in learning, by the formation of a model (or models) of its environment. (p. 89, italics added).[42]

In brief, the homeostatic model involves a "model" of the world, and this in turn has a predictive function. These models have much in common within a broader tradition of British cybernetics (Dewhurst, 2018). Nonetheless, the Conant and Ashby model has sometimes been characterised in a metalinguistic manner (i.e., in the manner of the language-of-thought hypothesis criticised in section 2.1). For example, Nelson and Narens (1990) invoke and interpret Conant and Ashby in terms of meta-sentences. Proust (2013) notes that Conant and Ashby's model need not be exclusively related to linguistic representation; indeed, on Proust's view, this metarepresentational view amounts to a misinterpretation of their model (p. 18). Conant and Ashby's model can be interpreted as entailing analogue and mathematical models. This interpretation is consonant with the way Dewhurst (2018) presents a case for embodied homeostatic mental representation models. On this interpretation, homeostatic models can be viewed as akin to analogue cognition rather than a digital or Fodorian language-of-thought form of mental representation.

### 2.3. (Non)Conceptual Content

Procedural metacognition has nonconceptual content (Proust, 2013). But what exactly is nonconceptual content? Nonconceptual content is typically characterised by the way in which perception can be non-conceptual and non-propositional. Evans (1982) observed that nonconceptual content is fine-grained; in other words, nonconceptual content outstrips what can be said about it. For example, the colour red has many more perceptible properties that we can ascribe names for it. The perceptual experience of the colour can be understood as outstripping any propositional account of the experience. Evans (1982) claims that another way of arguing for nonconceptual content is to focus on how nonconceptual content does not rely on the *generality* constraint, nor on the property of *objectivity*. The generality constraint refers to the ability of an agent to combine sentences systematically and in arbitrary ways using a rule-based system; for example, if a subject possesses the thought that *a* is *F*, then the subject will be able to generate the thought that *a* is *G* (p. 104). *Objectivity* is the property of a

[42] Williams and Colling (2018) note that predictive processing, while it may seem consistent with Ashby and Conant's (1970) theory, is nonetheless independent of the regulation of iconic representations.

representational system that refers to stable and permanent objects independently of what is proximately perceptible. A propositional format can be deemed to possess generality and objectivity. Procedural metacognition lacks these features as it is non-propositional and thus non-conceptual. Nonetheless, as Proust (2013) observes, procedural metacognition is sensitive to normativity – at least how veridical a thought is– such that there is a requirement for at least some form of content; this is where nonconceptual content enters the picture.

Various arguments have been advanced in support of nonconceptual content. A fairly representative list can be seen in the work of Bermúdez and Cahen (2020).While space doesn't allow me to go into any of these arguments with much specificity, I will briefly review them.

1. Nonconceptual features do not exhibit propositional attitudes. For example, nonconceptual content can simultaneously represent impossible or contradictory states of affairs. The canonical example is that of the waterfall illusion. Crane (1988) shows how opposing ideas can co-occur (this is where a motion aftereffect presents an illusion of movement such that an image of a waterfall can be perceived as going up and down at the same time). Crane has thus used this as empirical evidence that non-perceptual content exists, and as a way of contrasting nonconceptual content with conceptual content.
2. Another, second, line of evidence is that thinking is analog (as proposed by Dretske, 1981). This is the idea that more information is contained in information than is directly apparently at first. For example, the proposition *A* is *F*, in digital form, strictly carries the information contained in *A* insofar as it relates to *F*. In linguistic form, that 'a cup carries coffee' refers to nothing more than that the cup has coffee. Nonetheless, an analog representation (such as a picture) carries much more; in short, analog carries more content than digital representation.
3. Content is deemed to be unit-free (a point advanced by Peacocke).
4. Perception is fine-grained.
5. Learning concepts presupposes nonconceptual content.[43]
6. Concepts are independent of context (i.e., they're not demonstrative concepts).

[43] This sounds similar to what Macpherson (2015), in her list of characteristics that constitute nonconceptual content, refers to as how nonconceptual perceptual experiences lead to the acquisition of concepts such as shape, size, colour and pitch. Apparently missing on Macpherson's list is express mention of the context-dependence of nonconceptual content. Nonetheless, the other characteristics that feature on Macpherson's list imply that nonconceptual content is context-dependent.

> 7. Finally, infants and nonhuman animals have nonconceptual content. This line of argument is derived from empirical findings in comparative and developmental psychology. The central idea is that non-human animals and infants do not have propositional, conceptual capacities, and yet are able to perceive content. Invoking *nonconceptual* content is a way of accounting for this.

McDowell (1994) has famously argued against nonconceptual content; however, he has since revised his view. McDowell (2008) now maintains that conceptual content can contain non-propositional content; yet it is still conceptual because it is *conceptualisable*. Noë (2004) argues that on an enactivist reading there is a graded quality to concepts. Rather than concepts involving *explicit deliberative judgment*, according to Noë "conceptual skills can also enter thought as background conditions on the possession of further skills of one sort or another." (p. 187). Furthermore, "there may be no sharp line between conceptual and nonconceptual. Indeed, it may be that sensorimotor skills deserve to be thought of as primitive conceptual skills" (p. 31).

McDowell (2008) observes that not all the content in perception needs to be actually conceptual; the point is that the content is *potentially* conceptualisable. But, Crane (2013) wonders, if the content is not conceptual but only potentially conceptual, why shouldn't this content be referred to as *non*conceptual content. In this case it appears that advocates of nonconceptual content can agree with McDowell's (2008) non-propositional conception of content. Similarly, Thompson (forthcoming) notes that the way Noë connects skilled, sensorimotor engagement with conceptuality world "stretches the notions of understanding and conceptuality too far, that they become vacuous." (p. 44). Noë's definition of conceptual appears similar to what others would simply call nonconceptual. A nonpropositional definition amounts to a deflationary reading of conceptuality that in turn presents less of a threat to procedural metacognition. In any case, it all depends on how demanding one's definition concept is intended to be; if one sets a low threshold for conceptuality, then a lot of nonconceptual content will be conceptual (Bermúdez & Cahen (2020).

One of the arguments McDowell puts forward is that we need concepts so as to conceptually ground beliefs. But Proust (2013) responds to McDowell by observing that "rationality in non-verbal organisms should be characterised not in terms of explicit reason-giving ability, but in terms of entitlement." (p. 126). Infants and non-human animals can be granted warrant (even if not justification); here *warrant* pertains to a form of epistemic entitlement that does not presuppose that the agent is using reason. An infant or non-human

animal can be epistemically 'correct' by way of entitlement even though they do not possess the capacity to reason. This warrant involves a featural system that does not involve generality or objectivity.[44] Nonetheless, even if, for the sake of argument, one did argue for the existence of McDowell-style conceptual content, it wouldn't matter for Proust's thesis, because, as noted, McDowell no longer endorses the view that conceptual content needs to be propositional (McDowell, 2008). In effect, this non-propositional view on conceptual content is consistent with Proust's view that procedural metacognition is non-propositional. Nevertheless, I will follow Crane (2021) and Thompson (forthcoming), in the term non-conceptual as their definition aligns with the difference between propositional and non-propositional content.

*What does this mean for metacognition and extended cognition?*

As noted, Proust does defend a nonconceptual account, at the procedural level, and this, I propose has much in common with the extended cognition literature. See Figure 1 below.

**Figure 1**

*Nonconceptual and conceptual content as it relates to metacognition and extended cognition and mind*

| **Non-conceptual content** | **Conceptual content** |
|---|---|
| ↓ | ↓ |
| **Procedural metacognition** | **Analytic metacognition** |
| ↓ | ↓ |
| **Extended cognition processes** | **Extended mental states** |

As can be seen above, non-conceptual content is concerned with procedural metacognition and extended cognition processes. Conceptual content is epistemically and ontologically associated with analytic metacognition and extended mental states.

[44] On this view, representations can receive proximal and distal interpretations. Proust draws on Cussins' (1992) work on cognitive trails; "an organism that can pick up distal information is also able to store its knowledge not only in the form of its own dynamics, but also by relying on the organisation of the world itself." (Proust, 2013, p. 114).

The noteworthy point here is that procedural metacognition operates with nonconceptual content. Extended cognition also operates with nonconceptual content as extended cognition is procedural and non-propositional. Following Pöyhönen (2014), the concepts of extended *cognition* and extended *mind* can be viewed as referring to different domains. Extended cognition is conceptualised as a naturalistic domain in the sciences while the extended mind refers to a broader domain that includes folk-psychological discourse, propositions and intuitions. Bermúdez (2005) suggests that know-how can take place by way of imagistic reasoning, trial and error reasoning, analogical reasoning, and the exercise of complex bodily skills (p. 302). This know-how, with its non-conceptual format, has much in common with procedural metacognition and extended cognition. On the other hand, extended mind involves propositions, and extended analytic metacognition emerges from this relationship. Although of course Figure 1 is by necessity highly simplified. In fact most processes are interactive, as can be seen in Figure 2.

**Figure 1**

*Content (conceptual and nonconceptual), metacognition (procedural and analytic), and extension (cognitive and mental) in interaction.*

| | | |
|---|---|---|
| **Non-conceptual content** | ↔ | **Conceptual content** |
| ↕ | | ↕ |
| **Procedural metacognition** | ↔ | **Analytic metacognition** |
| ↕ | | ↕ |
| **Extended cognition processes** | ↔ | **Extended mental states** |

The non-conceptual can become conceptual, as we have seen; procedural and analytic metacognition reinforce each other; epistemic feelings offer an evaluative bedrock for analytic metacognition, and, in turn, analytic metacognition can correct for replace enrich, redescribe, and sometimes correct for erroneous cases of fluency and familiarity. Further, extended cognition and extended mental states (with propositional content) are also linked. For example, insofar as the propositional information in Otto's notebook is part of his 'mind,' it, too, involves extended cognition coupling processes (an implicit form of know-*how*), and a connection with extended mental states (knowing-*that*). The determinate content (for example, the specific fact

of where the museum is located) is connected with, and grounded in, the extended, coupling processes.

### *2.4.1. Conceptual binding and cognitive penetration.*

The interconnections between metacognitive nonconceptual content and conceptual content can be considered by way of cognitive penetration. Macpherson (2015) notes that psychologists have typically been interested in low-level (early visual) forms of cognitive penetration. Pylyshyn (1999)[45] rejects this concept of cognitive penetration and instead thinks that the effect can be explained by attentional effects. Philosophers, on the other hand, have typically been interested in the *experience* of cognitive penetration.

With the former, the idea of cognitive penetration is a rejection of the idea of Fodorian modularity (Fodor, 1983).[46] One of the key characteristics of modularity is that these hypothesised modules are informationally encapsulated; put differently, they are partitioned from thinking. On this view low-level perception is penetrated by cognition where cognition is defined as being aligned with propositions (beliefs and desires, intentional states more generally). For example, when I see a tree, I am presented with perceptual contents that are conceptually richer than the mere shape, texture and colour of the tree. At the early stages of vision concepts 'penetrate' perception, and thus, in some cases, we don't just see a series of shapes and colours that we then need to overlay with concepts; rather, we see the tree in the very perception itself. Many arguments for cognitive penetration have accepted some form of Pylyshyn's (1999) *semantic criterion* definition[47] of penetration; roughly speaking, this is where a belief penetrates perceptual experience *only if* the content of the belief is rationally related to the consequent experience. Macpherson (2012), for example, has an account of cognitive penetration that follows Pylyshyn's definition, albeit in a weaker sense (see Stokes, 2021b).

Fodor and others have suggested that attentional effects account for the apparent cases of penetration. The classic rejoinder to cognitive penetration that Fodor advanced involved an

[45] Pylyshyn coined the term in 1980. Stokes (2021) thinks it is an unfortunate term.
[46] There in are in fact nine characteristics that Fodor (1983) originally put forward when making a case for modularity: domain specificity; mandatoriness; limited access by central functions; fast processing; produce 'shallow' outputs; exhibit fixed neural architecture; have characteristic and specific breakdown patterns; and have characteristic and specific developmental pacing and sequencing. None of these characteristics are strictly necessary Fodor's (1983) modularity theory and Francis Galls faculty psychology. Also see Stokes (2021).
[47] Pylyshyn (1999) is against the idea of cognitive penetration, but nonetheless coined the term and put forward a definition that he believes has not been met by the evidence.

appeal to visual illusions such as the Müller-Lyer illusion. In brief, the Müller-Lyer illusion involves two identical lines, each affixed with arrows, that look different as a result of the arrows pointing either inwards (which makes the line look shorter), or outwards (which makes the line look longer). The idea Fodor was suggesting is that the persistence of the illusion where the two lines seem (perceptually) to be of different lengths – regardless of our (cognitive) beliefs about the lines actually having identical lengths – indicates that perception is partitioned from cognition. But, as Churchland (1988) argued, this is a strawman argument as no theorist is claiming that all perception is penetrated in every case. Moreover, Churchland draws on evidence that there is cultural variability in how the lines are perceived. But, more importantly, there are definitional issues at stake in cognitive penetration (Stokes, 2021b). Cognitive penetration can occur by way of the role of selective attention in cognition (Stokes, 2021a, 2021b), and the diachronic effects of sustained skilful practice on perceptual expertise (Fridland, 2015; Stokes, 2021a, 2021b). Such arguments also involve rejecting the Fodorian and Pylyshyn presupposition that there is a strict separation between perceptual modules and cognition (Shea, 2015). As Stokes (2021b) argues, there is, at best, limited evidence for cognitive modules; moreover, Stokes points out that there are no compelling empirical or theoretical reasons to assume that Fodorian-style modularity is the basis on which cognitive penetration is to be defended. In effect, the theorist advancing a case for cognitive penetration does not need to accommodate the Fodorian theory of modularity.

Stokes (2021b, p. 117) advances a case for cognitive penetration by reconsidering the role of attention.[48] A more standard view, endorsed by Firestone and Scholl (2016), Fodor (1983) and Pylyshyn (1999), claims that attention is an intermediary or gatekeeper between cognition and perception. For example, when Churchland (1988) invokes the duck-rabbit illusion - in which one's perception of the presented image flips between a duck and a rabbit – the standard line is that shifts in spatial attention change the input into the visual system; in other words, attentional effects can account for the supposed cognitive effects on perception. However, Stokes (2021b) has amassed a wealth of evidence that suggests attention is more than a gatekeeper. 1.). Attention is often 'guided' without the agent intending to do so. 2). It can change independently of spatial attention. 3). These non-spatial attention mechanisms are influenced by cognition, and 4). these attention mechanisms can influence conscious experience.

[48] Stokes (2021) also argues that cognitive penetration is an epistemic good in what he terms a consequentialist line of argument.

Fridland (2015) argues that there is an alternative way of examining cognitive penetration that emphasises non-propositional skill. Fodor (1975) equates cognition with propositional thought. However, Fridland (2015) notes that cognition need not be viewed as propositional. Propositional content – such as that which fulfils Evans' (1982) conditions of systematicity and generality – rests on more basic perceptual abilities. These basic perceptual abilities include making perceptual discriminations of similarities and differences, the ability to group features into a stereotypical pattern, or the ability to act skilfully (Fridland, 2015, p. 113) by way of a practical *know-how*[49] (in the style of Ryle, 1949). According to Fridland, the ability to perceive nonconceptual content is like the ability to exercise a skill; the fine-grained nature of skill is similar to the fine-grained nature of nonconceptual perception (p. 114).[50] Skill involves a sensitivity to context that context-independent conceptual content lacks.

### *2.4.2. Cognitive penetration, extended cognition and metacognition*

The idea I am proposing is that cognition can penetrate lower-level perception; thus, this is one way of connecting – or binding together – the procedural and analytic forms of metacognition. Nonetheless, what I am proposing is in opposition to what Proust (2015) states. Proust states that fluency is *never* cognitively penetrable. Proust cites a study by Nussinson and Koriat (2008) in which fluency remained unaffected by the higher-order knowledge. More specifically, participants were exposed to an unsolved anagram puzzle; next, they were exposed to anagram task that came with a solution. The participants were asked to rate the difficulty of the anagrams for participants who didn't have access to the anagram solution. The ratings were contaminated by the sense of fluency of the solved v. unsolved anagrams; further, this contaminating effect persisted, under time-pressure, even when the participants re-rated the perceived difficulty for others. Proust (2015) interprets this as meaning that the knowledge didn't penetrate fluency.

Nevertheless, I think Proust is a little too quick to reject the possibility of penetration. First of all, as was seen in the examples by Fridland (2015) and Stokes (2021b), diachronic skill is

[49] There is a debate as to whether knowing-how is derived from knowing-that propositions (Stanley, 2011) See Devitt (2011) for the view that there is empirical support for embodied, procedural knowledge not having propositional representations; and Schwartz and Drayson (2019) argue that the debate has meta-philosophical implications regarding the extent to which these metaphysical debates should be naturalised. They argue that Stanley (2011) is inconsistent in only appealing to cognitive science for a naturalised epistemological view when it suits his case; otherwise, Stanley mostly appeals to abstractions in the philosophy of language.

[50] Macpherson (2015) offers an account of how nonconceptual content can co-exist with a theory of cognitive penetration.

a good candidate for cognitive penetration. In the anagram tasks, the tasks were arguably too novel, and the participants lacked sustained experience with them, such that cognitive penetration never got a chance to develop. So, it appears that while Proust (2015) is rejecting cognitive penetration insofar as cognitive penetration consists of synchronic belief states, it is worth remembering that cognitive penetration can unfold in a skills-based, nonconceptual manner, and also by way of attentional mechanisms, which can be classed as cognitive (Shea, 2015; Stokes, 2021a, 2021b). Moreover, given that Goupil and Proust (2022) have recently proposed that curiosity is a form of epistemic feeling, and given that curiosity is related to attention, it could be that curiosity is a suitable target for cognitive penetration.

I present this focus on cognitive penetration as a potential way in which procedural and analytic metacognition can be bound together. This is important for an account of metacognition that involves the 'dual processing' of two forms of metacognition (the procedural and analytic). Indeed, Arango-Muñoz (2015) has observed that the connection between procedural and analytic metacognition is a weakness in Proust's (2013) account. Arango-Muñoz does not believe Proust offers as sufficiently comprehensive account of the connection. Given this gap, I propose that cognitive penetration can assist in connecting the two forms of metacognition; however, I do not mean that these two forms of metacognition will be always bound by way of cognitive penetration. Indeed, as Churchland (1988) pointed out, only in select situations will cognitive penetration occur.

### *2.4.3. Language and content*

How else is conceptual content connected with nonconceptual capacities and metacognitive feelings? Eleanor Gibson and Rader (1979) noted that attention is educated (as can be seen for example, by way of Sterelny, 2012; Tomasello, 2014). Bermúdez (2003) states that we are never conscious of propositional thoughts that have no *linguistic* vehicles (p. 60). Further: "We need, therefore, to distinguish weak and strong senses in which a representational vehicle might be structured." (p. 161). In the weak sense, structure exists when there is an isomorphism between vehicle and what the vehicle represents. On the other hand, in the stronger sense, structure requires "basic representational units combined according to independently identifiable combinatorial rules. Natural language sentences (or for that matter sentences in the language of thought) are clearly structured in the strong sense, whereas maps/models only possess structure in the weak sense." (Bermúdez, 2003, p. 161). Moreover, as O'Brien and Opie (2015) point out, linguiform content is the explananda of cognitive science and not the

'ground-floor furniture', thus language is an emergent feature. (p. 728). Huebner (2018) observes that conceptually-structured thought is a product of being socially situated, and conceptual categories are linked with what is most socially salient in the interactions people are engaged in (p. 24). Nonetheless, the public symbol system, in exploiting perceptual-motor mechanisms, can offer a way of thinking, and indeed thinking about thinking, that is unique to language-users. Language transforms passing thoughts into 'objects.' This is a view exemplified by Clark (1997) when he observes that:

> By 'freezing' our own thoughts in the *memorable*, *context-resistant, modality-transcending format* of a sentence, we create a special kind of mental object – an object that is amenable to scrutiny from *multiple* cognitive angles, *is not doomed to alter or change* every time we are exposed to new inputs or information, and *fixes* the ideas at a high level of abstraction from the idiosyncratic details of their proximal origins in sensory input. (p. 210, italics added).

Language stabilises and crystallises otherwise free-floating sensory impressions. This context-invariance grants us the capacities that Fodorian representations can supposedly provide for us, but in fact the language mastery available here does not presuppose a language-of-thought form of representation. (This will be further explored in specifically inner-speech and metacognition in section 4.4).

A point worth expanding on is what Clark (1997) means when he notes that a sentence becomes "an object that is amenable to scrutiny from *multiple* cognitive angles." (p. 210, italics added). One way of approaching Clark's point is to consider Tomasello's (1999) research on the *perspectival* nature of language. On Tomasello's view, language provides more than conceptually 'handy tags' for cognition; in addition to this, words are permeated with intersubjectivity and are, by necessity, perspectival. The language-user is 'implicitly' aware that "any given experiential scene may be construed from many different perspectives simultaneously", and so internalising a language also means internalising a multiplicity of viewpoints (p. 128). Furthermore, Tomasello points out that being enculturated in a language gives a person the conceptual resources to not only parse the world into events and participants, but also to re-interpret experiences such that actions can be conceptualised as objects, and objects as actions (p. 159). In terms of metacognition, around the ages of five to seven, a linguistically mediated form of metacognitive self-regulation occurs whereby the child appropriates and enacts for themself the instructions and rules that the adult has used toward

the child[51] (pp. 191-194). Language emerges from perceptual-motor systems and social-cultural practices, and in turn transforms these systems and practices. Moreover, Goupil and Proust (2022) hypothesise that the discrete, linguistic responses adults give in response to the non-verbal questioning behaviours of a child enrich the metacognitive, analog feelings of the child.

## 2.5.Conclusion

While it might seem strange that, in a thesis concerned with an extended account of metacognition, I have spent time overviewing some features of cognition that are decidedly brain-based, it is nonetheless important to do so. Extended accounts of course make space for neuroscience - even though there is a substantial debate about the existence and role representations – and this is as it should be (see Clark, 2008; Thompson, 2007). Nevertheless, it is crucial that an appropriate form of cognition is invoked, and I believe the most satisfying, conceptually reasonable, and empirically fruitful account is to be found in the literature detailing analogical cognition/representations that are non-propositional in nature rather than linguiform representations as proposed by the language of thought hypothesis. Analog cognition is consistent with extended cognition and procedural metacognition. Indeed, analog cognition is also consistent with subpersonal-level explanations; it is to this matter that the thesis turns in chapter three.

[51] Tomasello (1999) observes that the meta-discourses the adult engages in with the child play a role in the later forming self-regulation capacities of the child (p. 192). Goupil and Proust (2022) review a series of studies that demonstrate how conversational scaffolds can transform the epistemic feeling of curiosity into metacognitive questioning; for example, the asking of follow-up questions can be reinforced in the child if the adult offers *explanations* rather yes/no (or 'that's the way it is') responses to the child's question; additionally, children who hear adults offer explanations are more likely to create their own explanations when confronted with answers that don't satisfy them.

## 3. The Personal-Subpersonal distinction in relation to Metacognition and Extended Cognition

In the following chapter I focus on the role of the personal-subpersonal distinction as it relates to procedural metacognition and extended cognition thesis. This focus thus serves to provide evidence for claim one of the thesis: this is the claim that cognitive extension and metacognition are subpersonal-level phenomena. It will be argued that the personal-subpersonal distinction is important for understanding the ontological status of procedural metacognition extended cognition. I begin to make this case in section 3.1 by drawing upon the work of Drayson (2012, 2014) when considering the contested space in which the personal-subpersonal exists; to this end, I agree with Drayson that some of the confusion generated by the distinction is a result of conflating subpersonal-personal levels with Stich's (1978) doxastic-subdoxastic state distinction. In 3.2 I argue that extended cognition is best conceptualised as subpersonal, as are many of the metacognitive theories. In section 3.3 and 3.4, I bolster my claim that extended cognition is subpersonal by arguing – contrary to Roth (2015) - that extended cognition should be conceptualised as a subpersonal-level phenomenon.

This chapter builds on chapter two by continuing the idea that cognition is mechanistic and can be explained at levels that become progressively smaller and yet remain all part of the same cognitive phenomenon. More specifically, the analog representational format explored in chapter is consistent with the claim in the present chapter that procedural metacognition and extended cognition are subpersonal-level phenomena. The present chapter also provides a backdrop for the discussion of metacognitive norms in chapter four (specifically, see section 4.5 subpersonal metacognitive level).

### 3.1.Defining the Personal-Subpersonal Distinction

Dennett (1969) originally put forward the distinction between the subpersonal and the personal in *Content and Consciousness*. In many ways the distinction was intended to reconcile non-mechanistic vocabularies[52] of mental-states (as proffered by philosophers such as Ryle and

[52] Borderlines of psychological discourse have been historically contestable. Danziger (1997), when examining the nineteenth-century development of psychology as an autonomous science, observes that the language of the discipline had two challenges. First of all, it needed to differentiate itself from the everyday usage of psychological terms such as *attitude*, *motive*, and *temperament* – "terms that carried a mixed somatic and psychological meaning" (p. 52). Secondly, it needed to find a place for itself between the "non-communicating solitudes" of physiology and moral philosophy (p. 52) – the latter dominant at the time. But, in finding a place for itself, two research domains remained ambiguously at once physiological and psychological: *sensation* and *animate motion*. This physiological-psychological interface admitted of a good deal of porousness with "no sharp line" (p. 53).

Wittgenstein's late works) and sub-personal explanations that appealed to work conducted in the cognitive sciences of the time. Ryle and Wittgenstein tended to be weary of mechanistic explanations; indeed, Ryle (1949) thought it was a category error to look for the 'person' in the neurophysiology of a person. Dennett (1969) approached this problem by respecting the intuitions of these philosophers by partitioning non-mechanistic descriptions to the *personal-level*; at the same time, by making space for the *subpersonal-level*, gave Dennett the freedom to pursue mechanistic explanations.[53]

Drayson (2012) attempts to resurrect the original interpretation of Dennett's (1969) distinction as a vertical explanation that avoids the threat of a homunculus regress (more on this below). To that end, Drayson lucidly illustrates that the distinction does *not* necessarily map onto the distinction between conscious and unconscious. Nor does Dennett's distinction map on to the normative and non-normative (I have more to say about this in section 4.5). Indeed, Drayson notes that "The sub-personal level of *explanation* can posit conscious states, accessible to introspection." (p.22, italics added). Indeed, Drayson (2012, 2014) observes that the personal-subpersonal distinction is often confused with another distinction originally advanced by Stich (1978) between doxastic and subdoxastic states.[54] Briefly, doxastic states are states that we can access, and that are inferentially integrated with other doxastic states. *Sub*doxastic states, on the other hand, are inaccessible and non-inferentially integrated. Drayson uses the example of edge detection in vision or Chomsky's generative grammar as an example of subdoxastic states and processes. It is not possible to ever access these states and processes (assuming they even exist, in the case of generative grammar). Now, how does this relate to the personal-subpersonal distinction? Drayson's contention is that when philosophers and cognitive scientists invoke the term subpersonal they sometimes appear to really mean *subdoxastic* (as in inaccessible). In short, the personal and subpersonal levels do not neatly align onto a distinction between conscious and unconscious or doxastic and non-doxastic.

**Figure 2**

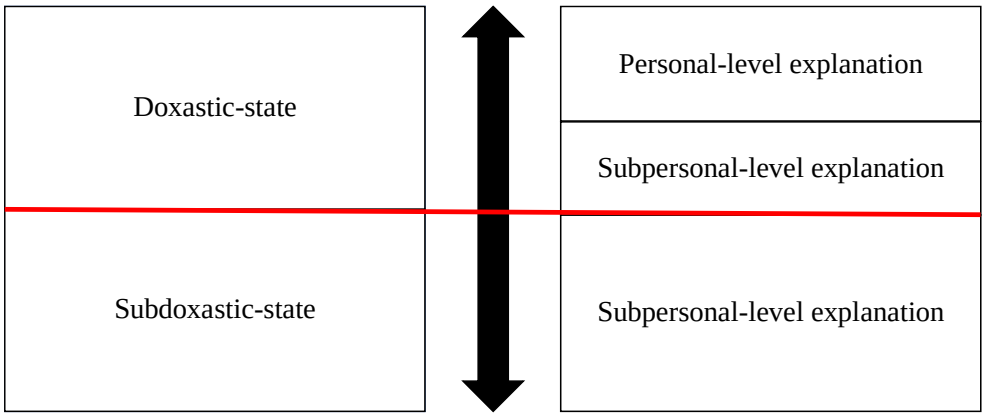
*Doxastic-Subdoxastic distinction and the personal-subpersonal level distinction.*

---

Indeed, Danziger notes that the conceptual development of psychology drew upon physiological concepts such as *organism*, *stimulus*, *reflex* and *energy*. In contemporary psychology, we can still see many areas of psychology straddling the boundaries of what we now call biology *and* folk-psychological, mentalistic concepts. Cognitive science, which encompasses neuroscience and psychology (among other disciplines), also admits of this range.

[53] See Dennett (2015) for reflections (close to fifty years after the publication of *Contents and Consciousness*) on the intellectual climate of time and how this influenced the development of the distinction.

[54] This can be seen in Thompson (2007) when he notes that: "Mental processes, according to cognitivism, are "subpersonal routines," which by nature are completely inaccessible to personal awareness under any conditions." (p 5 - 6).




Note: *Subdoxastic states are by necessity at the subpersonal explanatory level; however, doxastic states can be at the both the personal and subpersonal explanatory levels.*

As can be seen in Figure 3, Stich's (1978) doxastic-subdoxastic states are not equivalent to the personal-subpersonal level distinction. Instead, doxastic states can be personal and subpersonal. As Drayson (2014) points out, Stich's distinction is really between "those states posited by subpersonal explanation that are *also* referred to by personal explanation and those states that appear *only* in subpersonal explanation" (p. 343).

The original explanatory levels interpretation is connected to the avoidance of what Dennett (1991) has termed a 'Cartesian theatre;' namely the fear of the infinite regress wherein, inside a person, behind one centralised mental interpretation there is another interpreter, and behind that another interpreter and behind that one yet another and so on ad infinitum. Dennett (1978) articulated the a solution to this problem by way of *homuncular functionalism* whereby sub-systems of progressively specialised functions that bottom out at something unrecognisably human. Dennett notes: "If one can get a team or committee of *relatively* ignorant, narrow-minded, blind homunculi to produce the intelligent behaviour of the whole, this is progress," (p. 123). Dennett goes on to use the intuitive analogy of boxes within boxes, each box smaller than the last, that could be "replaced by a machine." (p. 124).[55] This view is consistent with, and indeed arguably vindicated by, multilevel neurocognitive mechanisms where structure and

[55] Dennett, inasmuch as he remains a functionalist, has arguably reconsidered his view in recent years so as to incorporate the biological implementation of some forms of cognition (specifically, the way in which the complexities of cells and tissues play functional roles). See: Levin and Dennett (2020) This view overlaps with that of Boone and Piccinini (2016): "structures constrain functions and vice versa" (p. 1522).

function constrain each other (Boone and Piccinini, 2016; Piccinini, 2020, 2022; also see chapter two of this thesis).

Is Drayson's (2012, 2014) concern with the (mis)use of the personal-subpersonal distinction a losing terminological battle? Drayson maintains that there are independent grounds for justifying that the personal-subpersonal distinction should be used so as to conform with Dennett's (1969) original meaning rather than Stich's distinction. These two distinctions are basically targeting different domains: levels of explanation (in Dennett's original account) and consciously accessible states and processes (in Stich's (1978) account). Moreover, the fact that theorists use the terminology inconsistently can lead to conceptual problems and invite confusion. Drayson (2014) cites examples of some prominent philosophers sliding between different uses of the distinction; in one example, the switching occurs over the course of two consecutive sentences. At an absolute minimum Drayson makes a plea for consistent and unambiguous usage of the distinction.

Drayson (2012, 2014), when defending Dennett's (1969) early distinction, draws on the work of Jaegwon Kim (2005). Kim (2005) For example, if someone broke a window, the diachronic (personal) explanation would best be covered by appealing to the series of person-level events that led to the breaking of the window (the picking up of the stone, the throwing of the stone, and so on); the window however, breaks in a way that is concerned with the molecular structure of glass and the structural laws of physics (which are synchronic and subpersonal).Nevertheless, Westfall (forthcoming) argues that it is wrong to follow Drayson (2012) in associating the personal-level with Kim's (2005) conception of horizontal explanation and the subpersonal-level with vertical explanation. Westfall draws upon textual evidence in Dennett's (1969) account to the effect that even Dennett seemed to suggest that, contrary to Drayson (2012), subpersonal explanations can be *horizontal* (i.e., involve citing earlier events to explain later events), and personal explanations can be *vertical* (i.e., explaining a capacity in terms of its components). Westfall also appeals to evidence from Marr's (1982) account of the early visual system (e.g., as it pertains to depth perception or edge detection) in which subpersonal explanations can take the form of vertical-explanation capacities *and* horizontal causal explanations. Similarly, Westfall notes that personal-level explanations – for instance, those concerning a person's attributes – could be explained both horizontally *and* vertically. Westfall uses as an example an imaginary philosopher called John. The fact of John being a good philosopher can be explained by way of a *horizontal explanation*; for example, a chain of events that includes his extensive reading and good supervision could be drawn on to explain the quality of John's work. But, additionally, a *vertical explanation* that includes sub-

capacities – traits of listening well, reasoning well, and so on – could be invoked to explain the fact of John being a good philosopher. The upshot of all this is that Westfall rejects Drayson's (2012) claim that the personal-subpersonal distinction should be interpreted in terms of horizontal (personal) and vertical (subpersonal) explanations. Even though Westfall attempts to rehabilitate the personal-subpersonal distinction, he offers good reasons for rejecting the vertical-horizontal criterion

Westfall's preferred way of examining the personal-subpersonal distinction is in terms of folk-psychology constructs (personal) and non-folk-psychological constructs (the sub-personal). This is view is similar to the way in which Zawidzki (2021) aligns procedural metacognition with the subpersonal-level, and the personal-level with folk-psychology.

The personal-subpersonal distinction is also appealed to in some of the enactivist literature. For example, Noë (2015) invokes what he calls the "embodiment level," and describes it as follows:

> The crucial thing about the embodiment level is that it is neither entirely *personal* (conscious, controlled, governed by thought and planning), nor is it properly *sub*personal (automatic, reflexive, independent of thought and understanding). Subpersonal-level activity unfolds on time-scales of milliseconds. Personal-level action, in contrast, takes place at much larger time scales of minutes, hours, days, weeks, and lifetimes. The embodiment level, as Dana Ballard, who first introduced this idea, understood, unfolds at an intermediate level, on the scale of seconds. 2015, p. 218).

It is important to notice the way in which Noë equates the subpersonal with that of the *automatic* and *non-conscious*. Here Noë appears to be confusing subpersonal-*level* explanation with subpersonal *states*. Moreover, Noë distinguishes the personal from the subpersonal by way of time-signatures[56] in which the subpersonal supposedly operates over milliseconds whereas the personal supposedly ranges from minutes to lifetimes. In contrast to what Noë proposes, I don't think there is a principled reason why the personal-subpersonal distinction needs to align with these time-signatures. In fact, even by the lights of Noë's own, by no means

[56] Noë's view overlaps with Gallagher's (2017) Varela-inspired view that different levels correlate with different time-signatures: the *elementary* level ranges from 10 – 100 milliseconds; the *integrative* level ranges from 0.5 – 3 seconds; and the *narrative* level is upwards of 3 seconds.

uncontroversial, definition of the distinction – where the personal is apparently conscious, controlled, thoughtful, and involves planning – it is unrealistic to assume that any of conscious, controlled thought cannot also occur by way of milliseconds (see section 5.2. in this thesis for examples of automaticity and conscious control operating by way of milliseconds; also see Kirsh, 2004). Conversely, the subpersonal-level – which Noë associates with automaticity, reflexes, and thought/understanding-independence – should by no means be restricted to a time scale of milliseconds. For example, subpersonal level explanations for phenomena can span over evolutionarily vast time-scales (see section 4.5 in this thesis).

Taken together, it is evident that there are various ways of characterising the personal-subpersonal distinction. To refresh, let's look at some of the views covered:

1. Drayson's (2012) view is that we should not confuse the personal-subpersonal distinction with doxastic states, consciousness, or normativity; however, Drayson does align the distinction with Kim's (2005) horizontal and vertical explanations.
2. Westfall (forthcoming) rejects Drayson's horizontal-vertical explanation interpretation. Instead, Westfall opts for a folk-psychological-non-folk-psychological interpretation.
3. Noë (2015) offers an interpretation that appeals to time-signatures (i.e., that the personal-level occurs over the course of minutes and longer, whereas the subpersonal-level occurs within milliseconds). I criticised this view a moment ago.
4. Rupert (2018) offers an account whereby personal-level and personal-states can be accommodated within the subpersonal ontology of cognitive science.

For the purposes of the present thesis, I endorse Drayson's return to Dennett's original (1969), explanatory-level account of the distinction; however, I also agree with Westfall's rejection of Drayson's attempt to align the distinction with Kim's (2005) horizontal-vertical distinction. Overall, I agree that the personal-level is folk-psychological in nature.

### 3.2. Metacognition, Extended Cognition, and the Personal-Subpersonal Distinction

*Metacognition*

Procedural metacognition can be considered as existing at the subpersonal-level and elements of analytic metacognition can be captured by the personal-level (Zawidzki, 2021). More

precisely, Zawidzki claims that person-level concepts, which he views as *socio-cognitive tools*, play an interpretive role in metacognition when understanding self and other. Zadwidzki cites concepts such as 'parent', 'spouse', 'teacher', and 'citizen' – concepts that have social salience – as providing a means of coordinating one's behaviour so as to make it more predictable to others. Because these concepts are frequently used they have a social currency and functionality; invoking these concepts assists in regulating the behaviour of self and other. For example, if someone talks about *parenthood*, the invocation of this concept makes certain types of behaviour more predictable. Zawidzki suggests that the mastery and possession of concepts has implications for metacognition. When a person believes-*that*, desires-*that*, or grieves-*that*, they, in doing so, play a social role and make use of concepts that have frequency within a particular social context.

What Zawidzki is suggesting here does not exhaust the entirety of metacognition – remember that Zadwidzki claims that subpersonal-level procedural metacognition also exists – but what Zawidzki is concerned with the conceptual form of metacognition that Proust (2013) labels analytic metacognition. Recall that in the previous chapter (chapter two) I examined in section 2.4.3. the role of language in metacognition. There I claimed that nonconceptual (procedural) metacognition is connected with conceptual content (as can be seen in the case of analytic metacognition) by way of mastery of language. I endorsed Clark's (1997) view that language 'freezes' thought (p. 210), and I noted that language, on account of its *perspectival* nature, amplifies the capacity to think about objects and actions from multiple points of view (as argued for by Tomasello, 1999, 2014). Furthermore – and importantly for my overall thesis – I suggested that language mastery is one way in which procedural and analytic metacognition are connected (claim two of this thesis). What Zawidzki (2021) proposes regarding the role of personal-level socio-cognitive tools is consistent with what I have proposed regarding the role of language. Moreover, Zawidzki's conceptualisation of the personal-level is aligned with 'folk psychology,' and this, in turn, is similar to the view of Westfall (forthcoming) who views the personal-subpersonal distinction as being primarily concerned with the folk-psychological (personal) and non-folk-psychological (subpersonal). Finally, it should come as no surprise that I agree with Zadwidzki on the point that procedural metacognition is a subpersonal phenomenon; indeed, Zadwidzki's view is consistent with at least half of claim one of my thesis, namely the claim that both metacognition and extended cognition are subpersonal-level phenomena.

Nevertheless, I believe it is worth pointing out that analytic metacognition – as present in the use of linguistically, conceptually-mediated metacognition – is not necessarily neatly

separable from subpersonal processes. There is not necessarily a sharp dividing line between the 'lower' and 'higher' metacognitive processes and their associated norms (Proust, 2013). More specifically, Proust notes that "It is true that analytic metacognition and reasoning are realized by 'subpersonal' processes, including ERN [error-related negativity] error signals: the latter help subjects to correct their behaviour, without awareness of error" (p. 299). Proust also observes that the flexibility of analytic metacognitive processes does not provide a principled way of distinguishing the personal from the subpersonal: "inflexibility has nothing to do with the fact that feelings are generated by subpersonal processes. All our flexible thoughts are also generated sub-personally." (p. 299). Subpersonal representations can be said to constrain the personal level (Proust, 1999). In summary, procedural metacognition is a subpersonal-level phenomenon, and - given the close and interactive relationship analytic metacognition has with procedural metacognition - analytic metacognition, at a minimum, is partially explained at the subpersonal level. Looming over these distinctions, however, are the contentious discussions about the roles of and relationships between cognitive (psychological) science and folk-psychology. Rupert (2018), for example, has offered a one-level, subpersonal account of cognition in which he goes so far as to claim that even subpersonal states should be located the subpersonal-level. It is worth noting that Rupert is not denying what may be typically deemed as personal-level beliefs, desires and visual experiences; rather, Rupert is following Drayson (2012, 2014) in viewing consciousness as orthogonal to the personal-subpersonal distinction. Indeed, what Rupert is appealing to is the way in which all of cognition is amenable to mechanistic description and explanation, "that if such *states* exist, we should expect to find them at the same *level* as mechanical, so-called subpersonal cognitive processing." (p. 17, italics added). Further, Rupert and Carter (2021) have put pressure on the notion that the subpersonal distinction should even exist; they argue that the personal-level is irrelevant to cognitive science. Or, put differently, Rupert (2018) notes that if folk concepts (beliefs, desires, etc.) are to be seen as relevant for cognitive science, these concepts need to fall within the naturalistic models and behavioural data of subpersonal-level cognitive science.

Elsewhere, Westfall (forthcoming) argues that it is wrong to follow Drayson (2012) in associating the personal-level with Kim's (2005) of horizontal explanation and the subpersonal-level with vertical explanation. Westfall draws upon textual evidence in Dennett's (1969) account to the effect that even Dennett seemed to suggest that, contrary to Drayson (2012), subpersonal explanations can be *horizontal* (i.e., involve citing earlier events to explain later events), and personal explanations can be *vertical* (i.e., explaining a capacity in terms of its components). Westfall also appeals to evidence from Marr's (1982) account of the early

visual system (e.g., as it pertains to depth perception or edge detection) in which subpersonal explanations can take the form of vertical-explanation capacities *and* horizontal causal explanations. Similarly, Westfall notes that personal-level explanations – for instance, those concerning a person's attributes – could be explained both horizontally *and* vertically. Westfall uses as an example an imaginary philosopher called John. The fact of John being a good philosopher can be explained by way of a *horizontal explanation*; for example, a chain of events that includes his extensive reading and good supervision could be drawn on to explain the quality of John's work. But, additionally, a *vertical explanation* that includes sub-capacities – traits of listening well, reasoning well, and so on – could be invoked to explain the fact of John being a good philosopher. The upshot of all this is that Westfall rejects Drayson's (2012) claim that the personal-subpersonal distinction should be interpreted in terms of horizontal (personal) and vertical (subpersonal) explanations. Even though Westfall attempts to rehabilitate the personal-subpersonal distinction, he offers good reasons for rejecting the vertical-horizontal criterion

Westfall's preferred way of examining the personal-subpersonal distinction is in terms of folk-psychology constructs (personal) and non-folk-psychological constructs (the sub-personal). This is view is similar to the way in which Zawidzki (2021) aligns procedural metacognition with the subpersonal-level, and the personal-level with folk-psychology.

*Extended Cognition*

Carter and Rupert (2021) note that progress has been made when theorists 'descend' to the subpersonal level. They refer to what Hutchins (1995b) work on the "steep gradients in the density of interaction [of external artefacts that]" that mark the boundaries of the cognitive system (p. 157); and that "centers and boundaries are features that are determined by the relative density of information flow across a system." (Hutchins, 2014, p. 37). These systems are often characterised as subpersonal (Pöyhönen, 2014). Similarly, Clark (1997) has argued that even higher cognition is constrained by perceptual recognition and decentralised mechanisms As was discussed in chapter one of this thesis, this focus on extended cognition is distinct from a focus on extended mind; the extended mind, being more folk-psychological in orientation, is arguably more amenable to the personal-level. Nevertheless, I claim that the extended cognition – and procedural metacognition – are best explained at the subpersonal-level.

### 3.3. Extended Cognition, The Personal-Subpersonal Distinction, and Non-derived Content

Roth (2015) claims that there has been "a conspicuous absence of discussion" (p. 137) concerning the relationship between the extended cognition/mind thesis and Dennett's (1969) personal-subpersonal distinction. Roth, I believe correctly, makes a case for why the personal-subpersonal distinction should be relevant for extended cognition. To that end, Roth (2015) seeks to defend the extended cognition thesis by way of Dennett's distinction. Roth claims that the extended cognition thesis should best be seen as operating at the personal level of processes (and explanation) rather than the subpersonal.

In light of this, my concern is that if one follows Roth's strategy of defending the extended cognition thesis at a *personal* level of explanation, then this will not align with procedural metacognition's status as *subpersonal*. This inconsistency would mean that metacognition and extension are working on two different levels. In effect, in opposition to Roth, I think it is preferrable to argue that extended cognition, like procedural metacognition, is at the subpersonal level. First, however, we need to examine what Roth claims.

Roth proposes that that the subpersonal processes are in the brain (and thus not extended) and that the personal processes are what arise when the agent interacts with external tools. The following quote captures his position:

> If we interpret BRAIN-BOUND as a claim about *sub-personal* processes, however, then, far from denying it, defenders of EXTENDED can say this is precisely how we should understand how certain *personal* level capacities are possible. Brain-bound cognitive processes are sub-personal cognitive processes instantiated in brains. The paper and pencil existing outside of the skull are not part of an extra-cranial cognitive process that interacts with a brain-bound cognitive process. Rather, it is the interaction between brain-bound *sub-personal cognitive processes* and external, non-cognitive tools that gives rise to exercises of personal-level cognitive processes. (p. 139).

Roth claims that "we need not deny that *neural* processes have a special priority over those that take place on paper." (p. 137) This is because the neural processes, according to Roth, are operating at the subpersonal-level. An example is instructive. Roth imagines that we are told to calculate 41 x 17, and then proceeds through a series of sub-goals to reach the answer (multiplying one and seven, etc.). Roth deems these goals to be *personal*, and it is not until, on

closer inspection, we examine the ability, that "the personal level drops out." (p. 137). Roth, playing devil's advocate, pre-empts how Adams and Aizawa would most likely approach the personal-subpersonal distinction – namely by invoking non-derived representations and states that are instantiated in the brain. The concept of non-derived builds on Adam and Aizawa's (2001, 2008, 2010) distinction between vehicles that are derived from conventions and those that are non-derived (because naturalistic). "Words, stop signs, warning lights and gas gauges mean what they do through some sort of social convention," Adams and Aizawa (2010) write, whereas "trees, rocks, birds, and grass mean what they do in virtue of satisfying some naturalistic conditions of meaning." (p. 70). Roth incorporates this derived-non-derived distinction into his case for extended cognition in which he maps extended mind/cognition on to the derived and personal and neural-based cognition on to the non-derived and subpersonal. I have reconstructed Roth's argument in a numbered format so as to make it clearer. We can see that:

1. Cognitive processes involve states / items of content.
2. Some of the content is derived from thoughts and not language.
3. Thoughts, unlike language, have non-derived content.
4. The non-derived content is at the subpersonal-level (per Roth's reading of Dennett).
5. Subpersonal content is brain-bound.
6. Subpersonal-level cognitive processes have non-derived content (per conditions 1, 2, 3) and are subpersonal (per 4), and the subpersonal is brain-bound (5).
7. **Conclusion**: brain-bound subpersonal-level cognitive processes have *special priority* over non-brain-based personal-level processes.

In effect, if non-derived representations have priority over derived representations; and, further, if these representations are in the brain, then this poses a serious challenge to the extended cognition hypothesis. It would mean that when, for example, someone is doing mathematics with pen and paper, the non-derived content of cognition is brain-bound, and the actions of using the non-neural tools and the markings of the mathematics on a page are derived, i.e. not *real* cognition. The boundary would exist at the non-derived content and that boundary exists at the cranium. Given that Roth *does* want to defend the extended cognition hypothesis, how does he propose to get out of this predicament? Roth's counter-response is to argue that we should concede that the non-derived content Adams and Aizawa argue for is indeed subpersonal (4) and brain-bound (5); however, *on Roth's view the extended view remains*

*personal*. In this way an extended cognition/mind theorist is immune from the non-derived content argument. Specifically, Roth writes that it is not correct to characterise thoughts in such an inclusive manner that non-linguistic thoughts constitute actual thinking. On Roth's view, we need the personal-level to actually explain genuine thought.

> "[It] is *not* harmless to use 'thought' in the expansive sense that includes sub-personal cognitive processes, go on to show that some thought content is not derived from language, and then directly conclude that the contentful states and activities attributable to *people* are not derived from language" (p. 140 – 141).[57]

Furthermore:

> "[b]y the lights of the explanatory project heralded by C & C [Clark and Chalmers], there is something fundamentally misguided about invoking this derived / underived distinction to argue for the priority of brain-bound personal level processes over extended, personal level processes." (p. 141).

Roth, in fact, claims that if we accept his argument that extended cognition is personal-level, one beneficial upshot is that Adam and Aizawa's anti-externalist position in a sense defeats itself. In other words, if we concede that Adams and Aizawa are correct that non-derived content is subpersonal cognition, and that it is brain-bound, then *we are not really explaining cognition anymore*. Roth notes: "As far as the debate over the extended mind hypothesis goes, however, such an admission would seem to undermine the dialectical force of invoking the distinction in the first place." (p. 141).

> [I]f non-derived content is the mark of the cognitive, and it turns out that the intentional capacities of people depend on sub-personal intentional capacities, then *extended or not*, the intentional capacities of people are not cognitive. In this way, invoking the distinction may threaten to prove too much. (p. 141).

[57] However, contrary to Roth, there are ways of arguing for non-linguistic thought. See Bermúdez (2003) for an account.

Roth appears to be wresting an idea of personhood - extendable personhood - away from Adams and Aizawa. He appears to be saying that Adams and Aizawa are focussed on the wrong target: they're focussed on in-the-head, non-derived, subpersonal cognitive processes rather than *thinking*, and it is only thinking that entails extended, derived, personal processes. According to Roth, even if we concede that cognition is bound by non-derived brain content, this at least frees up conceptual space to argue for the extended mind, and, after all, it is mind that we are most interested in anyway. In his conclusion, Roth reinforces this point by quoting from Dennett's (1969) *Content and Consciousness*: "The problem of mind is not to be divorced from the problem of a person. Looking at the 'phenomenon of mind' can only be looking at what a *person* does, feels, thinks, experiences." (p. 213). Thus, Roth is using the subpersonal-personal distinction to argue against critics such as Adams and Aizawa; and, in doing so, Roth draws out some of the Rylean ideas that inspired Dennett's early work. Thinking, feeling, acting *are what people do,* not subpersonal, intentional capacities. Per Roth, it would be a category mistake to search for thinking in the non-derived content of cognition. On Roth's view, the personal vocabulary is different from the subpersonal view of neural firings, and this difference must not only be preserved, but can actually help figure in extended cognition theories.

To reiterate, this arguably creates a dilemma for my thesis as I have so far supported the claim that procedural metacognition and extended cognition theories are *both* subpersonal level theories. I maintain that this is the right approach; however, if we follow Roth's way of identifying the subpersonal with brain-bound, then it would perhaps appear that we end up with this untidy result:

1. Procedural metacognition is subpersonal
1. The subpersonal is brain-bound (per Roth).
2. Therefore, metacognition is brain-bound.

I will go through my own objections to Roth. Indeed, there are three main ways one can oppose Roth's argument. First of all, Roth's characterisation of cognitive extension is problematic. Secondly, there is ambiguity around Roth's use of the subpersonal-personal distinction. Thirdly, there are problems with Roth's reliance on the derived-non-derived distinction.

### 3.4.What is extended: mind or cognition?

As I noted in chapter one, there is arguably a difference between extended mind and extended cognition hypotheses (see section 1.4). To reiterate, some theorists (Carter et al. 2018b; Drayson, 2010; Pöyhönen, 2014) would not go as far as theorising about the mind itself; rather, their target is cognition. More specifically, instead of *extended mind* (which is propositional and folk-psychological as exemplified by the famous Otto notebook case that is supposed to show that mental states can extend) their target is *extended cognition* (which is concerned with cognitive processes). Pöyhönen has made clear that extended mind and extended cognition are distinct: "Although the difference might appear insignificant, I suggest that the hypotheses apply to different domains, address largely different issues, and depend on different sources of evidence." (p. 737). I have followed Pöyhönen in honouring this distinction throughout this thesis. Roth (2015), however, appears to conflate the two terms. Roth consistently uses the term '*external mind*' rather cognition, and gives mind a Gilbert Ryle-inspired emphasis (i.e., thinking is what people do, not sub-agential functions).[58] Yet at the same time Roth also makes multiple references to *cognitive processes*. For example when illustrating how his preferred view of equating the personal with derived mind and the subpersonal with non-derived content, Roth notes that his view

> opens the door to a reproachment of two otherwise seemingly incompatible claims: (a) problem solving, calculating, figuring things out – these are all *cognitive processes*, and a lot of these processes take place outside of the head, e.g., on scratch paper, iPads, whiteboards, and so on; (b) problem solving, calculating, figuring things out – these processes are made possible by *cognitive processes* that occur solely within the confines of the skull. The key to reconciliation, of course, is to note that the *cognitive processes* mentioned in (a) belong to the personal level, while the *cognitive processes* mentioned in (b) belong to the sub-personal level. (p. 137, italics added).

[58] Near the end of their 1998 paper, Clark and Chalmers speculate on the idea of extended selfhood. This idea is somewhat orthogonal to the proposal of extended cognition. It is also an idea with a long history. Gallagher (2017) observes that "[William] James (1890) discussed extended aspects of self, suggesting that what we call self may include physical pieces of property, such as clothes, homes, and various things that we own, since we identify ourselves with our stuff and perhaps with the cognitive technologies we use, or the cognitive institutions we work within." (p. 63). Clark (2007) has written on soft-selves; specifically, the extent to which selfhood is assembled by a variety of nonconscious external props and artefacts, such that the sense of continuity relies on and emerges from the "de-centralised, distributed, heterogenous vision of the machinery of mind and self" (p.113) consistent with what Dennett (1991) has proposed. Of course, a more fine-grained analysis would also need to consider what constitutes selfhood; for instance, Neisser (1988) distinguishes between five different forms of selfhood: the *ecological self, interpersonal self, extended self, private self, and conceptual self*. Each of these forms relates to extension in different ways.

As the above quote illustrates, Roth is concerned with 'problem solving, calculating, figuring things out' and this in turn is the terrain of *cognitive extension* theorists. Such examples capture paradigm cases in the literature. Nevertheless, Roth's choice of examples (of calculation mainly) situate his view in the space of cognitive processes rather than mental-state extension. Yet Roth's consistent use of the term *extended mind* would imply that he is more interested in extended mental states. Now, one may if this is merely a terminological difference? I would answer in the negative, and say, alongside Pöyhönen (2014) that cognitive-mental distinction amounts to a substantive difference in the extended literature. To draw this out further, I would say that Roth has two options:

1. Roth could side with Clark (2008) in using cognition and mind interchangeably. If Roth endorses this interchangeability, he doesn't make this clear. The consequence of this for Roth's thesis is that there would be less autonomy for Roth's account of the *mental*. Clark's views cognition and the mental as synonymous (Clark, 2008). Clark is interested in cognition in a broad sense, including subpersonal capacities.
2. Roth could run with the line that there *is* a principled difference between cognitive processes and mental processes. In fact, it could possibly help Roth's thesis if he did make a distinction because some of the propositional, belief-style versions of extension arguably have more in common with personal-level phenomena (that's if Roth were to appeal to Drayson's (2012) line that personal-level explanations are horizontal, and more amenable to folk-psychology; whereas subpersonal explanations are vertical)[59] But in this case Roth would need to invoke different examples. This, however, leaves Roth in a strange place as *mind* is arguably stronger than *cognition*.

*1. The personal-subpersonal distinction*

The way in which Roth (2015) refers to the personal-subpersonal distinction invites ambiguity. Drayson (2012, 2014) has noted instances of theorists sliding between definitions: between subpersonal as a state and subpersonal as an (explanatory) level. This confusion is arguably present in Roth's account as well. Roth makes reference to levels over thirty times and yet, confusingly, doxastic and subdoxastic states enter the picture too (and is mentioned twenty

[59] Although recall that, as Westfall (forthcoming) shows, the idea that personal-level explanations can't be vertical and subpersonal-level explanation can't be horizontal is open to serious doubt.

times). For example, in the math example – multiplying 41 x 17 – notice that Roth refers to the *phenomenology* of reaching an answer. This would appear to be orthogonal to subpersonal explanation. So, at times, it is not entirely clear that one knows how Roth is intending to approach the distinction. If Roth means subpersonal *levels*, then he faces the problem of squaring his own rather idiosyncratic view of extended mind with the larger field of cognitive science that is concerned with subpersonal capacities (Rupert, 2018), and, more generally, that there is not a categorical distinction between the explanations for these subpersonal level capacities and the personal level (Bermúdez, 2000).[60] Indeed, it is quite an unusual, and by no means necessary, to claim that extended cognition is not subpersonal (Clark, 2015). As Ward (2012) notes, "[T]here is nothing about the notion of the sub personal that restricts it to the neural in particular." (p. 733). On the other hand, if Roth intends subpersonal to mean *subpersonal unconscious states*, then this is also a problem for his view. Extended cognition is orthogonal to questions of consciousness (Clark, 2008; Chalmers, 2019). Arguments for coupling – whether the original parity and trust-and-glue criteria in the first-wave or the mutual-manipulability arguments in the second-wave (Menary, 2007) – do not require consciousness. Even if the extended belief-states don't require consciousness to be extended because the beliefs are dispositional rather than occurrent. Of course, Roth could say that he is interested in subdoxastic states rather than unconscious states, but, as Drayson (2012; 2014) points out, subdoxastic states are not synonymous with the subpersonal.

### *2. Non-derived content*

A final issue with Roth's account is its reliance upon the non-derived-derived distinction. This distinction, as noted earlier, was proposed by Adams and Aizawa (2001, 2008, 2010). Recall that this is the idea that there is a qualitative difference between conventional signs and original content (with only the latter being brain-bound). Roth is conceding a lot to Adams and Aizawa by using this distinction in his own argument. There is conceptual space to take a different route. There are several options here:

1. Reject the non-derived-derived distinction. [61]

[60] One of the sources of the disagreement is that Roth believes that the personal-level has autonomy from the subpersonal (p. 136, footnote: 8).

[61] Greif, 2017 for an evolutionary argument for how all content is derived from evolutionary processes.

2. Accept it, but question-beggingly assume that non-derived content needs to be brain-based.
3. Even if one concedes that the non-derived-derived distinction exists, and that external resources do *not* have non-derived content; one need not assume that the 'mark of the cognitive' is determined by this distinction.

Beginning with option one, we can reject non-derived-derived distinction (see Clark 2005, 2010); Dennett (1990); Menary, (2010b)). This undercuts Adams and Aizawa's criticism, and thus is good for the extended cognition thesis. It would also be consistent with Allen's (2017) point that there is no mark of the cognitive. This strategy could also be consistent with those in 4-E who argue against the use of representations *tout court*. But insofar as it is worthwhile criticising the distinction, Clark (2005) argues that the existence of intrinsic content is "unclear" (p. 1), that it is "fuzzy and indistinct, and that (to put it bluntly) content is as content does" (p. 4). One example Clark gives is of someone imagining a Venn diagram; the content of the Venn diagram is at once a matter of convention and a bona fide cognitive process. Indeed, this is an example of how cultural practices permeate and transform cognition (see Menary & Gillett, 2017). This means that the conventional and the natural, 'original' content are entangled.

Secondly, there is the possibility of accepting the distinction of non-derived-derived content, *but* interpreting non-derived content so as to include more than brain-bound content (Gallagher 2017). Gallagher's strategy, inspired by Merleau-Ponty's (1945) motor-intentionality thesis, involves looking at how bodily, motoric acts possess a form of non-derived intentionality (p. 80); Gallagher also incorporates insights from the pragmatist tradition in his account. Indeed, according to Gallagher, Adams and Aizawa (2008) have too narrow a view on non-derived content as they only focus on propositional attitudes; however, a case can be made for a form of 'embodied intentionality' that grounds propositional intentionality (p. 2017, p. 62). The upshot is that rather than looking a reductive account that only focuses on brain states, belief can be connected with action tendencies that possess non-derived content (p. 63).

Thirdly, there is no need to assume that the non-derived-derived distinction matters for the mark of the cognitive. Allen (2017), for example, remarks that this distinction has 'no status' within cognitive science, and that the existence of this distinction emerged from introspective

claims with questionable empirical grounding (p. 4234). This is aside from larger questions about whether there even needs to be a mark of the cognitive (Allen, 2017).[62]

The point here is not to argue for a particular one of the above options; instead I am pointing out the conceptual space in which a variety of different responses can be draw upon. To summarise, Roth does not make a convincing case for extended cognition being subpersonal level or nor state. First of all, Roth's (2015) focus on extension at the personal-level seems to be a misreading insofar as it involves ascribing person-level predicates to cognitive processes. Second, extension is subpersonal insofar as it can be examined mechanically and is amenable to vertical explanation. Third, there is nothing to suggest that the derived-non-derived distinction upon which Roth (2015) relies should be accepted; and, alternatively, insofar as it could be accepted, making such a concession would not necessarily mean that the non-derived content needs to remain brain-bound.

### 3.5. Conclusion

The main emphasis in this chapter has been on the subpersonal-personal distinction. It has been proposed that both metacognition and the extended cognition thesis exist at the subpersonal level. To sharpen this point I have argued against Roth's (2015) conceptualisation of the extended mind thesis in relation to the subpersonal-personal distinction.

One thing that may have been clear to the reader is the place of metacognitive normativity. If, as Drayson (2012, 2014) has shown, the subpersonal-personal distinction is *not* a distinction that maps on to a distinction between causes and norms, then what is the status of norms in this account? Recall that normativity is part of what constitutes *mental action* (Proust 2013). Can metacognitive norms operative at the subpersonal-level? And how might this norms play a role in regulating subpersonal metacognitive processes that are extended across brain, body and world? The following chapter will consider these questions.

[62] Relatedly, even insofar as a mark of the cognitive *is* needed, it is not obvious that there needs to be just the *one* mark of the cognitive, rather than a pattern of marks. (Gallagher, 2017, p. 61).

## 4. Metacognitive Norms, Cognitive Action, and Extended Cognition

In the following I will examine how norms relate to metacognition and the extended cognition thesis. In section 4.1 I begin with a consideration of procedural and analytic metacognitive norms and the constitutive role these norms play in cognitive action. In section 4.1.1. I examine cognitive action and normativity. In 4.2 I will examine how metacognitive norms and extended cognition align with Menary and Gillett's (2017) account of cognitive norms and cultural practices. Finally, I examine two case examples that involve metacognitive normativity: one is non-linguistic, procedural and involves a nascent appreciation of consensus (tool-making in the Acheulean-period); and the second example is analytic involving a Vygotskian analysis of inner-speech and metacognition in relation to the norm of *relevance*. This chapter will at once build on ideas encountered in previous chapters while setting up a series of ideas that will be useful in chapter five of this thesis concerning the actual practice of external artefacts in relation to expertise, automaticity and control.

### 4.1.Metacognitive Norms

As we move through the world we are immersed in, and deeply affected by, norms. Norms regulate behaviour in conscious and unconscious ways. Norms also take a variety of forms. Aesthetic norms guide the creation and appreciation of art. Moral, ethical and legal norms guide conduct and the consequences of how we conduct ourselves; and a range of social, epistemic and cognitive norms shape how we interact with one another and the world. Nonetheless, as important as these norms are, I want to focus on a special range of norms that are concerned with metacognition.

Metacognitive norms can be seen as a special class of epistemic norms that, rather than directly acting on the world (i.e., instrumentally), act on and are integrated with the monitoring and control of one's cognitive processes (Proust, 2013). This capacity for norm-sensitivity, and the associated correctness conditions that come with the norms to which the sensitivity applies, varies depending on the context and activity. According to Proust (2013), procedural metacognition's primary, 'overarching norm' is fluency. Fluency is the nonconceptual norm associated with procedural metacognition and is the most foundational to the others. It is shared with non-human animals and pre-linguistic infants. Analytic norms, however, are conceptual, and the possession and use of these conceptual norms presupposes the agent has a language and can metarepresent. These metacognitive norms include: *Accuracy* (a norm corresponding

to a fact in the world where there is one correct answer); *Relevance* in conversation; and *Exhaustivity* in perception, memory and reasoning; for example, when trying to remember everything on a list there is the implicit potential of retrieving more than one needs. For the mathematician working on a proof, a norm of *accuracy* will be more important than *exhaustiveness* (the mathematician wants one correct answer, not a comprehensive series of answers that could include false-positive answers). For the novelist, a norm of *coherence* when telling a good fictional story may be more important than *accuracy* (after all, fiction comes at the expense of accuracy). A group may choose *relevance* over *consensus*. In a group situation - say a discussion about politics - *consensus* may be preferred over *accuracy*. In short, there are various trade-offs between norms that are context-sensitive and activity-dependent (also see Koriat & Levy-Sadot, 1999).

These norms are all linked with fluency. Indeed, unlike the other norms, fluency plays a foundational role at it is at once a norm and a causal-descriptive property (or function) of cognition. Fluency, as Alter and Oppenheimer (2009) point out, is always present."[E]very cognitive task can be described along a continuum from *effortless* to *highly effortful*, which produces a corresponding metacognitive experience that ranges from fluent to disfluent, respectively" (p. 220). They go on to note that a sense of fluency can take place across a range of dimensions: "Fluency experiences arise as a by-product of a wide array of cognitive processes, including but not limited to perception, memory, embodied cognition, linguistic processing, and higher order cognition." (p. 222). This is consistent with the way in which Proust (2013) distinguishes between *perceptual*, *memorial*, and *conceptual* fluency as it applies to different modalities and in conjunction with cognitive processes at varying levels of abstraction. Alter and Oppenheimer (2009) further divide what they term higher-order fluency into five forms: *conceptual fluency*, *diagnostic fluency*, *spatial reasoning fluency*, *imagery fluency*, and *decision fluency*. This is relevant here because this thesis is concerned with how fluency connects with analytic metacognition. Fluency denotes feelings rather than propositions. Fluency is at once objective and subjective; it crosses over from biology and embodiment into norms (Koriat, 2000; Proust, 2013). Fluency and familiarity can pull apart; fluency is the core feeling while familiarity is the interpretation of the core feeling of fluency (Goupil and Proust, 2022; Proust, 2015; also see Koriat, 2000). Familiarity is thus an interpretation of fluency. Indeed, this is evident in the fact that feelings of fluency, including disfluency, can occur when familiarity is not present.

**Table 1**

*Metacognitive norms according to Proust (2013)*

| Metacognitive norm | Description of the metacognitive norm |
|---|---|
| Fluency | Fluency is the property of a stimulus whereby it is processed quickly and adequately. |
| Consensus | Consensus refers to the access of shared information.[63] |
| Exhaustivity | Exhaustivity refers to a subject comprehending all that needs to be comprehended in a perceptual / memory / reasoning task. |
| Coherence | Refers to properties of consistence in demonstrative reasoning. |
| Accuracy | This refers to the property of a claim corresponding to a fact in the world, and independent of the beliefs an agent has about a proposition. |
| Relevance | This refers to a combination of fluency, informativeness, and exhaustivity. In social interaction relevance is used so as to understand a speaker's intention (Sperber & Wilson, 1995). |

### *4.1.1. Cognitive action and normativity:*

Not only do norms merge, as when fluency is conceptually enriched; crucially, norms play a constitutive role in cognitive action. A cognitive action is similar to a bodily action; both involve some form of adjustment and both are a natural kind on Proust's (2013) account; but while a bodily action is only instrumental, a cognitive action is directed toward one's own cognitive processes and is constituted by metacognitive norms.[64] Good examples are those of directed-remembering or perceiving which are captured in our folk-psychology '*trying to*' locutions (e.g. "I'm *trying to* remember / imagine / perceive / reason with"). Moreover, a

[63] Mercier and Sperber (2017) argue that reasoning developed not from solitary use, but rather as an outgrowth of justifying oneself and evaluating others in social contexts and public domains.

[64] There has been controversy in the literature regarding the status of epistemic norms. Dretske (2000), for instance, claims that only instrumental norms of rationality exist. Proust (2013), however, offers what is to my mind a convincing response to Dretske (see Proust, 2013, chapter 7). Due to considerations of space I unfortunately cannot go into detail here; however, the crucial claim Proust advances is that epistemic, metacognitive normativity is separable – and sometimes in conflict with – instrumental normativity.

cognitive action, in its most inclusive form, is defined more precisely by Proust (2013, p. 162, italics added) as:

> Being *motivated to have a goal G realized* → (= causes) *trying* to bring about H [a cognitive act] in order to see G realised by taking advantage of one's cognitive dispositions and *norm-sensitivity* for H reliably producing G.

As can be seen above, norm sensitivity and the goal-motivation are intertwined in cognitive action. More specifically, Proust (2013) emphasises that cognitive actions are regulated in two ways: *instrumental norms* and *epistemic norms*. The instrumental norm is motivational and is concerned with realising the act (goal G above); this norm is primarily concerned with utility (such as the ordering of preferences, and cost-benefit analyses). Epistemic norms are evaluative and and are concerned with fluency, accuracy, and other epistemic norms that feature in metacognitive actions (see Table 1). Further, these metacognitive actions can be prospective ('will I be able to perform the action?') and retrospective ('was the outcome reached?'). So, to reiterate, a cognitive action consists of both *instrumental* and *epistemic* norms (the epistemic norms are required for and are constitutive of action). To illustrate this, Proust puts forward an example, familiar to many, of a supermarket shopper who has forgotten their list. The situation can be functionally decomposed as follows:

i). *Instrumental norm*: there is the goal "remember the items on the shopping list."

ii). *Epistemic norm*: applying the right norm to the act, entailing:

a) *A Prospective* feeling concerning the feasibility of the act (*is it worth searching my memory*?) and epistemic feelings (e.g., a degree of memorial fluency will accompany a cost-benefit ratio of the worth of searching memory). A norm of *exhaustivity* may be followed (even if it comes at the cost of *accuracy*, meaning a few extra food items as might be bought.) *And:*

b) *A Retrospective* feeling: Post-evaluating the act's outcome. *Has the list been reproduced?* The agent is sensitive to the normative requirements here. Feelings of fluency (a proxy for 'accuracy') associated with feedback from the situation are involved in this self-evaluation, and problems can occur when the wrong norm has been used (e.g., if the shopper had used, for example, consensus – what others think should be purchased – rather than the norm of accuracy).

Crucially, the interplay of epistemic feelings – predicting and evaluating whether one is able to realise the goal - underpin the sense of effort or *trying*. This is captured by the experience of fluency (or disfluency) which serves as a sort of *knowing-how*, the bedrock, on which metacognition bottoms out. The agent's norm-sensitivity does not need to involve a concept of what the norm is (the activity and the environment will afford this). Proust (2013) observes that the very fact the list exists to begin with is that it was made in anticipation of the feeling of error for when the agent is in the supermarket with uncertain knowledge of what items need to be purchased. Proust observes that: "Externalising one's metacognitive capacities is a standard way of securing *normative requirements* as well as instrumental success in one's actions." (p. 166, italics added). The environment cues the agent so as to achieve metacognitive outcomes; moreover, and importantly for this thesis, is the idea that these external cues are part of the metacognitive process. More specifically, the cues are implicitly normative. As Proust writes:

> Agents thus rarely need to deliberate about the kind of accepting appropriate to a context, because *the selection is often dictated by the task* or triggered by the motivation for an outcome: at the supermarket counter, the exact opposite change is accepted; when doing maths, an *accurate* answer; at the bus stop, an appropriate waiting time; at a family meeting, a *consensual* conception of a situation. In this variety of contexts, no reflection is needed*:* [65] agents are trained, by prior feedback, to select the proper acceptance. (p. 181, 2013, italics added).

The activity-dependent and context-sensitive nature of metacognition can be appreciated in the above quote. The metacognitive norm of *accuracy*, for example, or the metacognitive norm of consensus inheres in the task itself. As Koriat (1993) empirically demonstrated, analytic metacognition has an engaged, activity-dependent basis. Furthermore, assessing the normativity can be unconscious or with exist with only minimal awareness. Metacognitive norms provide constraints to which we are sensitive and make adjustments; we need not be explicitly aware of metacognitive norms to adjust to them. As Proust notes, just as we do not need to be explicitly aware of the concept of gravity so as to adjust to a change in the force of

[65] In this passage Proust appears to be equating reflection with deliberation; however, while deliberation may entail reflection, reflection does not necessarily entail deliberation.

gravity when underwater, the same is true as far as epistemic norms are concerned (p. 153). It is also important to note that an unreflective action does not necessarily mean there is no monitoring and control (I say more on this in chapter five); it is rather the case that it is not *explicit* (i.e., deliberative) monitoring. Crucially, the agency that is involved here is not explicitly conscious nor is the cognitive action necessarily conceptual. Instead, the metacognitive agency is activity-dependent and context-specific (Proust, 2013). For example, a toddler, when retrieving toys he or she has lent to another, will need to remember *all* the toys she lent. She will not explicitly understand the norm of exhaustivity – i.e., she will not have conceptual awareness of the metacognitive concept of *exhaustivity* – even though she will be implicitly following the norm trying to retrieve her toys (2013, p. 157).

### 4.2.Cultural Practices, Metacognitive Norms and 4-E Cognition

Recall from chapter one that we encountered several waves of the extended cognition thesis. The second wave, as exemplified by Menary (2007, 2010a) and Sutton (2010), emphasises the role of norms and the complementarity between an agent and the artefacts they draw upon. I now want to focus on this and pay special attention to the role of normativity. Menary and Gillett (2017) put forward the following cognitive practices in relation to the way that the body manipulates the environment:

1. Biological Interactions
2. Corrective Practices
3. Epistemic Practices
4. Epistemic Tools and Representational Systems.
   a). Epistemic Tools
   b). Representational Systems

5). Blended Practices.

This series of cognitive practices can be briefly explicated in the following way:

1. *Biological interactions* involve sensorimotor coupling. These involve perception-action cycles. Menary and Gillett (2017) draw on the work of Noë (2004) to advance this point.

2. *Corrective practices* involve correcting a previously held thought. Menary and Gillett characterise these as entailing an 'exploratory inference.' It is associated with language, or at least is elaborated by way of language. Menary and Gillett cite a case study from Vygotsky (1978) whereby a child used private speech (spoken, but self-directed) when using a stick to obtain candy from a cupboard. When evaluating options, the private speech assisted the child in correcting her behaviour to achieve her goal.

3. *Epistemic Practices* are non-pragmatic (i.e., non-instrumental) actions that assist an agent in reaching a goal. These epistemic actions can simplify cognitive goals, allow for probing of the environment, involve 'epistemic diligence' whereby the quality of information in the environment is maintained, and, finally, entail the modification of environments to assist with cognitive tasks. Menary and Gillett draw upon Kirsh and Maglio's (1994) work to support this view.

4. *Epistemic Tools and Representational Systems* refer to extra-neural artefacts (such as computer devices, rulers) that assist in completing a task. Representational systems encompass mathematics, alphabets, and various artefacts.

5. *Blended Practices* entails cognitive practices 1 – 4 interacting. Blended practices can be defined by their cognitive complexity and, as the name hints at, a blending of practices that occur “in cycles of cognitive processing." (p. 75). This can entail hierarchical processing where various processes are operating simultaneously. (See chapter five of this thesis for a demonstration in relation to metacognitive skill, automaticity, and control).

#### *4.2.1. Connecting Menary and Gillet’s account to extended metacognition*

A question that now arises is how these metacognitive norms can be related to Proust's (2013) metacognitive norms and 4-E cognitive processes? In **Table** 2, below, I present these relationships.

**Table 2**

*Cultural practices, metacognition norms, and 4-E cognition processes.*

| Menary and Gillett (2017) | Metacognitive Norms Proust (2013) | 4-E Cognition processes |
|---|---|---|
| 1. Biological interactions | Procedural: Fluency | Embodiment (interoception and proprioception); affect. |
| 2. Corrective practices | Procedural: Fluency, Analytic: Accuracy, Consensus | Embodiment, enactivism, embedded cognition |
| 3. Epistemic practices | Fluency, Accuracy, Consensus, Exhaustiveness | Embodiment, enactivism, embedded cognition |
| 4. Epistemic Tools and representational systems | Fluency, Accuracy, consensus, exhaustiveness, | Embodiment, enactivism embedded cognition, extended cognition |
| 5. Blended Practices | All of the norms: analytical and procedural | A mixture of forms of 4-E cognition |

As can be seen in Table 2, fluency is a metacognitive norm that runs through all of the processes and practices. Fluency serves as a bedrock with which more abstract thoughts interact. Beginning at the earliest level, at sensorimotor practices, Menary and Gillet (2017) note that, phylogenetically, sensorimotor interactions are the most basic, and this level is redeployed so as to be useful for later cultural uses. These interactions involve the coupling and fluency mentioned earlier in this chapter. These interactions can be more or less difficult and effortful. Goupil and Proust (2022) make clear that sensorimotor cycles are not identical to fluency, but they enable it. In addition to perception-action cycles, (2) *corrective* practices, can involve language or at least some form of communication, that merge into (3) epistemic practices, which can include tools (4), and, finally (5) these practices can bind together: here, procedural

and analytic norms mesh with potentially all forms of 4-E cognition. To illustrate this in greater detail, I will now i) examine tool-making and metacognition, and ii) inner-speech and metacognition. In the both case examples I will weave together the considerations of normativity and cultural practice.

### 4.3. Toolmaking and procedural metacognition

Birch (2021) argues that the use of skill-based, norm-guided cognition developed independently of language use.[66] Birch argues that normativity arose through the fluent execution of skills when making tools in the Acheulean period[67] (roughly 1.7mya to 200kya). More precisely, Birch proposes that the stone hand-axe,[68] originally developed by the *Homo erectus* and used by the *Homo heidelbergensis*, presents not only a significant point in human evolution, but also the beginning of normatively-guided cognition. Birch's (2021) view on normative cognition is amenable to Proust's (2013) account of procedural metacognition; more specifically, it illustrates the way that procedural metacognition operates on a non-linguistic, skill-based, non-conceptual format with fluency as the overarching norm. Birch notes:

> Skill leads to *discontent* when the agent falls short of the standard of performance implicitly encoded in the control model. An incorrect adjustment, leading to a mismatch between the predictions of the cognitive control model and the agent's behaviour, *feels wrong* to the agent, independently of (and often temporally prior to) any physical discomfort the error may cause. Skill creates internal pressure to conform to an internalized standard of correct performance. (p. 7).

Here we have fine hand-eye coordination, and the sensorimotor cycles that this coordination entails, being redeployed toward a goal of making a tool (this echoes Menary and Gillet's (2017) focus on how sensorimotor systems are transformed when the agent into is enculturated). Also notice the Birch's use of the word *internalises*. Birch sees "no opposition" (p. 4) between the invoking the existence of internal models and relying upon external models.

[66] In terms of what language may have been present at this time, Sterelny and Planer (2021) claim that erectine hominims (around in the first third of the Pleistocene) had a rich protolanguage with structured signs and 'displaced reference' (i.e., the ability to refer to past events).
[67] Sterelny (2021) thinks it is "vanishingly unlikely" that the first stone tools were Acheulian hand-axes; Oldowan tools predate them (p. 35).
[68] These tools are also known as *cleavers* and as *Acheulean bifaces* (the tool were not always used as axes). Making them involved chipping away at stone in a precise, skilled way.

Indeed, internal and external models can be seen as complementary (I'll have more to say about this soon). In addition to this, we can see how feeling states are what ground the epistemic dissatisfaction. Birch relates this to Rietveld's (2008) work on *situated normativity*. More specifically, people who are skilled exhibit 'directed discontent' toward objects, and this discontent, in turn, motivates action. Importantly, this discontent manifests as a reactive feeling that is intertwined with possessing a standard of what *should* be done to correct for this perceived falling short. For example, Rietveld cites Wittgenstein's example of an architect looking at a doorframe and saying "make it higher, too low!" (p. 980). Implicit in this feeling is a standard that informs the reaction. The standard serves a correcting function that irons out variability and conforms to a consensus view on how something should be done. Birch (2021) notes that:

> Agents who possess a complex motor skill or craft skill possess a well-calibrated cognitive control model that accurately represents those aspects of the causal structure of the situation relevant to the successful execution of the skill; *anticipates* upcoming obstacles and problems; *predicts the flow of sensory feedback* that will occur if skill execution is successful; creates *affective pressure* to respond to mismatches between prediction and performance by adjusting one's technique; and represents *a norm of correct performance* in the pattern of mismatches that trigger affective pressure to make an adjustment. (p. 8).

What Birch is alerting us to is the affective and cognitive aspects of skill. The performance of a skill is not an entirely unconscious, inferential process; it is permeated with affect and normative-guidance. Error is not detected or corrected for in the manner of an unfeeling machine, and nor is it unconnected to the internalisation of social learning and its transmission within and between generations (Sterelny, 2012). Furthermore, while Birch does not refer to metacognition, this does not mean that metacognition is irrelevant here. Instead, as I suggested a moment ago, it can be posited that procedural metacognition plays a role in Birch's example. For the agent there is an experience of fluency that accompanies the performance at a procedural level. When this fluency is disrupted, the agent becomes aware of the sense of disfluency. Metacognitive norms, such as *accuracy* and *consensu*s - regardless of how nascent and inarticulable these norms may be – guide and partly constitute the agent's correction of the error. For example, if the agent making a tool is not conforming to correct standards – perhaps the tool is insufficient sharpened, or sculpted inadequately – then these precise defects are what

the agent is affectively pressured to alter. The norm of accuracy – wherein there is a 'objective' fact in the world, independent of the agent's belief-state – is what normatively guides the agent. Moreover, the norm of consensus merges with accuracy here as the standard of what constitutes an effective tool is not only determined by instrumental success (i.e., how effective the tool is when hunting), but also by socially-determined standards and cumulative cultural knowledge regarding the most effective way to produce a tool and what that tool should look like.

A critic could say that the Acheulean tool example does not really count as metacognition as the action is *world*-directed, and thus merely first-order and responsive to instrumentive concerns, rather than second-order and genuinely metacognitive. But this criticism would amount to a mischaracterisation of what is involved in the tool-making and procedural metacognition. The activity-dependence and world-involving nature of the task is part of what makes it procedural metacognition. The fluctuating sense of fluency, and the sense of *trying*, only makes sense within the context of the activity. As in Birch's description, the adjustments are made by way of not just what is happening with the tool, but also by way of the *affective pressure* that the agent experiences. At this affective level the loop between agent and tool is 'evaluated' and decisions are made about whether to stop or how to continue. It is as much a cognitive task as a constitutively world-involving task.

Lastly, I want to say something more about a comment I made a moment ago regarding *external models*. The models that an agent uses to correct errors need not be entirely internal. Rather, the artefacts – tools and objects – in the world itself can provide information on the correct standards for successfully producing a product. The *affective pressure* the agent experiences is partly in relation to externally available and readily accessible models. Sterelny (2021) notes that the learning does not need to occur by way of high-fidelity *imitation* learning (p. 36) in which the agent enacts the goal-states of another agent by way of metarepresentional capacities such as theory-of-mind. Instead, when cultural information is preserved across generations, the cumulative effect of this stored information results in available cultural resources. These are resources, such as tools, are observed and the processes of constructing them are emulated. Sterelny notes that "Artefacts are templates. Novices can attend to the production sequence – to the expert's products rather than his/her actions". (p. 36). Sterelny, thus means that those learning a skill can pay attention to the *product* and emulate the product rather than learning to make the tools by way of explicit, imitative learning. This learning is further amplified when the agent receives social scaffolding to explore, make errors and correct for them. In effect, insofar as Sterelny accepts high-fidelity cultural learning, he provides two caveats: a) there is no need to posit specific cognitive adaptations for individual learning, and,

b) even if the information flow is noisy, there can still be 'high fidelity' learning "if the incoming generation has the capacity to detect their own errors from signals in the world, or from their elders and the motivation and the support to correct them." (p. 9). Sterelny (2021) speculates that one of the teaching motivations in deep history came from intervening when seeing tool-making done poorly. The sense of intrinsic pride in one's work, although in some sense an individual experience, could have led to alerting others about mistaken ways of sharpening and refining stones and assembling tools. Consequently, collective intentionality and individual learning are not necessary as high-fidelity learning can take many forms. As far as metacognition is concerned, we can see a trajectory, located in deep history, leading from the procedural metacognitive norm of fluency to a nascent metacognitive appreciation of *consensus* that does not rely upon sophisticated metarepresentational abilities such as theory of mind.

### 4.4. Inner-Speech, Metacognition and the Norm of Relevance

We have seen normativity at a non-linguistic, procedural level, in relation to a nascent sense of consensus; now, I intend to look at the metacognitive norm of relevance in relation to inner-speech. It is worth noting again that procedural metacognition is still important here, only now procedural metacognition is elaborated with and enriched by words and concepts.

Vygotsky (1934/1962) presented an influential account of how inner-speech develops by way of internalisation and functions in cognitive performance. [69] Vygotsky (1934/1962) was very clear that – contrary to prominent behaviourists of the time – self-talk is *not* the same as mere *subvocal* speech (i.e., speech without sound). Specifically, inner-speech has unique syntactic and functional properties (Vygotsky, 1934/1962)[1] such that inner-speech is qualitatively different to ordinary speech. In addition, as well as distinguishing his theory from the behaviourists, Vygotsky was also arguing against Piaget, who considered private speech to be merely a "direct expression" of thought as thought takes on a social dimension. In this way, private speech is a mere *epiphenomenal* accompaniment of thought that plays no functional role; it is incomprehensible (even to the child), and disappears with socialisation. In effect, Piaget thought "it has no future." (See Vygotsky, 1934/1962; p. 130). This trajectory is in

[69] It is important to remember that Vygotsky originally intended for internalisation to mean *appropriation*. (See Kirchhoff and Kiverstein (2019) for a critique of internalisation; specifically, their critique of the idea that internalisation occurs without transformation.

striking contrast to Vygotsky's view where, in a converse manner to Piaget's theory, private speech *increases* with socialisation and becomes inner-speech.

According to Vygotsky's developmental view of inner-speech,[70] *inner speech is for oneself*, and *external speech is for others*. What is significant is the sequence in which these develop. The child is in external dialogue with others, beginning preverbally with a parent or caretaker, (see Lock, 1978), then by way of proto-conversations (Trevarthen 1979), before advancing on to more elaborate and complicated speech;[71] and this dialogical speech becomes progressively individualised as the child begins to direct these dialogical practices toward themself by way of private speech.[72] Private speech is still spoken aloud but nonetheless spoken for the benefit alone of the child (often while engaged in a task); and then, eventually, the speech 'goes underground' altogether, thus it is *internalised* (or *appropriated*, to use an alternative, and arguably more appropriate, translation; see Esteban-Guitart, 2014). Furthermore, as mentioned, Vygotsky proposed that inner-speech undergoes a transformation when it is internalised by the individual. First of all, there is an increase of *sense over meaning*; here, personal, idiosyncratic meanings that involve sense take precedence over conventional word meanings. Secondly, there is agglutination, where hybrid words form. Thirdly, there is an *infusion of sense*, so that one can be said to be thinking in "pure meaning." In additionally to this, there is a predicative structure without subject. Inner-speech thus has a curious phenomenology "speech almost without words." (p. 145).

One might ask how Vygotsky's theories have held up since they were first introduced in the 1930s. Contemporary empirical research is generally in favour of the theory (see Alderson-Day & Fernyhough, 2015; Alderson-Day et al. 2018).[73] It has been found that private speech tends to develop around two to three years of age, and occurs regularly around three to eight years; however, there is evidence that it continues to be useful past this point, but the benefits are mostly experienced under the age of five (Alderson-Day & Fernyhough, 2015). Wilkinson

[70] A brief word on terminology. Vygotsky is translated as calling inner speech *endophasy*, (although, for clarity, I won't use that term here). Similarly, *egocentric* speech was the original term for what is now *private speech.*

[71] Lock (1980) investigated the early, pre-verbal communicative gestures between child and parent; for instance, how action becomes gesture by way of ontogenetic ritualisation during interaction. For example, in infancy the 'pick me up' gesture begins as a functional climbing, arms over head action. Over repeated interactions, the parent learns to anticipate the child's action before the action is completed, and in turn the infant anticipates that the full action is not necessary to successfully communicate a request, hence the more abbreviated gesture. A similar account was put forward by Vygotsky that pointing is an abbreviated grasping action (see Tomasello 1999).

[72] Alderson-Day and Fernyhough (2015) point out methodological difficulties. For example, the presence of private speech could indicate the absence of inner speech; and, conversely, someone who does not display much private speech could still possess inner-speech. In consequence, the presence or absence of *private* speech should not necessarily be viewed as an index for the presence or absence of *inner* speech.

and Fernyhough (2018) present an array of neuropsychological findings, including the role of motoric processes in inner speech (i.e., the activation of speech and face muscles during inner speech). Additionally, there is fMRI evidence that supports the claim that inner speech is a genuinely productive speech act rather than a mere imaginative re-enactment of speech. Relatedly, and in terms of how metacognition fits with other cognitive processes, Vygotsky's contention that inner-speech is different from working memory has also been supported. Specifically, Vygotsky (1934) notes that inner speech is not equivalent to "verbal memory," because memory forms only one component of inner-speech (p. 130). Contemporary research supports this view; more specifically, while working memory is concerned with the function of verbal rehearsal – for example, the phonological loop which holds auditory information for short time periods - inner-speech is more concerned with how inner-speech can be used to enhance and transform cognitive capacities (Alderson-Day & Fernyhough, 2015). Arguably, inner speech is larger than the operation of the phonological loop. Finally, Gauker (2018) inner speech is different from auditory perception (consistent with Vygotsky's contention that inner-speech is, contrary to the behaviourists, not merely *sub-vocal talk*).[74]

What is inner-speech for? More specifically, how can other cognitive functions be augmented and transformed? It has been found that people tend to use 'inner speech' when encountering a problem (Sutton et al. 2011).[75] This is in line with original work of Vygotsky (1934) and more recently Fernyhough (2004). Indeed, inner-speech presents as condensed by default, but it *expands* when encountering stress and cognitive challenges (Fernyhough, 2004).[76] Now, what is of particular interest for this thesis is the functional role of inner-speech and how it emerges from cultural practice; specifically, how inner-speech operates when confronted with a *meta*cognitive problem, and how its existence reflects dyadic, communicate patterns in which the norm of *relevance* is central. In the next section I will briefly look at the principle of relevance as formulated by Sperber and Wilson (1986), and this will then be connected with the metacognitive norm of relevance (section 4.4.2).

### *4.4.1. Relevance and Speech*

[74] Hurlbert et al. (2013) note that studies show people tend to differentiate between *inner-speaking* (active) and *inner-hearing* (passive) Carruthers (2018) notes that this distinction may be related to the extent that a person is attentive toward somatosensory activation of the inner speech (active) or not (passive).

[75] Sometimes research does not differentiate between overt and covert (private) speech; this can be a methodological problem when reviewing studies (see Alderson-Day & Fernyhough, 2015).

[76] The most extreme version of this is *hyperreflexivity,* which is an exaggerated and alienating form of self-consciousness sometimes present in psychopathology (Sass, 1992).

To begin, let's return to the idea that Proust (2014) suggests, namely that *relevance* is the primary norm of conversation. The norm of relevance often takes place within the context of communication and all of the social-pragmatic constraints that a communication act entails. To this end, Sperber and Wilson's (1986) influential account of relevance is of primary importance here. Specifically, their account of communicative intentions builds on Paul Grice's (1975) [77] influential work on the co-operative character of conversation, and the regularities (or 'maxims') that conversations entail as they unfold between sender and receiver. The Gricean maxim that Sperber and Wilson focus on, and make central to their proposal, is *relevance.*

> *Principle of relevance*: Every act of ostensive communication communicates the presumption of its own relevance.

Briefly, to unpack this a little, the key idea here is that communication has a two-fold structure in which there is (1) the information to be communicated; and (2) a meta level in which the first-level of information is being intentionally communicated. The idea is that communication involves making something *relevant* to another person (p. 50). In short, there is an implicit sense that *I am trying to show you something*. Whether it turns out to be relevant is somewhat beside the point. What is important is that there is a background of relevance to which the communication refers; just like an assertion carries an implicit 'guarantee of truth' (p. 49); such an utterance presupposes relevance. Relevance can be viewed in the following exchange:

(1) *Peter*: Do you want some coffee?
*Mary*: Coffee would keep me awake. (p. 34).

Even though Mary's answer is not, strictly speaking, relevant (her answer does not *directly* respond to Peter's question), her answer is nonetheless relevant inasmuch as it provides an *implied* answer to Peter's question. So, rather than violating the principle of *relevance*, Mary is in practice fulfilling it. Indeed, implicit in the exchange is the fact that if Peter is working on the assumption (2, see below), then Peter can infer that Mary does not want coffee (3).[78]

[77] There are two main ways in which Wilson and Sperber's account can be distinguished from Grice's (1975) account. I). Grice was concerned with pragmatic inferences regarding implicit communication, whereas Sperber and Wilson (1986) believe that the explicit inferences are also important. ii). Utterances, by their very existence, presume their own relevance, even when a person is not aware of this norm of relevance; while Grice's account did incorporate relevance (it featured as one of his nine conversational maxims), but Sperber and Wilson see relevance as essential to communication as doing the explanatory work of the other eight maxims.
[78] This is a case of non-demonstrative inference (i.e., probabilistic rather than deductive, so defeasible).

(2) Mary does not want to stay awake

(3) Mary does not want coffee. (p. 35).

Grice (1975) termed these added assumptions and conclusions *implicatures*; this is the implied meaning in an utterance where context helps to determine conversational intent (rather than the meaning being encoded in the utterance). The assumption of (2) is implicated in what Mary uttered in (1), and thus conclusion (3) can be understood by Peter as Mary's implied answer. On Sperber and Wilson's (1986) view, utterances, by their very existence, presume their own relevance, such that listeners expect speakers to be relevant, and speakers expect that the listeners will detect relevance in the intended meaning of the spoken words. There is thus a mutual expectation of relevance; and if this presumption of relevance appears to be violated (perhaps the answer is logically irrelevant), the listener will alight upon a pragmatically relevant meaning. Given that many utterances *under*determine what can be logically inferred from them (i.e., the meaning of the utterances are consistent with a range of possible inferences), the presumption of relevance helps to narrow down what is implied by excluding other logically inferable but socially meaningless inferences. For example, when Mary initially states that coffee would keep her awake (1), this is consistent with Mary implying that (4) her eyes are open when she is awake, and that (5) coffee would help them to remain open. These assumptions form the background context in which the exchange unfolds, but are *not* relevant to the conversation. Mary is evidently wanting Peter to infer more than the effect of coffee on her eyes.[79] She is wanting to convey that she doesn't want coffee, and if Peter works on the assumption that Mary is trying to be relevant, that her utterance is presupposes its own relevance, then he can infer Mary's intended utterance as a 'no.' The other logically consistent, but pragmatically unimportant, inferences recede into the background. Moreover, even if the converse were true - if Mary *did* want coffee (perhaps Peter knows she has an exam the following day and needs to study late into the night) – then the point still holds that Peter is using the presumptive relevance of Mary's statement so as to not take her literally (and he can then discount a host of literal interpretations).

Sperber and Wilson are not only focussed on conversation; they are making a larger point about cognition itself: "Human cognitive processes, we argue, are geared to achieving the

[79] Another well known, and menacing, example, is that the of the stereotypical gangster who says, "Nice house you've got there." The intended effect of the utterance can be taken as meaning more than a simple comment on the quality of the house.

greatest possible cognitive effect for the smallest possible cognitive effort" (p. vii).[80] But there is a trade-off. Too much information and the processing demands will be too high. Conversely, too little information would be easier to process, but would come at the expense of offering sufficient relevance for the receiver to infer the sender's intended meaning. Sperber and Wilson make it clear that humans seek out relevance; humans are attuned to the *cognitive environment* of the other, hence they have mutual presuppositions about what the other will find relevant, how the environment will change that person's ideas, and also the implied premise in an exchange of which no explicit reference needs to be made.

The idea is that the shared cognitive environment - evident in instances of joint-attention, shared intentionality, and the communication that arises from the mutualism inherent in interaction - provides humans with the unique skill of sharing goals and intentions (Tomasello, 1999, 2014, 2019). Vygotsky (1934) also considered the importance of mutualism for social interaction. Much is presupposed when interacting face-to-face as there is a common ground between interlocutors. There is a presumption of linguistic relevance that exists between interlocutors such that short, efficient answers are often adopted (even though these answers can lead to misunderstandings of reference). (p. 139).[81]

> Pure predication occurs in external speech in two cases: either as an answer or when the subject of the sentence is known beforehand to all concerned. The answer to "Would you like a cup of tea?" is never "No, I don't want a cup of tea," but a simple "No." Obviously, such a sentence is possible only because its subject is tacitly understood by both parties. To "Has your brother read this book?" no one ever replies, "Yes, my brother has read this book." The answer is a short "Yes, he has." Now let us imagine that several people are waiting for a bus. No one will say, on seeing the bus approach, "The bus for which we are waiting is coming." The sentence is likely to be an abbreviated "Coming," or some such expression, because the subject is plain from the situation. (p. 139)

[80] This relates to the idea that cognition is time-pressured. See Clark (1997).
[81] This is common in adult mindreading where, in situations of time-pressure and uncertainty, egocentric epistemic bias (that others think like oneself) functions heuristically, and this of course can lead to misunderstandings. Correcting for one's egocentric bias can be cognitively effortful. See Apperly (2011, p. 89).

Vygotsky theorised that this mutualism carried into inner-speech: "the "mutual" perception is always there, in absolute form; therefore, a practically wordless "communication" of even the most complicated thoughts is the rule." (p. 145). This has implications for metacognition.

### *4.4.2. Metacognition, Inner-speech and Relevance*

Vygotsky (1934) pointed out that inner-speech has about it an "extreme, elliptical economy' (p. 45). As was mentioned earlier, inner-speech is fragmented, condensed, and abbreviated, but can also expand (Fernyhough, 2004). This is consistent with Proust (2007) when she notes that we don't think in terms of elaborate sentences. Inner-speech, too, has this low effort requirement, and can be viewed as a form of internal conversation, in which there is an 'economy' of words and syntax and an emphasis on semantics and relevance; the difference is that the communicator and receiver, rather than being two distinct individuals, are in fact the same person, and where inner speech becomes progressively more abbreviated. The norm of *relevance* (which combines informativeness and fluency) is central. These conversational dynamics are appropriated by the individual; and just as Vygotsky theorised, the content – including the public character of language – and the turn-taking dynamics of conversation are thus 'internalised.' But this is not all that happens. The *metacognitive character* of conversation *is also 'internalised.'* Norms – such as the norm of relevance – are thus embedded in the structure of metacognition, just as they are embodied in the turn-taking of internalised speech. Following Fernyhough (2004), inner-speech can expand and contract given the level of stress and the obstacle.

There is a need for low-effort, high-effect communication (i.e., a form of communication where relevant information is conveyed with the least amount of effort); in the case of inner speech, it is with oneself. But does this communication with oneself presuppose that we do or don't know our own communicative intentions? Views diverge here. Geurts' (2018), claims that inner speech is genuinely dyadic; inner-speech involves commitments: promises, statements (that commit oneself to current or possible state of affairs), questions, and orders.[82] These commitments are made to oneself in a way that resembles the commitments we make to others; however, Geurts claims that we know our own communicative intentions, and thus Deamer (2021), on the other hand, claims that we don't necessarily know our communicative

[82] Sutton (2007) observes the use of linguistic, 'instructional nudges' that can take place in skilled performance "stabilise the cognitive flow just enough to help us reorient it" (p. 774). (Also see, Sutton et al., 2011).

intentions when engaging in inner speech. We don't necessarily know our mental states directly until they are manifest in much the same way that we don't the mental states of others until they are manifest to us. Instead, there is an interpretative element involved. While Deamer allows for the commitment-function that Geurts argues for, she doesn't believe we know our communicative intentions. Similarly, according to Carruthers (2018), the communicative intention is tacit (i.e., it need not be explicit).[83] Moreover, while mental state attribution to self and other can be unreliable, the interpretation of semantic content, including its *relevance*, is reliable. Indeed, the reliability of semantic content, and the way this content is infused with meaning, is consistent with Vygotsky's characterisation of the inner-speech.

Readers will recall that earlier in the thesis (1.7.) I presented the evaluativist (i.e., non-conceptual, epistemic-feeling based) conception of metacognition, as presented by Koriat (2000), Proust (2013) and others, with the metarepresentational, attributivist account, one of the leading proponents of which is Carruthers (2011). This leads to a question: How could one continue incorporate the more attributivist (i.e., conceptual) approach to inner-speech advanced by Carruthers (2018), where communicative intentions are not known immediately, with a evaluativist conception of metacognition? I think it is worth recalling that on Proust's account (2007, 2013) there are procedural and analytic versions of metacognition. I propose that Carruthers' (2018) – and Deamer's (2021) – accounts, which I'm highly sympathetic towards, can be accommodated within what Proust calls analytic metacognition, which typically involves enculturation and participation in a public symbol system[84] (Proust, 2007; also see Bermúdez, 2003, chapter 8). The implication of this is that, contrary to Carruthers (2018), a greater space can be made for procedural metacognition. Proust (2007) views procedural metacognition as having a transparency and directness that metarepresentational (mindreading-style) metacognition lacks. While Carruthers (2018) restricts the sense of self-transparency to a variant of inner-speech that is synonymous with mindreading,[85] and rejects procedural metacognition, I think it is important to leave space for procedural metacognition.

Procedural metacognition, with its associated epistemic feelings and inflexible control structure, is experienced in a way that supplements the interpretive, stabilising functions of inner-speech. This interplay between the two is instructive. There is a sense of knowing and

[83] This is consistent with the idea that cognitively significant inner speech may not have auditory character (Gauker, 2018, p. 57).
[84] Most obviously, a language; but public symbol systems can include other systems of representation with recursive properties, such as mathematics, which can also figure in inner-speech.
[85] Carruthers (2018) does acknowledge that we do possess a sense of transparency to ourselves; however, this sense of the mind as being "infallibly self-presenting" is mostly a user-illusion, installed in the operation of mind-reading in order to facilitate its smooth functioning (p. 48).

transparency at the first recursive level (Proust, 2007) – I can know, believe, be aware of, feel, doubt – that is separate from metarepresentation. We can become aware of troubling mismatches that evoke feelings of disfluency, or, sometimes feelings of fluency that are tempered by semantic knowledge that something shouldn't be so easy, or shouldn't be a particular way. Inner-speech, like metacognition can be prospective and retrospective; it can provide a stabilising function, especially when encountering a cognitive challenge (see chapter five of this thesis). And while self-talk is not always metacognitive – it also consists of reactive, merely world-directed responses – it can entail a metacognitive aspect when dealing with obstacles, where there is some form of error and estrangement and need for recovery. Moreover - as we saw in chapter two of this thesis, and as Proust (2013) emphasises - linguistic redescription can enrich and replace sensitivity to epistemic fluency with the propositional, analytic correction that language-use facilitates.

Inner speech is a genuinely productive activity rather than merely recreative (Wilkinson & Fernyhough, 2018). The function of inner-speech is nicely captured by E. M. Forster's line "How do I know what I think till I see what I say?" (1927, p. 152). Our communicative intentions are not always transparent to ourselves. Instead, there is an unfolding of communicative intent in which the intentions becomes clear, and inner speech helps to facilitate this. The conception of cognition as decentralised is consistent with Dennett (1991) such that there is no 'central meaner,' no core, no 'Cartesian theatre,' where everything comes together. Indeed, if one follows Clark's (1996, 1997) line that language 'freezes' thought, and offers stable representations, then we can see how language offers a way of making sense of the multiple streams of perception and cognition. The ability to use language, to amplify and augment concepts and parse events and particulars into memories, in combination with epistemic feelings, increases the ability to evaluate one's own diverse, distributed cognitive subsystems.

### *4.4.3. Inner-speech and extension:*

It may seem paradoxical to talk of extension in relation to inner-speech. Is inner speech not an account, *par excellence*, of an internal-process? What could be more internal than inner-speech? Note that my claim here is that inner-speech is occurrent and not extended. Nonetheless, this should not be taken to imply that an externalist account cannot be given for it. I will go through several reasons as to why inner-speech can be viewed in an externalist light.

First of all, inner-speech can be used when doing a task that *can* be extended. In effect, the inner-speech is part of an overall task that is extended. (Recall that one of Vygotsky's insights was that inner-speech takes place when problem-solving or, more generally, when encountering some form of challenge or disruption). Thus, inner-speech can thus facilitate tool use and problem-solving. Indeed, as Shapiro (2019) has noted, many of the extended claims do not postulate that every part of the cognitive task needs to be realised externally for the *overall task* to count as external.

Secondly, the timescales at play with the development of inner-speech are of course on the level of months and years. Indeed, Kirchoff and Kiverstein (2019) gesture toward Hutchins (1995b) and Vygotsky (1978) in offering an account that examines not just the synchronic, but also the diachronic scales. Also recall how Kirchhoff and Kiverstein (2019) use this idea of diachronicity when neutralising Adams and Aizawa's (2008) causal-constitution objection, see section 1.6.).

### 4.5.Subpersonal norms?

Some philosophers[86] claim that normativity is personal-level and not subpersonal, and that the distinction between the normative and the non-normative defines the distinction between the personal and sub-personal (McDowell, 1994). This maps onto Sellars (1956) distinction between the personal located in what Sellar's refers to a normative *space of reasons* and the subpersonal, as it exists in a non-normative *space of causes.* Nonetheless, it is arguably mistaken to associate the personal-subpersonal and the normative-nonnormative in this way (Drayson, 2012). How do the norms reach down into the sub-personal components of cognition?

Proust (2013) speculates on whether normativity can work at the subpersonal level in contrast to the tendency of reducing norms to non-normative properties (p. 297). Proust cites Shea's (2012) case for seeing the dopaminergic system in the brain as carrying normative information regarding the expected reward of a given decision. Furthermore, Proust suggests that "cognitive systems need to respect epistemic norms at all levels to be viable." (p. 298). Clark (2015) has claimed that precision weighting – as it takes place in an active inference account – is "essentially meta-cognitive" (p. 3768); as we saw in section 1.8.1. of this thesis, this is not without problems. Boone and Piccinini (2016) warn that cognitive mechanisms

[86] These philosophers are often associated with the so-called 'Pittsburgh school.'

should not be reduced to mere brain-events; instead, cognitive mechanisms operate at multiple levels (see section 2.1. in this thesis). Furthermore, while we may follow Zadwidzki (2021) in viewing metacognition as operating along a continuum from non-metacognitive to metacognitive such that there is no bright-line, anything as narrow as what Clark is claiming as metacognitive will only be *minimally* metacognitive. Nevertheless, what Clark is hinting at is the extent to which norms can run deep into the subpersonal-level terrain.

Brandom (1994) has noted that there are "norms all the way down," (p. 61) although Brandom's emphasis is on social practices and actions. Gallagher (2017) extends Brandom's idea by looking at the normativity that exists in gestures. Meanwhile, Miłkowski (2010) proposes that epistemology should adopt the stance of cognitive science, and further, suggests that epistemology can be translated into cognitive science across subpersonal and personal levels and interactions among cognitive sub-systems. Miłkowski cites as an example the way Wheeler (2005) translated some of Heidegger's philosophical work into cognitive theories on embeddedness.

Thompson (2007) notes that normativity can exist at a basic level, and uses as an example the way in which *E. coli* swim towards sucrose. This action suggests, at a minimum, a basic level of reactivity to norms of correctness (p. 74). The central idea is that norm-sensitivity can occur at the subpersonal level; the idea that basic normativity exists at all levels of life is an idea shared by prominent enactivists (see Di Paolo et al., 2017). Nonetheless, whether these can classified as metacognitive norms is an open question.

Menary and Gillett (2017) view cultural practices as deeply normative, and as being subpersonal processes too. Indeed, on their view, these cultural practices span the social, the individual, and the subpersonal. Moreover, the public character of these processes ensure that over larger timescales there is "innovative alteration, expansion, and even contraction." (p. 72). For example, across deep evolutionary time, the metacognitive norm of *relevance* expanded and altered from earlier precursors; for example, Tomasello (2014) claims that joint-intentionality first existed in Homo heidelbergensis four-hundred-thousand years ago. On a smaller scale, at the more modest level of an individual life, we can see the expansion and contraction of the development and function of inner-speech; more specifically, the way people idiosyncratically appropriate the available linguistic resources and norms that pre-exist and will outlive the span of a human life. The precursors of metacognitive arguably emerge from the depths of evolutionary time, are embedded in cultural practices, and reach down into the subpersonal level of an individual.

### 4.6.Conclusion

I have so far sketched a way in which metacognition and the 4-E literature can be brought together via epistemic norms during (skilled) action. This is consistent with second-wave extended cognition (Menary, 2010). We are not just beholden to norms; rather we enact norms through skilled behaviour and metacognitive practices. For norms to be integrated into a cognitive system it is not a simple matter of memorising or being embedded in an environment that is rich with cognitive scaffolding; rather, cognitive action itself is constituted by epistemic norms. The norm of fluency, at once a descriptive-causal (i.e., functional) process and a epistemic-norm, finds itself distributed at various levels of cognitive activity, and how and when it binds with epistemic norms (such as accuracy and coherence). Finally, I have explored the idea that norms can exist at the subpersonal-level.

## 5. Metacognitive Skill and Recovery from Error

I have now highlighted that both procedural metacognition and extended cognition involve a mixture of non-conceptual and conceptual content with an analogical representational format (chapter 2), are subpersonal-level phenomena (chapter 3), and have explored how norms are integrated with metacognition (chapter 4). Now it is time to see how these operate in practice when encountering an obstacle. More precisely, when performing a task there is an interplay between automaticity and control, and this interplay has particular salience when there is some form of breakdown in one's relationship with a cognitive task. Moreover, while high levels of automaticity and control are not necessarily synonymous with an extended account of expertise, these capacities can be viewed as subpersonal capacities that assist in underpinning expertise. In this way, we can appreciate the way in which claim one of this thesis – that procedural metacognition and extended cognition are subpersonal-level phenomena – with claim three – that skilled metacognition can partially extend when appropriately coupled with the environment.

Here is the plan for this chapter. I will begin with a short review of individual differences in 5.1, then I will look automaticity and control; this focus on automaticity and control will be narrowed down even further regarding meshed cognition; next, I will look at a case example of how error is connected with metacognition and varieties of skilled recovery from error. Last of all, I will look at cognitive obstacles and recovery. This chapter builds on the previous chapters so as to present a framework for determining the occurrence of metacognitive extension.

### 5.1.Metacognition, Skill, and Individuality

There are differences in the metacognitive profiles of people. This, however, is separate to intelligence (Fleming, 2021). To approach this question of differences, I would like to take a different approach to what is see as the canonical study of individual differences in the literature: namely that of Dunning-Kruger effect (Dunning-Kruger, 1999). The effect is that "those with limited knowledge in a domain suffer a dual burden: Not only do they reach mistaken conclusions and make regrettable errors, but their incompetence robs them of their ability to realize it." (p. 1132). However, the empirical support for the Dunning-Kruger effect is problematic (Gignac & Zajenkowski, 2020; McIntosh et al., 2019). Some of the Dunning-Kruger effects can be accounted for by *regression to the mean* and the *better-than-average*

bias. Rozenblit and Keil (2002) present a phenomenon separate from the Dunning-Kruger effect, namely the 'illusion of explanatory depth.' The domain for this is in explanations rather than procedures or narratives.[87] One interpretation Rozenblit and Keil offer is that people remember theoretical relations in a skeletal, 'highly sparse' manner that is heavily reliant on environmental representations, and that this more sparse manner of understanding, although at times illusory, can be adaptive and efficient.[88]

I would like to focus on how differences in skill and enculturation change how metacognition operates in practice (Heyes et al., 2020; Kim et al., 2018; Menary & Kirchhoff, 2014), and, as we shall see, whether or not metacognition extends in particular situations. The notion of individual differences,[89] and the attendant fine-grained differences in cognitive profiles, is also relevant regarding the extended cognition literature. Sutton (2010), for instance, criticises the first wave parity principle:

> … Parity [principle] leaves no obvious space for investigating individual differences in relation to EM [extended mind and cognition], because it asks us to focus on generic features of cognitive states and processes, whether in the world or in the head. Yet *we often want to understand the specificities of particular embodied subjects*; just why and how one system – such as a particular embodied agent of one kind of another – can move between a variety of different artifacts (p. 199, italics added).

Of course there are also other ways that individual differences are relevant here; inequality, with the resultant disparities relating to possession of and access to particular resources, is undoubtedly a factor too. That aside, Sutton's criticism taps into longstanding debates regarding course-grained and fine-grained functionalism (see Clark, 2008 p. 94 for more details). One overarching lesson to draw from Sutton's criticism is that aspects of individual practice and skill matter. Indeed, Sutton (2010) makes it clear that there need not be any opposition between externalism and a focus on the specificities of the individual; if anything, externalist accounts can be strengthened when fine-grained examinations of individual variation are centred in the analysis rather being obscured or ignored. Furthermore, I believe

[87] Rozenblat and Keil (2002) also note that the effect cannot be accounted for by the complexity of the explanation nor its familiarity. Some of the factors that appear to account for the effect include the degree of visibility of parts of the mechanism.
[88] Indeed, this interpretation is consistent with Brooks' (1991) line that "explicit representations and models of the world simply get in the way. It turns out to be better to use the world as its own model" (p. 140).
[89] Talking about individual differences does not commit one to essentialist or genetic ideas about human ability.

that Sutton's point can be reinforced if we appeal to Cohen's (1994) distinction between *individualism* and *individuality*. While individualism is associated with methodological individualism[90] and, for our purposes, a variant of cognitivism that is too dismissive toward the environment; *individuality*, however, is concerned with the idiosyncrasies and subjectivity of the individual agent. As we will see below, this individuality is bound up with the experience, skill, and automaticity of the agent. Furthermore, this *individuality* is by no means trivial; not only do *the specificities of the embodied subject* (to borrow Sutton's line) affect how metacognitively successful the agent can be, but this specificity also affects the extent to which metacognitive control integrates with the tools – and the environment – and can thus be said to extend (as I will explore in section 5.3. – 5.5.).

## 5.2.Metacognitive Skill and Automaticity

Skilled behaviour entails automaticity, metacognition, and knowledge (Logan, 1985). Furthermore, although it is something of a truism that sustained practice leads to the development of both automaticity and skill (Montero, 2016), it is less clear what constitutes automaticity. More specifically, how does automaticity relate to *awareness*, *intention*, *attention* and *control*? Furthermore, how does automaticity relate to metacognition? Logan rightly observes that the concept of automaticity has a long history in psychology and has been "handed down from ordinary language." (p. 375).[91]. One problem Logan notes, citing Newell (1973), is the historical tendency within psychology to form conceptual binaries. Automaticity is no exception as it has often been pitted against control. A classic characterisation in literature by Schneider and Schiffrin (1977) is that "the activation of a sequence of nodes that (a) nearly

[90] In addition, Cohen (1994) points out that *individuality* can be contrasted with the *individualism* that inheres in some political ideologies.

[91] The way in which automaticity has been the preserve of folk-psychological conceptions with currency in everyday talk *and* has been a phenomenon of scientific interest can be related to the way in which scientific psychology has attempted, with mixed results, to differentiate itself and its concepts from folk-psychology (see Danziger, 1997, p. 52; also, see footnote 39 in this thesis). Although, as Danziger points out, that a psychological concept exists as a folk-psychological entity does not necessitate mapping on to a real phenomenon. To this end, Danziger draws on Frege's distinction between *sense* and *reference* when considering psychological phenomena. *Sense* is concerned with practices of naming and categorising; meanwhile, *reference* is a phenomenon of "sufficient distinctiveness and stability to warrant giving it a name." (p. 6). Sense and reference can potentially co-occur *or* pull apart. For example, to use an example relevant to this thesis, 'cognition' could be real (i.e., have reference), but have only weak sense (i.e., our ways of understanding cognition, and conceptions we attempt to map upon it, could be wrong such that cognition is *not* what we think it is); conversely, it could be that 'cognition' has sense within a research community but no reference in the *real* world. The map, in other words, doesn't necessarily reflect the territory (and sometimes there is just territory without a map, or a map without a territory). Ideally, the map and the territory will correspond; it is this correspondence is what researchers aspire toward, but this is not a given.

always becomes active in response to a particular input configuration and (b) the sequence is activated automatically without the necessity of active *control* or *attention* by the subject." (p. 2, italics added). This view of automaticity has been heavily criticised in recent years (Fridland, 2017; Montero 2010, 2016; Sutton et al, 2011). More precisely, Bargh (1994) has claimed that none of the four aspects typically invoked to characterise automaticity – *unawareness*; *unintentionality*; *inattention*; and *uncontrollability* – are strictly necessary. I will briefly look at each of these in turn with examples relevant for cognitive extension and metacognition:

*i). Automaticity and awareness*

Dreyfus and Dreyfus (1986) theorised that skilled behaviour amounts to a form of absorbed and 'skilful coping.' Dreyfus (2007) has likened this absorbed coping to a pilot guided by a radio beacon which only gives a warning signal when the plane goes off course. More concretely, by way of the skilful absorption that comes with being engaged in a task, we are, 'mindless.' This idea has traditionally been presented in the context of expertise whereby the expert, so engaged and automated in their task, becomes at one with the task and thus mindless (Dreyfus, 2007).

Many phenomenologists reject this mindless view and emphasise prereflective awareness as existing even during these supposed mindless tasks (Gallagher, 2021; Gallagher & Zahavi, 2021, Zahavi, 2005, 2014). *Prereflective awareness*[92] is a minimal form of awareness that is present during both ordinary life experience and skilled action. Prereflective self-awareness is both constitutive of and a condition of possibility for higher reflection. It is part of the very structure of subjectivity (Zahavi, 2005). Zahavi (2013) argues that the Dreyfus-style of 'mindless coping' leaves out the crucial element of first-person experience that a subject possesses. Similarly, flow states, as popularly characterised Csikszentmihalyi (1990), have sometimes been appealed to when making a case for the mindlessness of automaticity; this is where experts and performers are so enmeshed in their performances, so deeply lost, as it were, in the dynamics of activity that they enter a style of Dreyfus-style 'mindlessness.' But, as Andrada (2021) points out, on closer reader, even Csikzentmihalyi made space for some forms of awareness during these flow-states. Relatedly, Sutton et al. (2011) have noted that it is most probably confabulation that accounts for the supposed cases of 'mindlessness.' Further, this

[73] Prereflective self-awareness involves i). temporality and ii). builds on work in the phenomenological and existentialist tradition by Husserl, Sartre, and Merleau-Ponty (Zahavi, 2005).

notion of mindlessness may underly some of the cache that the idea of 'paralysis of analysis' has where overthinking is recognised as interfering with performance (Beilock et al., 2002, 2004). But, as Sutton et al. (2011) point out, there are problems with merely taking experts at their word.[93] In short, experts do not necessarily have the ability to articulate how and what they do during the intensity of a performance. Further, Montero (2016) notes that the participants in Beilock et al.'s studies were not were true experts; and Montero makes an argument for the importance of conscious control when performing (Also see Toner et al., 2022 for a wealth of evidence on conscious control and athletic performance). Finally, Kirsh (2004) observes that fast action does not necessarily equate to unconscious action; for example, players in a game will consciously scan using saccadic eye-movements for relevant cues.

*ii). Automaticity and intention.*

Fridland (2017) points out that intentions are not necessarily separable from automaticity, and invokes examples of a daily routines where automaticity and intentions are closely coupled. Or consider Clark (2007) on the planned generation of creative thoughts: "[It] should [not] come as a surprise to artists and scientists, who are often painfully aware that the bulk of their own (intentional, owned, self-expressing) creative activity flows from subterranean and nonconscious sources." (Clark, 2007, p. 110). Indeed, creative practices are often about setting up oneself in the right creative environment. As the daily practices of many writers and scientists attest (Currey, 2013), the habits and highly idiosyncratic environments can be geared to the sort of routines that will lead to the development of a creative work (a particular time of day, a particular work-station, an ambling stroll through a particular slice of wilderness, a particular style of coffee etc.). The conditions are set and fine-tuned so as to allow for spontaneity to occur within particular, and sometimes even quite predictable, bounds. While elements of these tasks could be seen as unintentional components in the larger scheme of creative processes, the scheme itself offers a form of unity and intention.

Another example is the way in which environments are structured so as to avoid risk. Environments are permeated with intentions: prior intentions can inform the epistemic structure of the environment. For example, in a vet clinic euthanasia drugs are often coloured

[93] Gallagher (2021) also offers grounds for scepticism regarding the claims of expert meditators who claim to access mindless, self-less states about which they then report. Gallagher notes that meditative practices can be pre-reflectively, minimally mindful; however, inasmuch as these states do approach the status of a blackout or trance, then reports of these states suggest that there was at least a minimal form of monitoring occurring (quite simply, if there is no experience at all, how could a person report on it?).

blue so the vet is not mistaken about what the drug is. The automaticity the veterinarian exercises when administering medication is constrained by intentions that are at once proximate, but also 'built into' the environment. Reminders of these intentions can work at a heuristic, procedural level to ward off dangerous mistakes.

It should be noted, however, that it is not the case an agent forms an intention and then automatic, intention-free behaviour inevitably follows. Instead, the automaticity is continuously implemented and controlled (J. P. Bermúdez, 2021; Fridland, 2021) even while the intention is being 'automatically' implemented. During the action of carrying out an intention there is a continuous sense of *trying*, of effort exerted with different grades of fluency, that attends the cognitive action (Proust, 2001). Moreover, Gallagher (2017) notes that intentional goals regulate bodily movement, even at the level of milliseconds (p. 147).

*iii). Automaticity and attention*

One of the canonical illustrations of the relationship between automaticity and the failure to inhibit cognitive interference is the Stroop test (originated by Stroop, 1935). Briefly, the Stroop test is predicated on an incongruency between the meaning of a presented word and the print colour of the word (for example, the word RED might be printed in blue ink); the incongruency typically results in slower reaction times and more errors (the Stroop effect), but these effects disappear when the word and colour are congruent. This Stroop effect has traditionally been of interest to psychologists as it is seen to tap into the relationship between an automated task (word recognition) and the more controlled process (naming a colour). On this basis, there is an asymmetry whereby the word recognition interferes with the ability to selectively attend to colour naming, but colour naming doesn't affect selectively attending to word recognition to the same extent (presumably because word recognition is more automated). Nonetheless, it's been demonstrated that the Stroop effect is open to interference (for example, when colour-neutral words are also presented), and in consequence the Stroop effect can be 'diluted' (Kahneman and Chajzyk, 1983). Logan (1985) carefully notes that this does not mean that the Stroop is not at least partially automatic; but it instead means that the Stroop effect is not *entirely* automatic, and, more importantly, that automaticity can co-exist with some forms of attention. In addition, Dishon-Berkovits and Algom (2000) point out that contextual variables modulate the Stroop effect such that the presence, magnitude, and direction of the effect can vary (such that colour-naming rather, than word-naming, takes precedence; a case of a reverse-Stroop effect). An experimental set-up that is biased toward the discriminability of the word-

stimuli can account for the asymmetry that appears to favour word recognition over colour naming.[94] Relatedly, the level of word salience, as shaped by font or print colour, can also affect the responses. In effect, the experimenter can manipulate the subject's attention by way of the characteristics of the stimuli. "[T]he experimental design – fixes the conditional probability of the color, given the word." (p. 1448). Dishon-Berkovits and Algom agree with Besner and Stolz (1999) that the Stroop effect is best viewed as reflecting the role of attention rather than the claim of strong automaticity. Algom and Chajot (2019) review a series of more recent studies that converge on this view.

*iv). Automaticity and controllable*

Finally, and importantly for this thesis, we turn to the relation between automaticity and control. Logan (1985) observes that automaticity and control co-occur in the case of skilled people. This is contrast to a more classic conception where automaticity is theorised to be distinct from control. The standard view can be seen in Figure 4 below.

**Figure 4**

*The standard model of automatic and controlled processing based on Schneider and Schiffrin's (1977) influential model.*

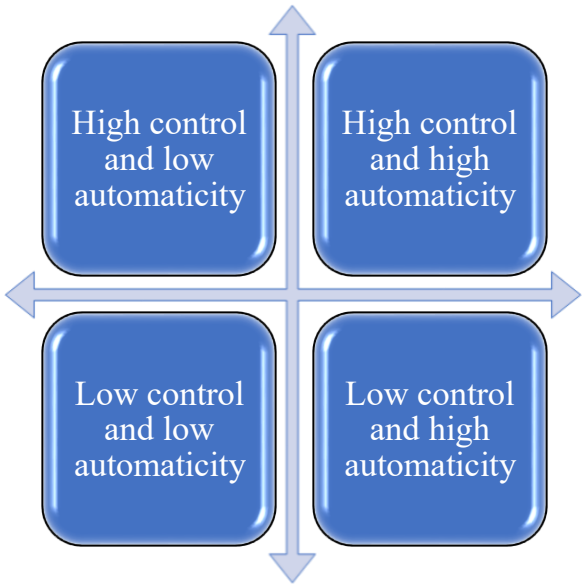

Instead of the standard view, Logan (1985) has argued on his skill-based account that greater automaticity *and* control comes with expertise (also see Toner and Moran, 2020). To this end, Bebko et al. (2005) offer an orthogonal model of control and automaticity. In Bebko et al.'s model (Figure 4), control and automaticity can co-occur, can occur individually, or not occur

[94] More specifically, Dishon-Berkovits and Algom (2000) note that a typical Stroop experiment will have four colours. Each colour word will correspond with its matching colour nine times (constituting a congruence set of 36 trials ). In the incongruent set, each colour will be matched with a different colour three times and once with a neutral word; (so there overall incongruent set is also 36). But the problem is that, even though the number of congruent and incongruent trials match, the matching word-colour pairs appear three times as often as the mismatched pairs. This biases the responses toward easily identifying colour words. Resultantly, Dishon-Berkovits this colour-word correlation in the congruence trials in experiment set-up better account for the Stroop effect than appeals to automaticity of reading or processing speed.

at all. For example, someone who is high in both control and automaticity would, on this model, be deemed highly skilled (upper right quadrant). Someone with low control and high automaticity (lower right quadrant) might, for example, be a typist who writes fast but is mistake-prone. A low control, low automaticity person (lower left quadrant) is often someone learning a new skill. Finally, a low-automaticity-high-control individual (upper left quadrant) might reflect a skill that is less automatic (i.e., more effortful) yet highly accurate (Bebko et al. suggest a 'hunt-and-peck' typist as one example of this).

***Figure 3.***

*Orthogonal model of control and automaticity (based on Bebko et al., 2005).*




*Note: As can be seen in Figure 4, control and automaticity are orthogonal. This has relevance for the extended cognition thesis because it indicates how integration can take place (see section 5.3.).*

One question that arises here is how Bebko et al.'s model relates to extended cognition and metacognition. To be clear, my claim is not that control and automaticity exhaust what it means to be an expert. Indeed, as Menary and Kirchhoff (2014) observe in relation to their account of *extended expertise*: extended expertise, on their view, can involve cognitive transformations by way of a sustained period of training, and extended expertise can involve expertise spread

across a collaborative working group. In effect, any account of (extended) expertise is bound to include a number of characteristics and components. In the present thesis, I am examining control and automaticity as specific components in extended expertise. Indeed, when considering automaticity and control, we can see how claim one of the thesis (concerned with the subpersonal-level status of extended cognition and metacognition phenomena) connects with claim three (that metacognition extends). As will be detailed further as this chapter proceeds, metacognition extends partially by virtue of the extended subpersonal automaticity and control capacities.

*Provisional summary*

In the preceding subsections I have followed Bargh (1994) in putting pressure on the idea that automaticity must by necessary include characteristics such as lack of awareness, lack of intention, inattention, and lack of control as necessary features of automaticity. In effect, automaticity can co-occur with varying constellations of the aforementioned characteristics. But does this mean that we should dispense with the concept of automaticity? Not necessarily. For example, there is the option of following Fridland (2017), Logan (1985), and Moors and De Houwer (2006) in adopting a conception of automaticity defined by characteristics that are not necessary and sufficient for a given phenomenon to count as an instance of automaticity. Just as the concept of *cognition* itself is difficult to characterise by way of an agreed-upon set of terms of that specify necessary and sufficient characteristics for its existence (Allen, 2017), automaticity also has similar definitional problems; however, these concepts can still be usefully preserved. Fridland (2017) advises that the concept of automaticity should be retained in a "qualified and contingent manner." (p. 4347). The concept of *automaticity* generally denotes fast, efficient, parallel and effortless processes that are resistant to interference by cognitive load – and may involve processes that *sometimes*, but not by necessity, involve the necessary features Bargh (1994) criticised, such as lacking in conscious control – but, in the end, the goals of the researcher determine how the concept of automaticity should be operationalised (Fridland, 2017). This is helpful for my purposes because it allows us to preserve the concept of automaticity, although with no necessary and sufficient characteristics.

### *5.2.1. Metacognitive control*

Fluency and epistemic feelings are inflexible and 'automatic'; however, that does not mean these feelings lack control for they can be integrated with control. Thus, inflexibility does not necessarily result from or imply the absence of control (Proust, 2013, p. 299). Proust (2007) observes that metacognitive control has a world-to-mind/cognition fit (the world is modified in light of control), and metacognitive monitoring has a mind-to-world fit (where the mind is being informed of and receiving feedback about an action). As was also outlined in section 4.1.1. cognitive action involves acting on cognitive processes rather than acting on the world; although of course when we act on cognitive processes we indirectly act on the world. Moreover, the structure of a cognitive action involves strategic control and an inherently normative structure (involving fluency, and, analytic norms such as accuracy, consensus, etc.). Cognitive action entails a sense of *trying* (Proust, 2001) and engagement that accompanies analytic metacognition too (Koriat, 1993).

I would now like to briefly examine how automaticity and control can be meshed together. This is of particular importance for section 5.4. that deals with cognitive obstacles and recovery from error.

### 5.2.2. *Knowledge, control, and automaticity as 'meshed.'*

The proposal in Christensen et al. (2016) builds on the ideas we have seen in the preceding chapters; this is namely that cognitive control does not disappear in advanced skill. To this end, Christensen et al. provide a hybrid theory of control and automaticity which they title *meshed control*. This involves top-down control, not just out the outset of an action (for example, deciding on a particular technique of golf-putting), but also on the execution of the action by way of non-explicit, situation awareness (this, in turn, overlaps with the minimally mindful account put forward by Gallagher, 2021). Crucially, Christensen et al. propose that as difficulty increases, the demands on the agent increase, and so control still contributes alongside automaticity (in line with Logan's (1985) account). When a performance is relatively simple, there is a greater degree of tolerance to distraction and the performance can 'appear' more automatic; however, when the task is more difficult there are greater demands, so situational awareness increases and there is less tolerance of distraction. Christensen et al (2016, p. 45) present nine forms of skill that are typically experienced during 'automatic' tasks. The initial five are as follows:

1). *Reduced attention* occurs after the skill has been learned. This of course does not mean there is a complete *absence* of attention; rather, the attention is less explicit.
2). *Multi-task tolerance* occurs when tasks can be combined with no deleterious impact on performance.
3). *Disruptive attention* occurs when attention to the performance can be disruptive.
4). *Sense of cognitive effort* can be low during performance.
5). *Reduced memory* refers to the way in which the performance of a well learned skill can occur with reduced or absent memory.

As an illustration of the aforementioned five forms of skill, Christensen et al. appeal to the familiar example of a skilled driver in a manual vehicle making their way around familiar streets. Such a person would generally exhibit the above five characteristics in a way that a beginner or a driver in an unfamiliar city would not. Nonetheless, these characteristics do not exhaust the various ways in which skill can manifest when confronted with a challenge. This is where *meshed cognition* enters the picture by way of the following factors that are less automated than factors 6 - 9:

6). *Strategic focus* amounts to enhanced attention to situations, methods and goals in the task performance. For example, an experienced driver who does not need to attend to the implementational mechanics of driving (gear-change, etc.) can give greater attention to the whole situation and higher task goals (such as the presence of other cars, lane-changing).
7). An *action slip* occurs when insufficient attention is given to the task (such as when a driver takes the wrong turn).
8). *Increased attention in response to challenge* involves awareness in demanding conditions. For example, during night-driving or on a busy highway increased attention is needed.
9). *Increased control in response to challenge* occurs when there is cognitive effort in response to a challenge. In such situations possibilities are evaluated and decided upon.

The motivation of Christensen et al.'s (2016) *meshed control* approach is intended to stand in contrast to the Dreyfus and Dreyfus (1986) model of skill. According to the Dreyfus and Dreyfus model the execution of a task is entirely automated; but, according to the meshed control model, cognitive control (by way of such factors as 6 - 9 above) is integrated with automated task execution. The key idea is that while some of the implementational components of skill (1 - 5 above) can be largely automated, there is nevertheless room for control by way

of factors 6 - 9. While an easy performance makes lighter demands on situational awareness (and distraction is more tolerable) the same is not so for more difficult problems in which mesh factors intervene (p. 44). Further, the Dreyfus-model views automation as the norm, and control as playing a role when learning a skill or in 'unusual conditions' in which the skill has not automated; Christensen et al. criticise the Dreyfus model for underestimating the extent to which unusual problems can occur (p. 48). In effect, the Dreyfus model sees control as existing outside automation.

Moreover, meshed skill leads to empirical predictions that differ from the Dreyfus model; specifically, whereas the Dreyfus-automated model predicts that distraction is less likely to disrupt a cognitive performance, and may even be beneficial, the meshed theory predicts that distraction will affect performance (p. 44). Another prediction the meshed approach makes is that experts encode more information in memory when the information is deemed relevant for future control whereas less experienced agents will have less ability, notwithstanding having good memory capacity, for predicting what is likely to be relevant in future situations (and thus worth remembering). Novices are more likely to encode more incidental information.

Meshed cognition has factors in common with the skills-based account of automaticity put forth by Logan (1985). On this view, automaticity exists on a continuum rather than a dichotomy between automated and un-automated skill (pp. 371-372), such as we saw with the Stroop effect. On Logan's view, although automaticity is acquired and modified with practice, there is no empirical reason to suggest that automaticity is ever complete; in other words, even though automaticity can *appear* to have reached a ceiling when speed, accuracy, and the reduction of dual-task interference is measured, automatisation can possibly continue to occur beyond observed measurements (p. 373). Moreover, Logan notes that a lot of skill entails the exploitation of dependencies such that the expert task-performer is anticipating more challenges than the less skilled performer (p. 381). This is consistent with Sutton et al.'s (2011) focus on how some forms of skill, control, and semantic knowledge are intertwined:

> Skill is not a matter of bypassing explicit thought, to let habitual actions run entirely on their own, but of building and accessing flexible links between knowing and doing. The forms of thinking and remembering which can, in some circumstances, reach in to animate the subtle kinaesthetic mechanisms of skilled performance must themselves be redescribed as active and dynamic. (p. 95)

More recently Christensen et al. (2021) have provided further evidence for the role of declarative knowledge in skilled action. Similarly, Logan (1985) captures the role and effects of declarative, metacognitive knowledge on automatic procedures:

> Skilled performers usually know more about their capabilities and their strategic options than do unskilled performers, and this *metacognitive* knowledge allows them to make better use of their automatic procedures. Skilled performers usually have a lot of *declarative* knowledge their skill that may not be relevant to performance of the skill. (p. 369).

Logan uses the example of skilled musicians for whom their declarative knowledge of their instruments exceeds that of the unskilled and "this *metacognitive* knowledge allows them to make better use of their automatic procedures." (p. 369). This means that when things go wrong, or if asked, the would be able to detail more than others about a particular topic, and perhaps use the knowledge to get out of a difficult situation. This is consistent with the claims of 'dual-process' style accounts set forth by Koriat and Levy-Sadot (1999) and Proust (2013) whereby metacognitive control constrains fluency at the procedural metacognitive level.

### 5.2.3. *Automaticity, cognitive penetration, and perceptual expertise*

Experts tend to be quick to identify a complex state of affairs. For example, the capacity of a radiographer can recognise, with some degree of automaticity, an aberrant feature on an scan, a skilled bird-watcher can immediately identify a species, or a poultry sorter who can 'automatically' determine the sex of a chicken (Stokes, 2021b; also, see Elzinga, 2018). While I won't rehearse what I noted in section 2.3.1, it is worth pointing out that, regardless, of the existence of cognitive penetration, perceptual expertise has been well argued for by Stokes (2021a, 2021b). Experts typically view scenes holistically, and their eye-saccades are suitably efficient so as to quickly take in the relevant aspects in a scene. The expert will perform rapid categorisations, even if they can't fully articulate these 'automatic' behaviours, but will often report that there is a phenomenology of a relevant feature 'popping out.' (2021a, p. 247). Additionally, there are distinct differences between novices and experts in how interference affects automaticity. While interference tends to affect context-sensitive attention in a novice; expert attention has been found to be more inflexible, automatic, to focus on more on spatial relations between features, and selective attention has been found to be given to only part of a

presented image (hence, if part of the image is affected, it is more likely to lead to interference effects).

An example of cognitive penetration can arguably be found in veterinary practice. More precisely, the skilled veterinarian's perceptual experience of stitching could be an instance of cognitive penetration. There are hundreds of stitch patterns that a veterinarian can use ranging from more common stitches including what are termed that of the *simple interrupted*, *simple continuous*, *horizontal mattress*, *vertical mattress*, *cruciate*; and also more complicated stitch patterns such as the *cushion suture* used to prevent leakage in hollow organs. The veterinarian, after sustained practice, will study and appropriate these techniques and this, in turn, has top-down effects when examining the instruments, thread, and the wounds that need healing. Whereas a novice might just see a thread and a wound, the veterinarian *sees more.* The novice might have a recollection of attempting, clumsily, to sew a button on a shirt, and, with this, a range of quite idiosyncratic associations that are diffuse and whimsical; these associations will be different to that of a vet. In consequence, unlike the novice, the skilled experience of a vet can result in top-down effects on perception.

*Section summary*

I have examined the various ways in which automaticity co-occurs in relation to awareness, intention, attention, intention, and control; I also looked at cognitive penetrability and the way knowledge affects control. This serves to demonstrate how automatic processes can be highly 'meshed', to borrow Christensen et al.'s (2016) term, and will be useful when considering the next section in which we examine a veterinarian using a medication dosage app.

## 5.3. The Veterinarian and the Medication Dosage Index App: An Example of Metacognitive Extension.

One of the tasks a veterinarian faces is determining which medication to give an animal and how much. The dose is how much should be given to the animal to produce the desired effect (e.g., kill bacteria and relieve pain). The species of animal, the weight of the animal, the precise presentation of symptoms and the age are relevant factors. Before the development of software accessible on a phone app[95], veterinarians exclusively needed to consult the medication

[95] There are various apps available. This example is not intended to describe a particular app; rather, it is designed to capture a composite idea of an app.

package dosage book (or a dosage chart on the wall for emergency drugs) for information on the medication, then the dose for the particular animal, and finally calculate the correct dosage for the animal. It can be time-consuming if multiple drug dosages need to be calculated, and the calculation processes are more prone to error. For example, let's imagine that the veterinarian needs to know the correct antibiotic dosage for a dog. A typical calculation proceeds by way of the following steps:

1. The first step is to weigh the animal (for example one might weigh a dog and determine that it is 10 kg).
2. Search for the adequate medication (for example, an antibiotic that is right for the dog).
3. Search for dosage for the species in a reference book or chart. In this case it is 10 mg/kg.
4. Find the concentration of the medication. (The medication could be in pill or liquid form). In this case, it is 100 mg/ml in liquid form.
5. Calculate 10 kg x dosage 10 mg/kg / concentration of drug 100 mg/ml = 1 ml.

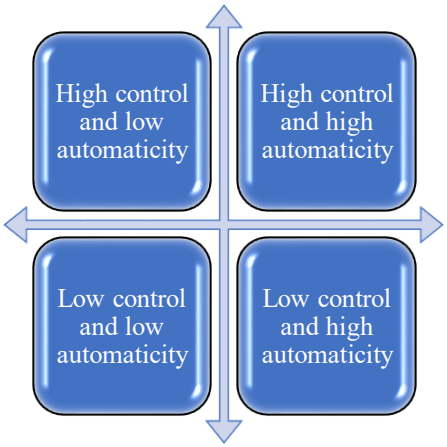
With an app, the veterinarian only needs to type in the species, drug type and weight, and the app will calculate it. The benefits of this include:

1. The vet does not need to search for the dosage and concentration of medication, nor calculate the dosage. This saves time – especially if more than one drug is required.
2. Minimises the potential error of not calculating the right dosage.
3. The dosage and concentration measurements can be more up to date.
4. The provision of updated information on side-effects is available.
5. The availability of updated information on drug interactions and contraindications (which saves time and reduces error). For example, griseofulvin should not be given to an animal with liver disease or to pregnant dogs.

Opportunity for lengthy contemplation is limited as the situation is time-pressured, and the information can be incomplete and uncertain. For instance, if a dog has an unexpected anaphylactic reaction, the vet has to make a rapid decision regarding medication to treat the symptoms. Moreover, using the app presupposes the vet has prior knowledge and experience with dosages and medications. Contrast this with someone who can use the dosage app work and will not be able to exercise the same level or character of cognitive control and monitoring.

He has no idea how to *read* the tools, even though he can, in a superficial sense, *use* them. It is not enough to simply go through the motions of using an app or tool more generally.

It is at this point useful to recall Figure 4, based on Bebko et al. (2005).




The use of the app, and the extent to which metacognitive integration occurs will depend on where the vet is on the quadrant. Let's look at how vets of varying levels of skill align with the four quadrants.

***Table 3***

*A framework for isolating metacognitive extension*

| Quadrant | Metacognitive Control | Automaticity | Extension | Metacognitive extension |
|---|---|---|---|---|
| 1. **(upper-right)** | High | High | High | High |
| 2. **(upper-left)** | High | Low | Marginal | Marginal |
| 3. **(lower-left)** | Low | Low | Marginal | Low |
| 4. **(lower-right)** | Low | High | Marginal | Low |

*Note: This table highlights the implications of control and automaticity for i). cognitive extension, and ii). cognitive extension. Each quadrant refers to a different veterinarian.*

Q.1). The highly controlled, highly automatic quadrant constitutes high level metacognitive extension. This vet at once exercises high control and high automaticity. Their behaviour involves what fluent, controlled coupling with the dosage app. This skilled coupling is consistent with Logan (1985) account of skill with regards to the way in which control and automaticity co-occur.

Q. 2). The highly controlled, low automatic quadrant constitutes a more marginal case of metacognitive extension as it is high in metacognitive control, but low in automaticity.

Q. 3). The low control and low automaticity quadrant doesn't constitute metacognitive extension.

Q. 4). The low control and high-automaticity quadrant constitutes a form of marginal extension but not metacognitive extension.

I will now turn to a consideration of metacognitive control and automaticity in regard to a veterinarian recovering from an obstacle. The aim is to demonstrate how metacognitive skill can assist with recovery.

### 5.4.Obstacles, Error and Recovery

Gallagher (2017) writes about Dewey's concept of *the situation*. Dewey's characterisation of cognition entails the organism-environment such that the two cannot be distinguished. Nonetheless, "Dewey's concept of the situation arises when the coupling of the organism-environment becomes problematic or starts to break down … *when it starts to go wrong we have what Dewey calls a problematic situation and it calls for a re-pairing, a reestablishment of a workable coupling*. (p. 55, italics added). Dewey's idea of the *situation*, with I break-down in coupling and need for repair, is relevant here because metacognition usually occurs when there is a cognitive obstacle or disruption (Proust, 2015). Furthermore, insofar as metacognition is skilful, Fridland (2015) reminds us that part of learning a skill is adapting to the variable conditions that present as novel yet relevant in the environments in which these skills unfold. More precisely, repetition, even when successful, does not qualify as skill if it leads to behaviour that is too rigid and inflexible; instead, the test of skill is in situations of disruption where there is a need for repair (Fridland, 2019; Sutton et al., 2011). As was illustrated section

5.2, skills are not entirely automated, and it is exactly this flexibility, this lack of total automation, that can present as metacognitive skill in situations where there is an obstacle. What is of importance is *if*, and *how*, an agent adapts to and recovers from the obstacle.

The problem, the *situation* as it were, for the veterinarians I will consider is that the drug dosage index we were introduced to is not working as it should be. The index, for expository purposes, has been affected by a bug, and so its output is unreliable. This being so, I will hold two assumptions constant:

1. For each of the veterinarians I will be considering in detail (Q 1 - 4) there is some form of awareness of the potential error as manifested by a disrupted sense of fluency.
2. There is no error signal – i.e., no explicit warning or failure signal – that flashes on the screen. The veterinarians need to detect that there is an error in the absence of an express suggestion that something is wrong.

I intend to conceptually isolate the specific ways in which the veterinarians attempt to recover from the obstacle.

**Table 4**

*A framework for isolating metacognitive extension*

| Veterinarian | Error Detection | Metacognitive Control | Automaticity | Cognitive Extension | Metacognitive Extension |
|---|---|---|---|---|---|
| **Q. 0.** | No | N/A | N/A | Possibly | No |
| **Q. 3** | Yes | Low | Low | Low (marginal) | Low |
| **Q. 4** | Yes | Low | High | Low (marginal) | Low |
| **Q. 2** | Yes | High | Low | Intermediate (marginal) | Intermediate (marginal) |

| | | | | | |
|---|---|---|---|---|---|
| **Q. 1** | Yes | High | High | High | High |

*Note.* The roles of control, automaticity, metacognition and extended cognition in relation to veterinarians of mixed experience. For expository purposes, the veterinarians are ordered from not exhibiting metacognitive extension (Q3) through to exhibiting metacognitive extension (Q1).

I will now review each veterinarian from least to most metacognitively extended. In each vignette I will focus on how each veterinarian uses their skill to assess and deal with the error, and the implications of this for metacognitive extension.

*Q. 0. Veterinarian*

As can be seen above, Q.0. vet, whom I've placed here for expository purposes, does not detect any errors. This being the case, the existence of metacognitive extension is beside the point. As for the existence of some form of non-metacognitive extension, it is possible that this vet could, on some minimal definitions of cognitive extension, be classified as being coupled to her tool such that it could amount to a form of bona fide extension; but it is not extension of the sort that is of interests as it is not metacognitive.

*Q. 3. Low-control, low automaticity veterinarian.*

The vet detects an error and experiences disfluency. She re-types the species, drug-type and weight into the app, and the number presents as it did initially. Even though the app is wrong, the vet doesn't recognise the error, and assumes, despite initial reservations, that because she re-typed the correct inputs into the app, the app result must be correct. The metacognitive extension is low.

*Q. 4. Low control, high automaticity.*

The vet detects an error and experiences disfluency. She checks to see that the inputs (species, drug-type, and weight) and these are the same as before. Satisfied, she continues with the dosage recommendation, even though it may present as somewhat unusual. In contrast to the previous vet (Q.3), this vet operates in a manner where there is greater automaticity, but the

metacognitive control is not thoroughgoing, and the vet's metacognitive activity is brought to a premature end. Perhaps this greater automaticity allows for her behaviour to be classified as a case of extension. There is coupling, and the process is more rapid, but the metacognitive extension is still low.

*Q. 2. High control, low automaticity vet*

The vet detects an error and experiences disfluency. She checks the inputs, and these are correct, yet the app continues to produce a result with which she is dissatisfied. Nonetheless, unlike vets Q.3 and Q.4, vet Q.2 does not accept the dosage that the app presents her with. Instead, she overrides the app result. This can take various forms. The vet may manually calculate the inputs on a calculator (or pen and paper); alternatively, or in addition to this, another vet (a colleague or supervisor), may assist. This being so, the process of recovering from the error may involve collaboration. Moreover, Hutchins (1995b) observes:

> Not every recovery from error is instructional in intent or in consequence. Some are simply what is required to get the job done. *There may be no need to diagnose the cause of the error in order to know how to recover from it*. (p. 276, italics added).

The vet does not necessarily need to make a diagnosis of the dosage app problem. To successfully correct the problem does not necessarily entail knowing *why* the app is not working. Moreover, while the vet is exercising metacognition, it is arguably happening in a way that is not automatic. The implication of this can lead to trade-offs. For example, Hutchins (1995b) observes that the costs associated with correcting errors need to be traded off with the benefits that come with "detecting, diagnosing, and correcting errors." (p. 279). Hutchins's points out that intercepting errors can be costly because during the time allocated to detecting and correcting for an error, more errors can be made; moreover, when the agent is focussed on detecting and recovering from an error, they can be distracted and thus make them more error-prone. This is a point relevant for the Q. 2 vet I'm currently considering whose lower automaticity could result in being more likely to make more errors even though this vet has a high level of control.

*Q. 1. High-control, high automaticity vet*

The final vet I am considering (Q. 1) partially recapitulates the response of the Q. 2 vet (above). Like, Q. 2, the Q. 1 vet is able to detect a problem and correct for it by going beyond the information the tool is supplying her with. What is distinctive of Q. 1, however, is that her case entails high control *and* high automation, and it is here that we see an interaction between the procedural and the analytic forms of metacognition. The vet notices that the dosage number doesn't look right. She may not be able to articulate what it is that is wrong about the initial, erroneous result, but her epistemic feeling, with its attendant experience of disfluency, plays a guiding role in this instance (consistent with Koriat & Sadot, 1999). She thus chooses to override the app and rely on her own judgement she might re-type the weight to check the input, or possibly skip this step entirely. The automaticity of her responses, as manifest by their rapidity, and the way in which the cognitive sub-capacities are implemented, means that the vet can exercise *strategic focus*, and increased *attention*, and *control* when encountering an obstacle. Furthermore, she can avoid the trade-offs that Hutchins (1995b) refers to whereby detecting errors and managing errors can unfortunately entail granting less time to other tasks. She can anticipate problems, so her situational awareness is heightened; this is important as other errors, beyond problems with dosage information, can proliferate, including, information on side-effects that could be wrong.

The vet is able to integrate conceptual knowledge with a series of analytic metacognitive norms; and thus correct for more 'automatic' responses.[96] This is consistent with claims of Christensen et al. (2021) whereby skill involves declarative knowledge and does not totally automate.

## 5.5.Discussion, potential criticisms, and clarifications

### *5.5.1. Success conditions and meta-cognition:*

As can be seen, I have arranged the cases so as to form a hierarchy where the vets become progressively more successful. As such, there is a connection between the presence of metacognition and success. Nevertheless, there are times when metacognitive skill and successful goal-completion can pull apart. With less experienced veterinarians, it is possible that by way of luck they can arrive at the correct result. Conversely, there are times when metacognitive expertise leads to inadequate responses. In effect, metacognitive extension does

[96] We can also expect the presence of top-down, cognitive penetration effects.

not hinge on metacognition *always* leading to the objectively right decision. Metacognition, including metacognition that entails an extended component, can still be wrong, but is not less extended for that. Nonetheless, the general expectation is that the more skilled the metacognition, the more likely the agent is to be successful.

Elzinga (2018) notes that someone who can be classified as knowing how to perform a task:

1. Can reliably live up to normative standards (i.e., conform with the standard of what constitutes a successful performance). This reliability does not amount to "perfect conformity" of the standard at all times (p. 121). Elzinga cites the example of an experienced chess player who may be inattentive at times and make poor moves, but can still be said to exhibit high levels of know-how despite the occasional errors. Moreover, inasmuch as the skilled performance is norm-governed, the norms are not necessarily verbalisable.
2. Skilled performances involve resilience when encountering resistance; but this does not mean the skilled performer will always need to perfectly adapt. Even experts sometimes fail to adapt. "The idea of trying to perform and trying to adapt is enough" to still be successfully qualified as 'knowing-how' rather than knowing-that. (p. 124).

### 5.5.2. *Continuities and cut-off points*

The above account of veterinarians and their tools is, by necessity, a stylised account in which I have distilled what I believe to be the most relevant elements. Situations, obstacles, and (attempted) recoveries unfold in ways that are not always neatly categorisable. This being so, rather than a bright-line separating extension from non-extension, and, specifically, metacognitive-extension from non-metacognitive extension, we should instead conceive of the framework I have offered in Table 4 as working along a sliding-scale in which boundaries can be porous. This conception is consistent with literature in which cognitive extension can be viewed as dimensional (Sutton et al., 2010), in which procedural metacognition operates by way of degrees (Zawidzki, 2021), and automated skill exists on a continuum with no end-point (Logan, 1985).

One's level of automaticity and control is not fixed. Training and experience can mean one moves to different levels; alternatively, or in addition, collaborating with people who are more experienced can expand one's metacognitive capacities. It can also be the case that working with people who make mistakes can also be useful (Hutchins, 1995b, p. 277).

### *5.5.3. What is being extended?*

What is being extended? In the example of the metacognitive usage of the dosage app, there are at least two aspects of extension present. First, the process of calculating; and, secondly, the information itself, including belief states.

1. *Coupling and controlled action*

A critic could point out that what is taking place is *embedded* rather than *extended* cognition (for example, Goldinger et al., 2016). For example, a critic could say that the app is triggering (meta)cognitive activity in the head of the veterinarian (the embedded view) rather than the activity with the app constituting (meta)cognitive activity (the extended view). (See section 1.6. for general arguments in favour of an extended rather than an embedded view). If one invokes the Mutual Manipulability (MM) criterion (Menary, 2006, 2007). Recall that "X is the manipulation of the notebook [or device] reciprocally coupled to Y – the brain processes – which together constitute Z, the process of remembering." (2006, p. 334). It would appear that the use of the app fulfils the MM criterion as the app is highly tangible, requiring dynamic sensorimotor activity and continuous, reciprocal causation. Thus, the use of the dosage app is arguably more *active.* In contrast to this, looking at a wall-chart - with information on drug concentration, species dosage, the kg dosage-table with the advised amount of drug for an individual animal – is arguably a more *passive* process as there is less sensorimotor engagement with the resource. So, there are *varieties of coupling* of varying levels of activeness and passivity. The type of coupling that takes place between the vet and the app is arguably different from the coupling that takes place between a vet and a wall-chart; however, this is not to say that the use of the wall-chart does not require sensorimotor engagement (e.g., by way of quick eye saccades when examining the chart). Nonetheless, the sensorimotor skill-set will be different when the coupling is with an app rather than a chart. Furthermore, these varieties of coupling involve varying levels of fluency as the app or wall-chart will afford varying levels of ease and resistance in terms of speed and access to information. Now, one could still make a case for including using a wall-chart as part of an extended cognitive process, although that is not the case I making here. Instead, I am proposing that the use of the app entails a stronger form of coupling and thus extended cognition because the app, unlike the wall chart, involves more tactile behaviour than the use of the wall-chart.

2. *Conceptual beliefs*

Firstly, the information the app provides is propositional, determinate content (a weight, a particular medication). Is this content part of extended cognition? Extended belief-states are more aligned with folk-psychological intuitions (Pöyhönen, 2014); these are exemplified by the Otto-notebook style thought experiment in Clark and Chalmers (1998). There are ongoing debates as to whether these beliefs are sufficiently fine-grained to count as beliefs in the way that we folk-psychologically characterise beliefs. However, Shea (2020) suggests that there can be genuinely psychology states that do not conform to what we have traditionally conceived of as belief-states. Shea cites interactions with written artefacts as an example of forms of cognition that may elude standardly characterised belief states; in consequence, there are forms of extended cognition that involve propositional states for which we lack folk-psychological terms. What Shea is referring appears to resemble Hutchins' (2014) focus on the relative densities of information as it flows across the agent-environment system (p. 37).

Secondly, what should be said about information that the vet has not encountered before? For example, a vet could know of a drug side-effect with a particular dog breed; but, until the point of using the dosage app, she may not know that this particular side-effect applies to another breed (it could be that the side-effect has only recently been discovered for this breed or that it had long been discovered but the veterinarian simply had not come across it in relation to this breed until now). Would we want to say that this relatively novel information counts as part of the vet's extended cognitive process? Perhaps one way of approaching this question is to consider the difference between *dependence* and *understanding*. Shea (2018) distinguishes between dependence on metacognitive concepts and understanding of metacognitive concepts. Shea notes that usually dependence[97] and understanding are aligned; however, they can also come apart. Presumably the more information a vet has – and the greater her understanding – then, even if it is not a proposition that has been actively considered before, it nonetheless emerges out of other propositions (knowledge of dog breeds, knowledge of drugs) that the vet *does* understand and is disposed to understand. Moreover, this understanding implies a

[97] Although there are some social epistemologists who argue dependence can in some circumstances constitute knowledge. Hardwick (1985) argues for *epistemic dependence* where a person can know something by proxy by way of deferring to the authority of someone who does know (for example, a layman can 'know' via an expert). Because of divisions of cognitive labour, even within an area of expertise not even the experts themselves will possess direct knowledge of everything they know and thus need to rely on knowledge-by-proxy (Hardwick, 1985).

metacognitive appreciation of uncertainty; for example, part of this understanding means blood-test before the drug is administered

It's also worth remembering that many even supposedly 'brain-bound' propositions are not explicitly computed in any sense. Dennett (1978), for example, notes "the propositional attitudes we *have* far outstrip those we (in some sense) actively entertain" (p. 104). The proposition that wild zebras don't wear overcoats, for example, is not something that most people have thought about. Nowhere in the brain does this thought 'exist' as such; rather, such propositions fall out of other propositions – of wild animals, of garments – that we possess. This has implications for the extended cognition hypothesis as it suggests that the threat of cognitive bloat can be diffused by considering that many thoughts do not need to be endorsed in the past, nor actively considered, to count as elements as part of a cognitive system. For a vet, with sufficient understanding, any 'new facts' that emerge from the output of an app will be so integrated with previous understandings that the information can be said to be densely interconnected.

3. *Metacognitive effort*

The focus on metacognitive extension brings us full-circle with the concerns in chapter one. Recall that I was concerned the account of extended metacognition advanced by Kirsh (2004) was too focused on first-order reacting that the metacognitive processes became too effortless. Further, I related this to a problematic set up by Clark (2015) where he was concerned that extended cognition can be either too reliable (and so not extended) or too effortful (and thus too deliberative and not extended).

Metacognitive action has an element of continuous trying (Proust, 2001, 2013). Indeed, metacognition is not just world-directed, but directed at one's cognitive sub-systems. However, in the case of extended cognition, there is a sense in which the processes are both world and cognition-directed; however, this is because parts of the local environment with which the agent is coupled are part of cognition, albeit only when coupling holds, and only derivatively. What matters is the effort, the *trying*, the hybridity of planning; it is the sense of effort, or *trying*, that is metacognitive. The agent is seeking to repair the 'workable coupling' so as to resecure smooth contact with the world.

## 5.6. Conclusion

This chapter began with a consideration of the importance of individual skill in relation to metacognition. Individuality has been a throughline of this chapter; more specifically, the effects of enculturation and sustained practice have an effect on expertise (Menary & Kirchhoff, 2014; Montero, 2016). Even when considering the value of automaticity and skill, and how explicit knowledge is meshed with procedural metacognition, it has been emphasised that individual skill matters for cognitive extension.

In the second half of this chapter I created a framework for examining instances of metacognitive extension. To this end, and consistent with claim *three* in this thesis, I examined four vets as they detected errors, yet responded in different ways. It became clear that the sort of close coupling seen in cognitive extension could be present, but was insufficient to secure metacognitive extension. Instead, exhibiting high automaticity and metacognitive control are needed.

## 6. Conclusion, Relevance and future directions

To conclude I will re-examine the three claims I introduced at the beginning of the thesis; finally, in 6.3 I will briefly state a few applications of and possible future directions for this line of research.

*The three claims*

In chapter one I made three claims that I proposed would form three overarching themes in this thesis.

Claim *one*: Procedural metacognition and extended cognition are both primarily phenomena that have subpersonal-level mechanisms. This was explored in chapter three where it was suggested that possibly the personal-level should be abandoned. It was decided that even if one does pursue this option, the thesis fares no worse for that. This is well demonstrated by the interaction between the procedural and the analytic (claim two, below); and the role of normativity, cultural practices, and metacognition (in chapter four). Chapter four provided evidence that perhaps what we call cognition, and specifically our capacity for metacognition, is the result of mutually reinforcing interactions between subpersonal-level (brain-bound and extended processes) and cultural practices. To the extent that the personal-level, with its attendant folk-psychological concepts, is not relevant for procedural metacognition, that is because these concepts are absorbed into a larger, and arguably superior account that seeks explanation in cultural practices and subpersonal mechanisms rather than mentalistic, personal-level explanations.

Claim *two*: This claim is concerned with the binding of procedural and analytic metacognition during cognitive action. This claim was in part motivated by Arango-Muñoz's (2015) observation that one of the weaker points in Proust's (2013) is on how procedural metacognition and analytic metacognition relate to each other. On Proust's (2013) account, procedural metacognition can be conceptually enriched, and this conceptual enrichment can serve as a corrective when encountering a cognitive obstacle. Yet, at the same time, the fluency provides something of a bedrock, a *knowing-how*, on which control can be guided, and epistemic acceptances formed. I have taken this further by examining (in chapter two) analog cognition, and by arguing – in opposition to Proust (2015) – that cognitive penetration can

occur as way of binding together epistemic feelings at the procedural level and analytic norms. I also examined the role of language in binding the procedural and analytic forms of metacognition. Furthermore, this idea was developed further in chapter five in relation to skill, and, specifically, metacognitive skill. This involved looking at how, on a meshed account of control, automatic processes are intricately connected with processes of control.

Claim *three*: This was the claim that metacognitive extension can occur. Here, I differentiated extension from merely being actively coupled with a device. As I examined in my framework (Table 4) a subject can conceivably demonstrate a high amount of automaticity with a device, but this alone does not constitute metacognitive extension. Furthermore, while I don't propose that the presence of a high level of control and a high levels of automaticity jointly exhaust the concept of extended expertise, I nonetheless do propose that control and automaticity can be conceptualised as subpersonal-level phenomena. In fact, it is on this final point that claim three of this thesis builds on the claim one (concerning subpersonal-level phenomena).

In conclusion, while the above three claims are linked, they are not logically dependent on each other; however, they nonetheless mutual reinforcing. The subpersonal-level explanations (claim one) assist with explaining how procedural metacognition and analytic metacognition bind together (claim two); and these two claims assist in providing a conceptual space in which the existence of metacognitive extension can be defended (claim three) as involving subpersonal norms (see section 4.5) and subpersonal phenomena such as control and automaticity.

***Relevance and future directions:***

First of all, this research has relevance insofar as it provides conceptual clarification and a framework within which extended metacognition can be understood and in which empirical hypotheses can be developed. Indeed, as van Rooj and Baggio (2021) have demonstrated, a lot of cognitive psychology has unfortunately been focused on effects rather than theorising about psychological capacities. The current work, rather than attempting to present a series of effects, has attempted to organise these effects into a framework that has explanatory value. Moreover, this framework focusses on a specific form of cognitive capacity rather than cognition tout court. As Allen (2017) has observed, *cognition* is an umbrella concept that encompasses a range of more specific (sub)capacities. That being so, it makes sense to characterise cognitive

extension in terms of more specific psychological capacities such as transactive remembering as advanced by Wegner (1987); or 'cognitive reserve' in a rehabilitative context (Drayson & Clark, 2020). The research I have presented in this thesis goes in this direction by examining the capacity of metacognition in both its procedural and analytic forms. In effect, I have presented a more specific, fine-grained way of looking at extension that is consonant with the unique and variable cognitive profiles of individuals. Furthermore, the focus on specific capacities is extended, offers a fruitful way of apply extended ideas that are more refined than simply talking of cognitive extension tout court.

Secondly, one of the advantages this work has in relation to cognitive extension – and distributed, 4-E, and externalist approaches in general, and of metacognition in particular – is that it makes the cognitive processes more explanatorily tractable when the unit of analysis is the agent-environment system rather than only focusing on what is occurring inside the agent's head. This explanatory tractability has implications for cognitive scientists studying metacognition. More generally, educationalists, health-professionals, and indeed, anyone exercising any sort of expertise, or encountering any sort of cognitive obstacle, can benefit from understanding the phenomenon of metacognition as existing as extended rather than merely located inside the head.

More broadly, when people with technical expertise are using tools, there are questions of the extent to which they can be responsible for 'decisions' made by tools. I don't propose that there are any easy answers to such concerns, but without a doubt these concerns will be of growing relevance. I would suggest, however, that clues to understanding these concerns can be found when considering metacognitive capacities, specifically, how metacognition enables and constrains tool-use. Further, there are questions to be asked about the extent to which we are personally responsible, and on an extended view, perhaps *collectively responsible*, for our metacognitive capacities. Future research could pursue this idea.

There has also been recent work looking at the connection between psychiatry and enactivism (De Haan, 2020; Nielsen & Ward, 2020), and, more generally non-reductive accounts of psychopathology (Borsboom et al., 2019). It would be useful to explore this further. A recent meta-analysis found that all psychopathologies involve a metacognitive component (Sun et al., 2017). Relatedly, the connection between extended metacognition and psychotherapeutic models and techniques[98] and cognitive extension has the potential to be a

[98] There is a lot of overlap among different forms of psychotherapy and one of the common factors, I would propose, is that all of them feature metacognition in some way. Specifically, most forms of psychotherapy,

growing area of research. Of particular interest is the growing use of technology in psychological-therapeutic practice; whether it is a case of patients using mental-health apps, or professionals using software to assist with diagnosis, the emerging consensus is that the practice of professional psychology is increasingly intertwined with the use of technology. This area would also be fertile ground for further exploring the ideas in this thesis; namely, the myriad ways in which metacognitive loops extend across brain, body, and world.

whether more cognitive, psycho-dynamically, or behaviourally-oriented entails some form of *corrective* experiences involving metacognitive monitoring and control.

## References


Adams, F. & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology, 14*(1), 43 - 64.

Adams, F. & Aizawa, K. (2008). *The Bounds of Cognition*. Oxford: Blackwell.

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin, 141*(5), 931–965. https://doi.org/10.1037/bul0000021

Algom, D., & Chajut, E. (2019). Reclaiming the Stroop effect back from control to input-driven perception attention and perception. *Frontiers in Psychology, 10*, article 1683, https://doi.org/10.3389/fpsyg.2019.01683

Alksnis, N., & Reynolds, J. A. (2021). Revaluing the behaviorist ghost in enactivism and embodied cognition. *Synthese*, *198*(6), 5785-5807. https://doi.org/10.1007/s11229-019-02432-1

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*(3), 219-235. https://doi.org/10.1177/1088868309341564

Allen, C. (2017). On (not) defining cognition. *Synthese, 194*, 4233-4249. https://doi.org/10.1080/09515089.2013.766789

Allen, M. & Friston, K. J. (2018). From cognitivism to autopoesis: towards a computational framework for the embodied mind. *Synthese, 95*, 2459-2482. https://doi.org/10.1007/s11229-016-1288-5

Andrada, G. (2021). Mind the notebook. *Synthese, 198*, 4689-4708. https://doi.org/10.1080/00207727008920220

Apperly, I. A. (2011). *Mindreaders*. New York, NY: Psychology Press.

Arango-Muñoz, S. (2013). Scaffolded memory and metacognitive feelings. *Review of Philosophy and Psychology, 4,* 135-152. https://doi.org/10.1007/s13164-012-0124-1

Arango-Muñoz, S. (2015). Review of the book *The philosophy of metacognition: Mental agency and self-awareness,* by Joëlle Proust. *Mind and Machines, 25*(3), 297-300. https://doi.org10.1007/s11023-015-9376-8

Baggio, G. (2021). Imagery in action. G. H. Mead's contribution to sensorimotor enactivism. *Phenomenology and the Cognitive Sciences, 20*, 935-955. https://doi.org/10.1007/s11097-021-09784-5

Baggs, E. and Chemero, A. (2021). Radical embodiment in two directions. *Synthese*, *198*, 2175-2190. https://doi.org/10.1080/00207727008920220

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*(4), 577-660. https://doi.org/10.1017/S0140525X99002149

Beck, J. (2018). Analog mental representation. *WIREs cognitive science, 9*(6), 1-10. https://doi.org/10.1002/wcs.1479

Barwich, A. S. (2019). The value of failure in science: The story of grandmother cells in neuroscience. *Frontiers in Neuroscience, 13*, e:1121. https://doi.org/10.3389/fnins.2019.01121

Beilock, S. L., Carr, T. H., MacMahon, C., & Starkes, J. L. (2002). When paying attention becomes counterproductive: Impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills. *Journal of Experimental Psychology: Applied*, *8*, 6-16.

Beilock, S. L., & Gray, R. (2004). Why do athletes choke under pressure? In G Tenenbaum & R. C. Eklund (eds.), *Handbook of sport psychology* (3rd ed). (pp. 425-444). Hoboken, NJ: Wiley.

Bergson, (1896 / 1991). *Matter and Memory*. (M. M. Paul & W. S. Palmer, Trans). Zone Books.

Bermúdez, J. P. (2021). The skill of self-control. *Synthese, 199*, 6251-6273. https://doi.org/10.1007/s11229-021-03068-w

Bermúdez, J. L. (1998). *The paradox of self-consciousness*. Cambridge, MA: MIT Press.

Bermúdez, J. L. (2000). Personal and subpersonal: A difference without a distinction. *Philosophical Explorations, 3*(1), 63-82.

Bermúdez, J. L. (2003). *Thinking without words.* Oxford University Press.

Bermúdez, J. L. (2005) *Philosophy of Psychology: A contemporary introduction.* Routledge.

Bermúdez, J. L. Cahen, A. (2020). Nonconceptual mental content, *The Encyclopaedia of Philosophy*, (E. N. Zalta (Ed.), Stanford University. https://plato.stanford.edu/entries/content-nonconceptual/

Bernard, S. Proust, J. & Clément, F. (2015). Procedural metacognition and false belief understanding in 3- to 5- year old children. *PlosOne, 10*(10), https://doi.org/10.1371/journal.pone.0141321

Besner, D. & Stolz, J. A. (1999). What kind of attention modulates the Stroop effect? *Psychonomic Bulletin & Review, 6*, 99-104.

Birch, J. (2021). Toolmaking and the evolution of normative cognition. *Biology & Philosophy*. *36*, article 4. https://doi.org/10.1080/09515089.2013.766789

Blakemore, S. J., Frith, C. D., & Wolpert, D. M. (2001). The cerebellum is involved in predicting the sensory consequences of action. *Neuroreport, 12*, 1879-1884. https://doi.or/10.1097/00001756-200107030-00023

Boone, W., & Piccinini, G. (2016). The cognitive neuroscience revolution. *Synthese, 193* 1509-1534. https://doi.org/10.1080/09515089.2013.766789

Borsboom, D., Cramer, A. O. J., & Kalis, A. (2019). Brain disorders? Not really... Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences,42*(2*),* 1-63. https://doi:10.1017/S0140525X17002266

Brandom, R. B. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment.* Cambridge University Press.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*, 139-159

Burge, T. (1979). Individualism and the mental. Reprinted in T. Burge (ed.). *Foundations of mind* (pp. 100-150). Oxford: Clarendon Press.

Burge, T. (1986). Cartesian error and the objectivity of perception. Reprinted in T. Burge (ed.) *Foundations of mind* (pp. 192-207). Oxford: Clarendon Press.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences, 32*(2), 121-138. https://doi.org/10.1017/S0140525X09000545

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

Carruthers, P. (2018). The contents and causes of inner speech. In P. Langland-Hassan & A. Vicente (Eds.), *Inner speech: New voices* (pp. 31-52). Oxford: Oxford University Press.

Carter, J. A., Clark, A., Kallestrup, J., Palermos, S. O., & Pritchard, D. (Eds.). (2018a). *Extended epistemology*. Oxford: Oxford University Press.

Carter, J. A., Clark, A. & Palermos, S. O. (2018b). New humans, ethics, trust. In Carter, J. A., Clark, A. Kallestrup, J., Palermos, S. O. & Pritchard, D. (Eds.), *Extended epistemology.* (pp. 331-352) Oxford: Oxford University Press.

Carter, J. A., & Rupert, R. (2021). Epistemic value in the subpersonal vale. *Synthese, 198*(10), 9243-9272. https://doi.org/10.1080/00207727008920220

Chalmers, D. (2019). Extended cognition and extended consciousness. In M. Colombo, E. Irvine, and M. Stapleton, (Eds.), *Andy Clark and his critics* (pp. 9–20). New York: Oxford University Press.

Chemero, A. (2009). *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press.

Christensen, W., Sutton, J. McIlwain, D. J. F. (2016). Cognition is skilled action: Meshed control and the varieties of skill experience. *Mind & Language, 31*, 37-66.

Christensen, W., Sutton, J., & Bicknell, K. (2019). Memory systems and the control of skilled action. *Philosophical Psychology, 32*(5), 693-719. https://doi.org/10.1080/09515089.2013.766789

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy, 78*, 67-90.

Churchland, P. M. (1988). Perceptual plasticity and theoretical neutrality: A reply to Jerry Fodor. *Philosophy of Science, 55*(2), 167-187. https://doi.oresg/10.1086/289425

Churchland, P. M. *Plato's camera: How the physical brain captures a landscape of abstract universals*. Cambridge: MIT Press.

Churchland, P. S., & Sejnowski, T. (1992). *The Computational Mind.* Cambridge, MA: MIT Press.

Clark, A. (1996). Dealing in futures: Folk psychology and the role of representations in cognitive science. In Robert N. McCauley (ed.), *The Churchlands and their critics*. (pp. 86-103). Cambridge, MA: Blackwell.

Clark, A. (1997). *Being there: putting brain, body, and world together again.* Cambridge, M.A, MIT Press.

Clark, A. (1998). Magic words: How language augments human computation. In P. Carruthers and J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 162-183). Cambridge, UK: Cambridge University Press.

Clark, A. (2005). Intrinsic content, active memory and the extended mind. *Analysis, 65*, (1), 1-11.

Clark, A. (2007). Curing cognitive hiccups: a defense of the extended mind. *The Journal of Philosophy, 104*(4), 163-192. https://doi.org/10.1080/00207727008920220

Clark, A. (2008). *Supersizing the Mind:* Embodiment*, Action*. Oxford: Oxford University Press.

Clark, A. (2010). Coupling, constitution and cognitive kind: A reply to Adams and Aizawa. In R. Menary (Ed.), *The Extended Mind*. (pp. 81-99). Cambridge, MA: MIT Press.

Clark, A. (2015). What 'extended me' knows. *Synthese*, *192*(11), 3757 - 3775. https://doi.org/10.1080/00207727008920220

Clark, A. (2016) *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.

Clark, A. (2020). Beyond desire? Agency, choice, and the predictive mind. Australasian *Journal of Philosophy, 98*(1), 1-15. https://doi.org/10.1080/00207727008920220

Clark, A. and Chalmers, D. (1998). The Extended Mind. *Analysis*, (58). 7-19.

Cohen, A. P. (1994). *Self consciousness: An alternative anthropology of identity*. London, UK: Routledge.

Conant, R., & Ashby, W. R., (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science, 1*(2), 89–97. https://doi.org/10.1080/00207727008920220

Craik, K. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.

Crane, T. (1988). The waterfall illusion, *Analysis, 48*, 142-7.

Crane, T. (1990). The language of thought: No syntax without semantics. *Mind and Language, 5*(3), 187-212.

Crane, T. (1992). The nonconceptual content of experience. In T. Crane (ed.) *The Contents of experience: Essays on perception*. (pp. 136-57). Cambridge: Cambridge University Press,

Crane, T. (2013). The given. In J. Shear (Ed.) *Mind, reason and being-in-the-world: The McDowell-Dreyfus debate*. (pp. 229 – 249). London: Routledge Press.

Csikszentmihalyi, M. (1990). *Flow: the psychology of optimal experience.* New York: Harper and Row.

Cummins, R. (1996). *Representations, targets, and attitudes*. Cambridge: MIT Press.

Currey, M. (2013). *Daily rituals: How artists work*. New York: Knopf.

Cussins, A. (1992). Content, embodiment and objectivity: The theory of cognitive trails. *Mind,* 101, pp. 651-688.

Cussins, A. (2012). Environmental representation of the body. *Review of Philosophy and Psychology, 3*(1), 15-32. https://doi.org/10.1080/00207727008920220

Deamer, F. (2021). Why do we talk to ourselves? *Review of Philosophy and Psychology, 12*, 425-433. https://doi.org/10.1007/s13164-020-00487-5

De Beauvoir, S. (1949). *The second sex*. New York: Vintage Books.

De Haan, S. E. (2020). *Enactive psychiatry*. Cambridge: Cambridge University Press.

Dennett, D. C. (1969). *Content and Consciousness*. London: Routledge and Kegan Paul.

Dennett, D. C. (1978). *Brainstorms: philosophical essays on mind and psychology.* Cambridge, MA: MIT Press.

Dennett, D. (1990). The myth of original intentionality. In K. A. Mohyeldin Said, W. H. Newton-Smith, R. Viale & K. V. Wilkes (Eds.), *Modelling the mind* (pp. 43–62). Oxford: Oxford University Press.

Dennett, D. C. (1991). *Consciousness explained.* Boston: Little Brown.

Dennett, D. C. (1996). *Kinds of minds: Towards an understanding of consciousness.* London: Weidenfeld and Nicolson.

Dennett, D. C. (2003). *Freedom Evolves.* London: Penguin.

Dennett, D. C. (2015). *Forward: Writing Contents and Consciousness*. In F. De Brigard and C. Muñoz-Suárez (Eds.), *Content and Consciousness Revisited.* (pp. v-x). New York: Springer

Devitt, M. (1990). "A narrow representational theory of mind." In *Mind and Cognition*, Lycan, W. (Ed.). (pp. 371-398). Oxford: Basil Blackwell.

Devitt, M. (2011). Methodology and the Nature of Knowing How. *The Journal of Philosophy, 108*(4), 205-218. https://doi.org/10.1080/09515089.2013.766789

Dewhurst, J. (2018). British Cybernetics. In Sprevak, M. & Colombo, M. (eds). *The Routledge Handbook of the Computational Mind.* (pp. 38-51). Abingdon: Routledge.

Di Paolo, E., Buhrmann, T., & Barandiaran, X. (2017). *Sensorimotor life: an enactive proposal.* Oxford: Oxford University Press.

Di Paolo, E., Thompson, E., & Beer, R. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences, 3*. https://doi.org/10.1080/00207727008920220

Dishon-Berkovits, M. & Algom, D. (2000). The stroop effect: It is not the robust phenomenon that you have thought it to be. *Memory & Cognition, 28*(8), 1437-1449.

Drayson, Z. (2010). Extended cognition and the metaphysics of mind. *Cognitive Systems Research*. *11* (4), 366-377.

Drayson, Z. (2012). The uses and abuses of the personal / subpersonal distinction. *Philosophical Perspectives 26*(1), 1-18. https://doi.org/10.1080/09515089.2013.766789

Drayson, Z. (2014). The personal/subpersonal distinction. *Philosophy Compass, 9*(5), 338-346. https://doi.org/10.1080/09515089.2013.766789

Drayson, Z. and Clark, A. (2020). Cognitive disability and embodied, extended minds. In Wasserman, D. and Cureton, A. (eds.) *Oxford Handbook of Philosophy and Disability*. 580-597. Oxford University Press. https://doi:10.1093/oxfordhb/9780190622879.013.10

Dretske, F. L. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.

Dretske, F. L. (2000). Norms, history, and the constitution of the mental. In F. L. Dretske (Ed.) *Perception, knowledge and belief: Selected essays*, 242-58. Cambridge: Cambridge University Press.

Dreyfus, H. L. and Dreyfus, S. (1986). *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York: Free Press.

Dreyfus, H. L. (2007). The return of the myth of the mental. *Inquiry, 50*(4), 352-365. https://doi.org/10.1080/00201740701489245

Dunning, J., & Kruger, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121-34. https://doi.org/10.1037/0022-3514.77.6.1121

Elton, M. (2000). The personal/sub-personal distinction: An introduction. *Philosophical Explorations, 3*(1), 2-5. https://doi.org/10.1080/13869790008520977

Elzinga, B. (2018). Self-regulation and knowledge how. *Episteme, 15*(1), 119-140. https://doi:10.1017/epi.2016.45

Elzinga, B. (2021). Intellectualizing know how. *Synthese, 198*(2)*,* 1741-1760. https://doi.org/10.1080/09515089.2013.766789

Esteban-Guitart, M. (2014). Appropriation. In T. Teo (Ed.), *Encyclopedia of Critical Psychology*. New York: Springer. https://doi.org/10.1007/978-1-4614-5583-7_616

Evans, G. (1982). *The Varieties of Reference*. Oxford: Oxford University Press.

Evans, J. S. B. T, & Frankish, K. (Eds.). (2009). *In two minds: Dual processes and beyond*. Oxford, UK: Oxford University Press.

Fernyhough, C. (2004). Alien voices and inner dialogue: Towards a developmental account of auditory verbal hallucinations. *New Ideas in Psychology*, *22*, 49-68. https://doi.org/10.1016/j.newideapsych.2004.09.001

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences, 39*:e229, 1-72. https://doi.org10.1017/S0140525X15000965

Fish, W. (2009). *Perception, Hallucination, Illusion*. Oxford: Oxford University Press.

Flavell (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry', *American Psychologist* 3*4,* 906-11.

Forster, E. M. (1927). *Aspects of the Novel*. London: Edward Arnold.

Fleming, S. (2021). *Know thyself: The science of self-awareness*. Basic Books.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8* article: 443. https://doi.org/10.1080/00207727008920220

Fodor, J. A. (1975). *The language of thought.* Cambridge, MA: Harvard University Press.

Fodor, J. A. (1983). The modularity of mind: An essay on faculty psychology. Cambridge, MA: MIT Press.

Fridland, E. (2015). Skill, nonpropositional thought, and the cognitive penetrability of perception. *Journal for General Philosophy of Science, 46*(1), 105-120.

Fridland, E. (2017). Automatically minded. *Synthese,* 194, 4337-4363. https://doi.org/10.1007/s11229-014-0617-9.

Fridland, E. (2019). Longer, smaller, faster, stronger: On skills and intelligence. *Philosophical Psychology, 32*(5), pp. 760-784. https://doi.org/10.1080/09515089.2019.1607275

Fridland, E. (2021). Skill and strategic control. *Synthese*, 199, 5937-5964. https://doi.org/10.1007/s11229-021-03053-3

Frith, C. (2012). Explaining delusions of control: the comparator model 20 years on. *Consciousness and Cognition, 21*(1), 52-54.

Gallagher, S. (2005). *How the Body Shapes the Mind.* Oxford: Oxford University Press.

Gallagher, S. (2017a). *Enactivist Interventions: Rethinking the Mind*. Oxford: Oxford University Press.

Gallagher, S. (2018). The Extended Mind: State of the Question. *The Southern Journal of Philosophy*, *56* (4), 421-427. https://doi.org/10.1111/sjp.12308

Gallagher, S. (2021). *Performance/Art: The Venetian lectures*. Milan: Mimesis International Edizioni.

Gallagher, S. & Allen, M. (2018). Active inference, Enactivism, and the Hermeneutics of Social Cognition. *Synthese. 195*(6), 2627-48. https://doi.org/10.1080/09515089.2013.766789

Gallagher, S. & Zahavi, D. (2021). *The Phenomenological Mind: An introduction to philosophy of mind and cognitive science* (3rd edition). London: Routledge.

Gauker, C. (2018). Inner speech as the internalization of outer speech. In P. Langland-Hassan & A. Vicente (Eds.), *Inner speech: New voices* (pp. 53-77). Oxford: Oxford University Press.

Gertler, B. (2007). Overextending the mind. In B. Gertler & L. Shapiro (Eds.), *Arguing about the mind* (pp. 192-206). Routledge.

Geurts, B. (2018). Making sense of self talk. *Review of Philosophy and Psychology, 9*(2), 271-285. https://doi.org/10.1007/s13164-017-0375-y

Gibbard, A. (1975). Contingent identity. *Journal of Philosophical Logic, 4,* 187-221.

Gibson, E., & Rader, N. (1979) Attention. In Hale G.A., Lewis M. (eds) *Attention and Cognitive Development*. Springer, Boston, M.A.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston, Houghton Mifflin.

Gignac, G. E, & Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence, 80,* article: 101449. https://doi.org/10.1080/00207727008920220

Goldin-Meadow, S. (2003). *Hearing Gesture: How Our Hands Help Us Think*. Cambridge, MA: Harvard University Press.

Goldinger, S. D., Papesh, M. H., Barnhart, A. S., Hansen, W. A., & Hout, M. C. (2016) The poverty of embodied cognition. *Psychonomic Bulletin & Review, 23,* 959-978. https://doi.org/10.1080/09515089.2013.766789

Gopnik, A. I. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioural and Brain Sciences, 16*(1), 1-14, 29 - 113.

Goupil, L. & Proust, J. (2022). Curiosity as a metacognitive feeling. PsyArXiv. https://doi:10.31234/osf.io/c8a6t

Graham, G. (2019). Behaviorism. *Stanford encyclopedia of Psychology* (E. N. Zalta (Ed.). Stanford University. https://doi.org/10.1080/09515089.2013.766789

Greif, H. (2017). What is the extension of the extended mind? *Synthese, 194*(11), 4311-4336. https://doi.org/10.1080/09515089.2013.766789

Grice, H. P. (1975). Logic and conversation, in P. Cole & J. L. Morgan (eds.) *Syntax and Semantics* 3*: Speech Acts* (pp. 41-58). New York: Academic Press.

Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing ourselves together: The cultural origins of metacognition. *Trends in Cognitive Sciences, 24*(5), 349-362. https://doi.org/10.1016/j.tics.2020.02.007

Huebner, B. (2018). Picturing, signifying, and attending. *Belgrade Philosophical Annual*, *31,* 7-40. https://doi.org/10.1080/00207727008920220

Hurlbert, R. T., Heavey, C. L., & Kelsey, J. M. (2013). Toward a phenomenology of inner speaking. *Consciousness and Cognition, 22*(4), 1477-1494. https://doi.org/10.1016/j.concog.2013.10.003

Hurley, S. L. (1998). *Consciousness in Action*. Cambridge, MA: Harvard University Press.

Hutchins, E. (1995a). How a cockpit remembers its speeds *Cognitive science, 19*, 265-288.

Hutchins, E. (1995b). *Cognition in the wild*. Cambridge, MA: MIT Press.

Hutchins, E. (2014). The cultural ecosystem of human cognition. *Philosophical Psychology, 27*(1), 34-49. https://doi.org/10.1080/09515089.2013.766789

Hutto, D. D. (2008). *Folk psychological narratives: The sociocultural basis of understanding reasons.* Cambridge, MA: MIT Press.

Hutto, D. D., & Satne, G. L. (2015). The natural origins of content. *Philosophia, 43*(3), 521-536. http://dx.doi.org/10.1007/s11406-015-9644-0

Hutto, D.D. & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. Cambridge, MA: MIT Press.

James, W. (1890). *Principles of psychology*. New York: Dover Publications.

Kelso, J. A. S. (1995). Dynamic patterns: The self-organization of brain and behaviour. Cambridge, MA: MIT Press.

Kim, J. (2005). *Physicalism, or something near enough.* Princeton: Princeton University Press.

Kim, S., Shahaeian, A. & Proust, J. (2018). Developmental diversity in mindreading and metacognition. In J. Proust & M. Fortier (Eds.), *Metacognitive diversity: An interdisciplinary approach* (pp. 97-133). Oxford: Oxford University Press.

Kim, S., Sodian, B., Paulus, M., Senju, M., Okuno, A., Ueno, M., Itakura, S., & Proust, J. (2020). Metacognition and mindreading in young children: A cross-cultural study. *Consciousness and Cognition, 85*, 103017. https://doi.org/10.1016/j.concog.2020.103017

Kirchhoff, M. D. (2015). Extended cognition & the causal-constitutive fallacy: in search for a diachronic and dynamical conception of constitution. *Philosophy and Phenomenological Research, 90*(20), 320-360. https://doi.org/10.1080/00207727008920220

Kirchhoff, M. D., and Kiverstein, J. (2019). *Extended consciousness and predictive processing: A third-wave view*. Oxford: Routledge.

Kirchhoff, M. D., and Kiverstein, J. (2020). Attuning to the world: The diachronic constitution of the extended conscious mind. *Frontiers in Psychology, 11*:1966 https://doi.org/10.1080/00207727008920220

Kirsh, D. (2004). Metacognition, distributed cognition and visual design. In P Gardinfors & P Johansson (eds.), *Cognition, education, and communication technology*. (pp. 147-180). Hillsdale, NJ: Erlbaum.

Kirsh, D. and Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*(4), 513-549. https://doi.org/10.1207/s15516709cog1804_1

Knappett, C., Malafouris, L., & Tomkins, P. (2010). Ceramics (as Containers). In D. Hicks & M. C. Beaudry (Eds.), *The Oxford Handbook of Material Culture Studies*. (pp. 588 – 612). Oxford University Press.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*(4), 609-639. https://doi.org/10.1037/0033-295X.100.4.609

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition, 9*(2) 149-171. https://doi.org/10.1006/ccog.2000.0433

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M, Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness*. pp. 289-325. Cambridge University Press. https://doi.org/10.1017/CBO9780511816789.012

Koriat, A., & Levy-Sadot, R. (1999). Processes underlying metacognitive judgements: Information-based and experience-based monitoring of one's own knowledge. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology*, pp. 483-502. Guilford: New York Publications.

Langland-Hassan, P. (2014). Unwitting self-awareness? *Philosophy and Phenomenological Research*, *89*, 719-726.

León, F. and Zahavi, D. (2022). Consciousness, philosophy and neuroscience. *Acta Neurochir*. Advance online publication. https://doi.org/10.1080/00207727008920220

Levin, M. & Dennett, D. C. (2020). Cognition all the way down. *Aeon* https://aeon.co/essays/how-to-understand-cells-tissues-and-organisms-as-agents-with-agendas. Accessed 9th March 2022.

Lock, A. (1980). *The guided reinvention of language*. London: Academic Press.

Logan, G. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology, 39*(2), 367– 386.

Macpherson, F. (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research*, *84*(1), 24-62. https://doi.org/10.1111/j.1933-1592.2010.00481.x

Macpherson, F. (2015). Cognitive penetration and nonconceptual content. In J. Zeimbekis & Raftopoulos (eds.), *The cognitive penetrability of perception: New philosophical Perspectives* (pp. 331-59). Oxford: Oxford University Press.

Maley, C. J. (2011). Analog and digital, continuous and discrete. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition,* 155(1), 117-131. https://doi.org/10.1007/s11098-010-9562-8

Maley, C. J. (2018). Toward analog neural computation. *Minds & Machines, 28*(1). 77-91. https://doi.org/10.1007/s11023-017-9442-5

Maley, C. J. (2021). The physicality of representation. *Synthese, 199*(5), 14725-14750. https://doi.org/10.1007/s11229-021-03441-9

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

McCulloch, W., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology, 52*(1-2), 99-115. https://doi.org/10.1007/BF02459570

McDowell, J. (1994). *Mind and World*. Cambridge, MA: Harvard University Press.

McDowell, J. (2008). Avoiding the myth of the given. In J. Lindgaard (ed.)*, John McDowell: Experience, norm, and nature.* (pp. 1-14). New York: John Wiley & Sons.

McIntosh, R. D., Fowler, E. A., Lyu, T., & Della Salla, S. (2019). Wise up: Clarifying the role of metacognition in the Dunning-Kruger effect. *Journal of Experimental Psychology: General*, *148*(11), 1882-1897. https://doi.org/10.1037/xge0000579

McGeer, V. (2007). The regulative dimension of folk psychology. In D. D. Hutto & M. Ratcliffe (Eds.). *Folk-Psychology re-assessed* (pp. 137-156). Dordrecht: Springer.

Menary, R. (2006). Attacking the bounds of cognition. *Philosophical Psychology, 19*, 329-344. https://doi.org/10.1080/09515080600690557

Menary, R. (2007). *Cognitive integration: Mind and cognition unbounded.* Basingstoke: Palgrave, Macmillan.

Menary, R. (2010a). Cognitive integration and the extended mind. In R. Menary (ed.), *The extended mind* (pp. 227-44). Cambridge, MA: MIT Press.

Menary, R. (2010b). The holy grail of cognitivism: A response to Adams and Aizawa. *Phenomenology and the Cognitive Sciences, 9,* 605-618. https://doi.org/10.1007/s11097-010-9185-8

Menary, R. (2015). Mathematical cognition - A case for enculturation. In T. Metzinger & J. M. Windt (Eds.) *Open MIND: 25*, 1-22. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570818

Menary, R. & Gillett, A. J. (2017). Embodying culture: integrated cognitive systems and cultural evolution. In J. Kiverstein (Ed.), *The Routledge Handbook of Philosophy of the Social Mind* (pp. 72 - 87). New York: Routledge.

Menary, R. & Kirchhoff, M. (2014). Cognitive transformations and extended expertise. *Educational Philosophy and Theory, 46*, 610-623. https://doi.org/10.1080/09515089.2013.766789

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Cambridge, MA: Harvard University Press.

Merleau-Ponty, M. (1945 / 2012). *Phenomenology of Perception.* (Donald Landes, Trans.). Routledge.

Miłkowski, M. (2010). Making Naturalised Epistemology (Slightly) Normative. In K. Talmont-Kamiński & M. Miłkowski (Eds.) *Beyond description: Naturalism and normativity.* pp. 73-84. London, UK: College Publications.

Miłkowski, M., Clowes, R., Rucińska, Z., Przegalińska, A., Zawidzki, T., Krueger, J., Gies, A., McGann, M., Afeltowicz., Wachowski, W., Stjernberg, F., Loughlin, V., and Hohol, M. (2018). From wide cognition to mechanisms: a silent revolution. *Frontiers in psychology,* 9, 2393. https://doi.org/10.1080/09515089.2013.766789

Montero, B. G. (2010). Does bodily awareness interfere with highly skilled movement? *Inquiry, 53*(2) 105-122.

Montero, B. G. (2016). *Thought in action: Expertise and the conscious mind*. Oxford: Oxford University Press.

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin, 132*(2), 297-326.

Neisser, U. (1988). Five kinds of self-knowledge. *Philosophical Psychology, 1*(1), 35-59.

Nelson, T. O. & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *Psychology of Learning and Motivation, 26.* pp. 125-173. New York: Academic Press.

Newell, A. (1973). You can't play twenty questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing*. (pp. 1-26). New York: Academic Press.

Newen, L. De Bruin, & S, Gallagher (2018), *The Handbook of 4E Cognition*. (First edition) Oxford, UK: Oxford University Press.

Nicholson, T. Williamson, D. M., Grainger, C., Lind, S. E., & Carruthers, P. (2019). Relationships between implicit and explicit uncertainty monitoring and mindreading: Evidence from autism spectrum disorder. *Consciousness and Cognition, 70*, 11-24. https://doi.org/10.1016/j.concog.2019.01.013

Nielsen, K. & Ward, T. (2020). Mental disorder as both natural and normative: Developing the normative dimension of the 3e conceptual framework for psychopathology. Journal of Theoretical and Philosophical Psychology, 40(2), 107-123. https://doi.org/10.1037/teo0000118

Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.

Noë, A. (2012). *Varieties of Presence*. Cambridge, MA: Harvard University Press.

Noë, A. (2015). *Strange Tools: Art and Human Nature.* New York: Hill and Wang.

Nussinson, R. & Koriat, A. (2008). Correcting experience-based judgements: The perseverance of subjective experience in the face of the correction of judgement. *Metacognition and Learning, 3*(2), 159-174. https://doi.org/10.1007/s11409-008-90

O'Brien, G., & Opie, J. (2015). Intentionality lite or analog content? A response to Hutto and Satne. *Philosophica, 43*, 723-729. https://doi.org/10.1007/s11406-015-9623-5

Perner, J. (1991). *Understanding the Representational Mind*. Cambridge, MA: MIT Press.

Piccinini, G. (2020). Neurocognitive mechanisms: explaining biological cognition. Oxford: Oxford University Press.

Piccinini, G. (2022). Situated neural representations: Solving the problems of content. *Frontiers in Neurorobotics*. 16:846979. https://doi.org/10.3389/fnbot.2022.846979

Pöyhönen, S. (2014). Explanatory power of extended cognition. *Philosophical Psychology*, *27*(5), 735-759. https://doi.org/10.1080/09515089.2013.766789

Proust, J. (2001). A plea for mental acts. *Synthese, 129*, 105-128. https://doi.org/10.1023/A:1012651308747

Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese, 159*(2), 271-295. https://doi.org/10.1080/09515089.2013.766789

Proust, J. (2009). My answers to five questions on agency. In J. Aguilar, & A. A. Buckareff (Eds.) *Philosophy of action: 5 questions*.

Proust, J. (2013). *The Philosophy of Metacognition: Mental Agency and Self-Awareness.* Oxford: Oxford University Press.

Proust, J. (2014). Epistemic action, extended knowledge, and metacognition. *Philosophical Issues, 24*(1)*,* 364-992. https://doi.org/10.1080/00207727008920220

Proust, J. (2015). Time and action: Impulsivity, habit, strategy. *Review of Philosophical Psychology, 6*(4), 717-743. https://doi.org/10.1080/00207727008920220

Putnam, H. (1975). The Meaning of Meaning in K Gunderson (ed.) *Language, Mind and Knowledge, Minnesota Studies in the Philosophy of Science*, 7, (pp. 131-93). Minneapolis: University of Minnesota Press.

Pylyshyn, Z. W. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences, 22*(3), 341-365.

Ramsey, W. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Rietveld, E. (2008). Situated normativity: the normative aspect of embodied cognition in unreflective action. *Mind, 117*, 973-1001.

Risko, E. J. and Gilbert, S. J. (2016). Cognitive Offloading. *Trends in Cognitive Science*, *20* (9) 676-688. https://doi.org/10.1080/00207727008920220

Rorty, R. (1979). *Philosophy and the mirror of nature*. Princeton: Princeton University Press.

Roth, M. (2015). I am large, I contain multitudes: The personal, the subpersonal, and the extended. In F. De Brigard and C. Muñoz-Suárez (Eds.), *Content and Consciousness Revisited.* (pp. 129-142). New York: Springer.

Rowlands, M. (1999). The body in mind: Understanding cognitive processes. Cambridge: Cambridge University Press.

Rozenblit, L. & Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science, 26*(5), 521-562. https://doi.org/10.1080/00207727008920220

Rupert, R. D. (2009). *Cognitive systems and the extended mind.* Oxford: Oxford University Press.

Rupert, R. D. (2018). The self in the age of cognitive science: Decoupling the self from the Personal level. *Philosophic Exchange, 47*(1), article 2, 1-36.

Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.

Sass, L. (1992). *Madness and Modernism: Insanity in the light of modern art, literature and thought*. New York: Basic Books.

Schneider, W. & Shiffrin, R. M. (1977). Controlled and automatic information processing. I. Detection, search, and attention. *Psychological Review, 84*(1), 1-66.

Schwartz, A., & Drayson, Z. (2019). Intellectualism and the argument from cognitive science. *Philosophical Psychology, 32*(5), 661 - 691.

Sellars, W. S. (1956). Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science, 1,* 253-329.

Shapiro, L. (2019). *Embodied cognition* (2nd edition). London, UK: Routledge.

Shea, N. (2012). Reward prediction errors are meta-representational. *Nous, 48*(2), 314-341. https://doi.org/10.1111/j.1468-0068.2012.00863.x

Shea, N. (2018). *Representation in cognitive science*. Oxford: Oxford University Press.

Shea, N. (2020). Functionalist interrelations amongst human psychological states inter see, ditto for martians. In J. Smortchkova, K. Dołęga, & T. Schlicht, (Eds.), *What are mental representations?* pp. 242-253. Oxford: Oxford University Press.

Shvarts, A. & Bakker, A. (2019). The early history of the scaffolding metaphor: Bernstein, Luria, Vygotsky, and before. *Mind, Culture, and Activity, 16*(1), 4-23. https://doi.org/10.1080/00207727008920220

Simon, H. (1969). *The science of the artificial*. Cambridge, MA: MIT Press.

Smart, P. R. (2022). Toward a mechanistic account of extended cognition. *Philosophical Psychology.* Advance online publication. https://doi.org/10.1080/00207727008920220

Sperber, D. and Wilson, D. (1986/1996). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Sprevak, M. (2009). Extended cognition and functionalism. *Journal of Philosophy, 106*(9), 503-527. https://doi.org/10.1080/00207727008920220

Sprevak, M. (2019). Extended Cognition. Published in T. Crane (Ed.) *The Routledge Encyclopedia of Philosophy Online*. London: Routledge.

Stanley, J. (2011). *Know how.* Oxford: Oxford University Press.

Sterelny, K. (2003). *Thought in a Hostile World: The Evolution of Human Cognition*. Oxford, UK: Blackwell.

Sterelny, K. (2010). Minds: extended or scaffolded? *Phenomenology and the Cognitive Sciences 9*(4), 465-481. https://doi.org/10.1080/00207727008920220

Sterelny, K. (2012). *The Evolved Apprentice*: *How Evolution Made Humans Unique.* Cambridge, MA: MIT Press.

Sterelny, K. (2021). *The Pleistocene Social Contract: Culture and Cooperation in Human Evolution.* Oxford: Oxford University Press.

Sterelny, K., & Planer, R. (2021). *From signal to symbol: The evolution of language*. Cambridge, MA: MIT Press.

Stich, S. (1978). Beliefs and subdoxastic states. *Philosophy of Science 45*(4), 499 - 518.

Stokes, D. (2021a). On perceptual expertise. *Mind and Language,* 36(2), 241-263. https://doi.org/10.1111/mila.12270

Stokes, D. (2021b). *Thinking and perceiving*. London: Routledge.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18(*6), 643-662. https://doi.org/10.1037/h0054651

Sun, X., Zhu, C., & So, S. (2017). Dysfunctional metacognition across psychopathologies: A meta-analytic review. *European Psychiatry, 45*, 139-153. https://doi:10.1016/j.eurpsy.2017.05.029

Sutton, J. (1998). *Philosophy and memory traces: Descartes to connectionism*. Cambridge: Cambridge University Press.

Sutton, J., Harris, C. B., Keil, P. G., & Barnier, A. J. (2010). The psychology of memory, extended cognition, and socially distributed remembering. *Phenomenology and the Cognitive Sciences, 9,* 521-560.

Sutton, J. (2010). Exograms and interdisciplinarity: History, the extended mind, and the civilising process. In R, Menary (Ed). *The extended mind*, (pp. 189-225). Cambridge, MA: MIT Press.

Sutton, J. (2015). Remembering as public practice: Wittgenstein, memory, and distributed cognitive ecologies. *Mind, language, and action,* 409-433.

Sutton, J., McIlwain, D., Christensen, W., & Geeves, A. (2011). Applying intelligence to the reflexes: Embodied skills and habits between Dreyfus and Descartes. *Journal of the British Society for Phenomenology, 42*(1) 78-103. https://doi.org/10.1080/09515089.2013.766789

Thomas, N. J. T. (1999). Are theories of imagery theories of imagination? An active perception approach to conscious mental content. *Cognitive Science, 23*(2), 207-245.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology and the Sciences of Mind.* Cambridge, MA: Harvard University Press.

Thompson, E. (forthcoming). What's in a concept? Conceptualising the nonconceptual in Buddhist philosophy and cognitive science. In C. Coseru (Ed.), *Reasons and empty persons: Mind, metaphysics, and morality: Essays in honor of Mark Siderits*. London: Springer.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review, 55*(4), 189-208. https://doi.org/10.1037/h0061626

Tomasello, M. (1999). *The cultural origins of human cognition.* Cambridge: Harvard University Press.

Tomasello, M. (2014). *A natural history of human thinking.* Cambridge: Harvard University Press.

Tomasello, M. (2019). *Becoming human: A theory of ontogeny*. Belknap Press of Harvard University Press.

Toner, J., and Moran, A. (2020). Exploring the Orthogonal Relationship between Controlled and Automated Processes in Skilled Action. *Review of Philosophy and Psychology*, 1–17. https://doi.org/10.1007/s13164-020-00505-6

Trevarthen, C. B. (1979). Communication and cooperation in early infancy: A description of primary intersubjectivity. In M. Bullowa (ed.), *Before speech* pp. 321-348. Cambridge, MA: Cambridge University Press.

Turvey, M. T., Shaw, R. E., Reed, E. D., & Mace, W. M. (1981). Ecological laws of perceiving and acting: In reply to Fodor and Pylyshyn. *Cognition, 9*, 237-304.

Van Rooj, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682-697. https://doi.org/10.1177/1745691620970604

Varela, F., Thompson, E., and Rosch, E. 1991. *The embodied mind.* Cambridge, MA: MIT Press.

Vuorre, M., & Metcalfe, J. (2021). Measures of relative metacognitive accuracy are confounded with task performance in tasks that permit guessing. *Metacognition and Learning.* Advance online publication. https://doi.org/10.1007/s11409-020-09257-1

Vygotsky, L. S. (1962). *Thought and Language.* (E. Hanfmann and G. Vakar, Trans.) Cambridge, Mass: MIT Press. (Original work published 1934).

Vygotsky, L. S. (1978). *Mind in Society: The development of higher psychological processes.* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Trans.) Cambridge, MA: Harvard University Press. (Original work published 1930).

Ward, D. (2012). Enjoying the spread: Conscious externalism reconsidered. *Mind, 121*(483), 731-751. https://doi.org/10.1093/mind/fzs095

Ward, D., Silverman, D., & Villalobos, M. (2017). Introduction: the varieties of enactivism. *Topoi, 36,* 365-375. https://doi.org/10.1080/09515089.2013.766789

Wegner, D. (1987). Transactive memory: A contemporary analysis of group mind. In B. Mullen & G. R. Goethals (Eds.), *Theories of group behaviour*. pp. 185-208. New York: Springer-Verlag.

Westfall, M. (forthcoming). Constructing persons: On the personal-subpersonal distinction. *Philosophical Psychology*. Advance online publication. https://doi.org/10.1080/09515089.2022.2096431

Wheeler, M. (2005). *Reconstructing the cognitive world: the next step.* Cambridge, MA: MIT Press.

Wheeler, (2011). In search of clarity about parity. *Philosophical Studies, 152*, 417-425. https://doi.org/10.1080/09515089.2013.766789

Wheeler, M. (2019). The reappearing tool: transparency, smart technology, and the extended mind. *AI & Society*, *34*, 857-866. https://doi.org/10.1080/09515089.2013.766789

Wilkinson, S. (2020). The agentive role of inner speech in self-knowledge. *Teorema, 39*(2) 7-26.

Wilkinson, S. & Fernyhough, C. (2018). When inner speech misleads. In P. Langland-Hassan & A. Vicente (Eds.), *Inner speech: New voices*. (pp 244-260). Oxford: Oxford University Press.

Williams, D. (2018). Predictive processing and the representation wars. *Minds and Machines, 28*, 141-172. https://10.1007/s11023-017-9441-6

Williams, D. (2020). Predictive coding and thought. *Synthese, 197*(4), 1749-1775. https://10.1007/s11229-018-1768-x

Williams, D., and Colling, L. (2018). From symbols to icons: the return of resemblance in the cognitive neuroscience revolution. *Synthese, 195*, 1941-1967. https://doi.org/10.1007/s11229-017-1578-6

Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks, 11*, 1317-1329.

Young, I. M. (1980). Throwing like a girl: A phenomenology of feminine body comportment motility and spatiality. *Human Studies, 3*(2), 137-156.

Zahavi, D. (2005). *Subjectivity and selfhood: Investigating the first-person perspective.* Cambridge: MIT Press.

Zahavi, D. (2013). Mindedness, mindlessness and first-person authority. In J. K. Schear (ed.), *Mind, Reason, and Being-in-The-World: The McDowell-Dreyfus Debate* (pp. 320-40). London: Routledge.

Zahavi, D. (2014) *Self and Other: Exploring Subjectivity, Empathy, and Shame*. Oxford: Oxford University Press.

Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. Cambridge: MIT Press.

Zawidzki, T. W. (2021. A new perspective on the relationship between metacognition and social cognition: metacognitive concepts as socio-cognitive tools. *Synthese*, 198, 6573-6596. https://doi.org/10.1007/s11229-019-02477-2

Zeki, S. (1999). Splendours and miseries of the brain. *Philosophical Transactions: Biological Sciences,* 354, (1392), 2053-2065. https://doi.org/10.1080/09515089.2013.766789