Chapter 1

Digital Epistemology: A Research Programme Motivated

1.1 Introduction

The very idea of 'digital epistemology' might sound perplexing. Epistemology – the philosophical theory of knowledge – is about things in the *head*, right? It certainly has been so far[1], but it's less clear that's how things should stay, especially given where most of our information is – and how it's being generated – these days.

This chapter aims to give the reader a sense of what 'digital epistemology' is all about, and why it is a research programme worth pursuing *alongside* traditional epistemology, as a way to theorise in a principled way about knowledge as a multiply realisable kind, one that is often but *not* always stored (or in some cases, not even produced) in the head.

[1] Questions under the description of 'epistemology' – in the tradition of mainstream analytic epistemology – of the 20th and 21st centuries (to the point of writing, in 2021) – have invariably concerned brainbound cognition. This is reflected in the choices of topics covered in leading epistemology anthologies and textbooks e.g., (Sosa, Fantl, and McGrath 2019; Pritchard 2013; Steup and Neta 2005; Bernecker and Dretske 2000; Dancy, Sosa, and Steup 1992; Dancy 1985; Chisholm 1977; Littlejohn and Carter 2019).

Here is the plan. §1.2, 'Where is all the knowledge?' discusses some motivations for broadening our toolkit in epistemology in order to address new questions raised by the largely digitised ways in which we are nowadays storing old information and also generating new information. §1.31 subjects the 'intracranialist dogma' in mainstream epistemology to some critical scrutiny, by showing why it is that a traditional (brainbound) picture of cognition needn't be assumed at all to make sense of the entailment relationship between knowledge and belief. §1.32, 'The Falling Pillars of Cartesianism', explains how 'extruding' epistemology from the skull and skin is really just a natural point on a progression we're already on, and which has – since the mid 1970s – seen other forms of internalism (content internalism and epistemic internalism) lose the grip they once had on our thinking about knowledge. §1.4, 'Desiderata for a 21st Century Epistemology: A Shortlist' - lays out a research agenda for 'digital epistemology', by identifying a range of key questions that this book will set out to answer (or at least begin to answer), unshackled by the cognitive internalist dogma that knowledge-apt cognition is materially realised always and only by brainbound processes.

A problem that will be a central theme of the book (Chapters 2-4) is to make sense of how *bona fide* propositional knowledge can be stored outside the head, in our gadgets, and how such 'extended' knowledge differs from mere digital information we possess, including accurate digital information, that falls short of knowledge. However, as the 'short list' of questions outlined in §1.4 indicates, the book will also tackle epistemological questions related to the cognitively outsourced *generation* of new knowledge through through AI-driven deep reinforcement learning (e.g., Google DeepMind) and via big data and text mining.

Let's begin, though, by thinking a bit about where our most important information is right now. Where exactly are *you* keeping it?

1.2 Where is all the knowledge?

A lot of knowledge is in our *brains* of course, which can store an impressive amount. Let's consider how much, just to get this out of the way.

On the conservative[2] assumption that a human brain has about a billion neurons (each with 1,000 connections to other neurons), a human brain hosts about a trillion connections between the billion neurons it has. How powerful is this, in terms of what we are capable of remembering?

According to cognitive neuroscientist Paul Reber, it's a *lot* of storage: roughly enough to hold 300 million hours of video footage:

> If each neuron could only help store a single memory, running out of space would be a problem. You might have only a few gigabytes of storage space, similar to the space in an iPod or a USB flash drive. Yet neurons combine so that each one helps with many memories at a time, exponentially increasing the brain's memory storage capacity to something closer to around 2.5 petabytes (or a million gigabytes). For comparison, if your brain worked like a digital video recorder in a television, 2.5 petabytes would be enough to hold three million hours of TV shows (Hawes 2010).

With over 7.65 billion people in the world, we have a worldwide (human) brain storage capacity of 19.1 billion petabytes[3], enough to hold the equivalent of 2.3 *million billion* hours of YouTube videos – which is a lot of stored knowledge[4] capacity between us.

This sounds impressive at first. But suppose we ran a search on that 19.1 petabytes. What would we find in there? Or, more to the point, what would we *not* find in there?

[2] For less conservative estimates, see Herculano-Houzel (2009).

[3] For a sense of scale, there are 1,024 terabytes in a petabyte.

[4] Not all of this is knowledge of course; some of us use our storage space to host conspiracy theories, as well as to store a wide variety of other kinds of information, which is more or less useful.

Let's start with simple information that we use to structure our lives across time. For example: think about your plans for the next several weeks. Where are you supposed to be, who are you supposed to meet up with and where, what projects we have agreed to start and to finish (and by when), when does your family needs us to help them out with an errand, what does the task involve, etc. A *lot* of this information is *not* going to turn up in any brain storage search.

As of 2018, nearly 70% of adults store the above kind of life-structuring information in online calendars, which are accessed most regularly via smartphones.[5] A simple reason for using our smartphones this way is that, by offloading this kind of information from brain to digital storage, it frees up to use our on-board cognitive resources for other things.

Using, e.g., Google Calendar or Apple's iCloud Calendar to store life-structuring information is a paradigmatic example of what cognitive scientists refer to as *cognitive offloading*. Put very generally, cognitive offloading can be defined as the 'use of physical action to alter the information processing requirements of a task so as to reduce cognitive demand' (Risko and Gilbert 2016, 677). In the simple case of digital calendars, this physical action (of entering relevant info into our calendars, inviting others to join, etc.) is becoming increasingly fluent and seamless[6] – as Andy Clark notes (2015), the more we offload certain kinds of tasks, the less we notice that we're doing it.[7]

Cognitive offloading helps us overcome capacity limitations, and also to

[5] As reported by an ECAL survey from May 2018. http://ecal.com/70-percent-of-adults-rely-on-digital-calendar/ Accessed on 16 April 2021.

[6] According to studies reported by Grinschl et al. (2020), offloading via mobile phones is more effective when there is no stylus or buttons, but through the kind of touchscreen interfaces that are now standard.

[7] Compare: we don't stop and think 'I will now store this information in my biomemory'. A central idea of Clark's – one that he first developed with David Chalmers in their (1998) paper 'The Extended Mind' (which we will discuss later in the book, in Chapters Two and Three) is that our use of gadgets for cognitive offloading increasingly 'mimics' the transparency by which we rely (uncritically, and unreflectively) on our own biomemories. For discussion of other factors that explain our increased reliance on external memory storage, see Clowes (2013),

minimise computational effort, in order to do things we couldn't otherwise do.[8] By making us more efficient, cognitive offloading has helped improve performance in cases of (along with memory[9], which is freed up) perception[10], spatial reasoning[11], mathematics[12], and even bartending.[13]

Dealing with capacity and computation limitations through cognitive offloading is actually a very old strategy. The Ancient Romans resorted to using other people for this purpose – memory slaves called *graeculi*[14] (who would follow their masters around, letting them borrow their brain storage); likewise, ancient Peruvians used knots (rather than people) called *quipus* as external memory aids.[15]

Plausibly, the ubiquity of cognitive offloading nowadays to our gadgets is best understood not as some revolutionary strategy as some popular writers have positioned it (e.g., Harari 2016), but rather as just the latest manifestation of very natural way to increase (as humans historically always have) our cognitive performance in light of our limitations.[16] A central difference is that, our *opportunities* are now much greater given the power and portability of smartphones, and the consistent improvement of interface design (especially now that touchscreens are the norm — on this see Grinschl et al. (2020)).

For our present purposes, an appreciation of commonsense reasons we have to offload certain cognitive tasks – along with the increased ease by which we can do this efficiently – is important in that it reveals a certain *mismatch between theory and practice*. Whereas memory has always been

[8] (Risko and Gilbert 2016, 676).

[9] Although most statistics about memory offloading pertain to adults, offloading has also increased performance for children in working memory tasks. See, e.g., Berry et al. (2019).

[10] (Ha et al. 2014; Jeffri and Rambli 2021).

[11] (Chu and Kita 2011).

[12] (Goldin-Meadow et al. 2001; Costa et al. 2011).

[13] (Beach 1993).

[14] See Nestojko et al. (2013). For a discussion of Roman slaves as a proto-form of extended cognition, see Wheeler (2018).

[15] For an overview of external memory aids used in ancient Peru, see de Acosta (2002).

[16] For a sustained presentation of this idea, see Clark (2003).

important in *epistemology* – the epistemology of memory is a thriving subfield in its own right[17] – the *kind* of memory that the tools of mainstream epistemology are suited to tackle is *biomemory*. It is specifically biomemory with reference to the main debates about memory-based knowledge are framed.[18] Such debates attempt to explain, among other things, why *some* information stored in your head is *bona fide* knowledge, whereas some is not, and why.

Here, though, is where we can appreciate what is an increasing mismatch between theory and practice. Our best epistemological theories of memory helps us sort the good (knowledge) from the bad (unknown, incorrect info) and *both* from *merely* correct but unknown stored information when it comes to what's stored in *biomemory*. Problem is – and this is where cognitive offloading comes in – a huge chunk of information that we actually rely on to structure our lives simply *isn't stored there anymore*. This is a problem. In order to address it (closing the gap of this theory/practice mismatch), we need to either stop offloading (a bad idea) or expand our theory, in such a way that we can, by expanding it, better distinguish the good from the bad from the merely correct information stored not in our heads *but in our gadgets* (where we're likely to actually find it!).

But here is a further important point worth taking in. *Even if* cognitive offloading is oftentimes strategical and smart (e.g., in so far as it helps us to overcome capacity limitations and minimise computational effort, thus to increase cognitive performance), it is *not* always the optimal strategy for us.[19] In some circumstances, it is better (with respect to optimising performance at the cognitive task) to forego offloading and use biomemory. Although we often opt for what is best between these options, but as

[17] A central dividing issue in the debate concerns whether memory is best understood as preserving or generating positive epistemic status of beliefs stored in memory. For discussion, see Bernecker (2010, 2011), Michelian (2011), Frise (2017), and Senor (2005).

[18] This is particularly the case with the debate between generativism and preservantism; though proponents and opponents of the epistemic theory of memory also take this assumption to be in the background. See Bernecker (2008) for an overview of the metaphysics of memory and its role in these debates.

[19] On this point, see especially Risko and Gilbert (2016, 383–5).

empirical work on offloading shows, sometimes we just plain get it wrong, often for reasons that are unclear.[20] As Timothy Dunn and Evan Risko (2016) describe the state of play, ‘One of the major theoretical tasks in understanding cognitive offloading is to determine how individuals decide on-the-fly whether to incorporate an external strategy into an ongoing cognitive act’. At least in some cases, the decision (explicit or implicit) to offload a given cognitive task or not is influenced by metacognitive evaluations of our mental abilities.[21] And when these metacognitive evaluations of our abilities are inaccurate, this can lead to suboptimal offloading behavior.[22]

Relatedly, offloading itself – even when it *is* the optimal strategy – inevitably leaves us subject to various new kinds of ‘memory manipulation’ and in ways that raise epistemological problems.[23] (Imagine, for example, that a glitchy iCloud very easily could have distorted offloaded information in your digital diary, but just by luck left it intact.[24])

Putting this all together, the sheer *extent* of our present-day cognitive offloading habits – and the digitally framed epistemological problems posed by such offloading (in current and future forms[25]) – suggests some kind of expansion of the tools we use to investigate the epistemology of memory, and of dispositional knowledge more generally (viz., our knowledge that is non-occurrent[26], and which we retrain through epistemically hygienic

[20] See, e.g., Sachdeva and Gilbert (2020).

[21] See, e.g., Dunn and Risko (2016), Gilbert et al. (2020), Risko and Dunn (2015), and Weis and Weise (2019).

[22] See Risko and Gilbert (2016).

[23] See Risko et al. (2019) and Carter (2020) for discussion of cases.

[24] This kind of case – raised initially in Carter (2013, 2017) will be a focus in Chapter Four.

[25] See Carter (2021, Ch. 1) for a discussion of some future forms of cognitive enhancement that we can expect on the horizon, and which alter how it is possible to represent the world. For some overview of current cognitive enhancement technologies and methods, see Sandberg and Bostrom (2006), Bostrom and Sandberg (2009), and Armstrong et al. (2012).

[26] Though we’ll discuss this idea more in Chapter Two, it should be clarified here that dispositional knowledge is meant to line up with (i) dispositional belief; as opposed to

storage). The longer we wait to expand these tools while at the same time continuing to offload even more extensively, the wider the mismatch between theory and practice becomes.

At this point, our motivations for taking seriously the idea of 'digital knowledge' – its nature and the norms governing how we manage it – are sourced in the sheer influx of offloading, which principally concerns how we nowadays *store* information. A reader who grants what's been said so far about storage (and the related point about mismatch, when it comes to theory) might draw an important line between knowledge *storage*, on the one hand, and knowledge *generation*, on the other. Maybe, as a tempting line of thought goes, we really do need some theory to help us to make sense of 'digital knowledge' stored outside the head as something other than mere information. *But*, even so, it surely remains that all knowledge is at least initially *generated* squarely inside the head; our tech can (when all goes well, by way of offloading) store and preserve knowledge, and we can of course (either digitally or through traditional testimony) share knowledge with others, but our offloading gadgets – viz., our *cognitive scaffolding*[27] – can't *itself* generate it. In *this* respect – the traditionalist might point out – all knowledge is brainbound. All the world's knowledge is *generated by brains*.[28]

Up until about five years or so ago, the above kind of line on knowledge storage vs generation – where just the latter is taken to be necessarily

mere (ii) dispositions to believe. For the canonical discussion of this distinction, see Audi (1994).

[27] For discussion of various forms of cognitive scaffolding within the extended cognition framework, see Kiverstein (2018).

[28] Perhaps no where is this idea – viz., that knowledge is a brainbound *production* of a thinker – more prevalent than in contemporary virtue epistemology, on which knowledge is fundamentally understood as a kind of achievement in thinking – creditable to a (biological) subject, who knows in virtue of exercising her abilities to get to the truth. See, e.g.,Greco (2010), Sosa (2010), and Zagzebski (1996) for some canonical ways of thinking about knowledge in this way. Cf., however, for some more recent trends in virtue epistemology that are more open to liberalising the idea of the knowing subject and the material realisers of her knowledge-generating cognition, see, e.g., Pritchard (2010), Palermos (2014), Kelp (2013), and Carter (2018).

brainbased – might have seemed compelling, even somewhat progressive. However, even this sort of narrative (liberal about knowledge storage, conservative about knowledge generation) is quickly going out of date, thanks to a recent revolution in artificial intelligence that has, especially since the rise of Google DeepMind in 2017, upended what was previously understood about how to most effectively learn from experience.[29]

The key to this revolution has been to exploit new breakthroughs in two areas: reinforcement learning, which is a particular type of machine learning[30], and systems neuroscience, which studies the structure and function of neural circuits and systems.[31] Generic reinforcement learning algorithms – along with quickly reaching pinnacles that have never before achieved in chess and other games such as Go and Shogi[32] – are now *better than humans* at detecting breast cancer[33] and eye disease[34], at folding proteins[35], developing quantum algorithms[36], and even at and producing mathematical proofs[37]. In each of these cases, intelligent machines utilising machine learning have made new intellectual breakthroughs – at the frontiers of all of these subject areas – that, for purely human intelligence, would have been out of reach.

To give a sense of this power – Google DeepMind's Alpha Zero was, in

[29] For some representative work here, see, e.g., Boughton et al. (2020), De Fauw et al. (2018), Evans et al.(2018), McKinney et al. (2020), Powles and Hodson (2017), Sadler and Regan (2019), and Silver et al. (2017).

[30] For a helpful introduction to reinforcement learning in artificial intelligence, see Sutton and Barto (2018).

[31] For a helpful discussion, see David Silver (and colleagues') (2017) discussion of how Google's Alpha Zero.

[32] See Silver et al. (2017); for a sustained discussion specifically of Google Deep Mind's success with chess, see Sadler and Regan (2019).

[33] (McKinney et al. 2020).

[34] (De Fauw et al. 2018).

[35] (Evans et al. 2018).

[36] (Broughton et al. 2020).

[37] (Kaliszyk et al. 2018). Note that, as of April 2021, there has been a new breakthrough in which neural nets have succeeded in quickly solving what are regarded as the most difficult equations in mathematics – viz., partial differential equations – which describe complex phenomena involving many independent variables. See Li et al. (2021).

December 2018 – and after only four hours of playing itself, and with no pre-programmed information about the value of any of the chess pieces – able to crush the world's leading computer chess programme, Stockfish, in a 100-game series by using strategies that chess experts described as 'alien.'[38]

The capabilities of this new form of superhuman AI raises its own distinctive epistemological challenges for the theorist. For one thing, as deep neural networks like Google DeepMind are increasingly taking over the knowledge discovery work at the frontiers of different subject areas, the more we human thinkers are under pressure to not only *rely* on these intelligent machines to keep making further advances (beyond what they've already done), but also *to verify* the accuracy and reliability of these results which we were (via human cognition) unable to achieve in the first place. But what is the epistemically optimal way for us to go about doing this – and what balance of our own cognition versus cognition outsourced to intelligent machines is appropriate for such verification?[39] This is entirely open terrain. As is the related issue of how we might harness certain strategies (e.g., 'strategic forgetting') used by neural networks like Google DeepMind in order to think smarter and learn better ourselves.

Let's take a step back and register where we've now gotten to. We've seen that, just as important information we use to structure our lives is increasingly digitally *stored*, so is it now *also* – across a wide range of domains of inquiry – digitally *generated*. What this means for epistemologists is this: in order to meet new epistemological challenges that these new habits of storing and creating information, we need to find some way to move beyond the dogma that all human knowledge storage and generation is brainbound. The cost of holding on to this dogma seem to be an even wider mismatch between theory and practice.

Let's now look more squarely at the dogma itself, consider why we've

[38] See, e.g., Wright (2017).

[39] The challenge here is one of working out how to balance (i) epistemic dependence with (ii) demands for epistemic gatekeeping (see Greco 2020a, 2020b) that we did not face when human epistemic labour could play relevant gatekeeping functions.

latched on to it in the first place, and whether there remains any good reason to do so.

1.3 On an Intracranialist Dogma in Epistemology

Epistemologists aren't usually very explicit in defending any specific thesis about the nature of cognition which would be incompatible with the ideas motivated in the previous section – viz., that knowledge might at least sometimes, and to some extent, be digitally stored and digitally generated. After all, it's rare that we find epistemologists (as opposed to philosophers of mind[40]) defending specific views about the metaphysical nature of knowledge-apt representational states. However, mainstream work in epistemology – especially on the nature of knowledge – almost invariably[41] *presupposes* a certain background picture of cognition that fits very well with the idea that knowledge is necessarily stored and generated in the head.[42]

In the philosophy of mind and cognitive science, this background picture has a name – *cognitive internalism* – the view that, necessarily, cognition supervenes on brainbound, biological properties of the cogniser.[43] Cognitive internalism is, as Fred Adams and Kenneth Aizawa (2009) put it, tantamount to a dictum of commonsense – viz., that the 'mind is in the head'.

When articulated as a view about cognitive *processes* (i.e., memory storage

[40] For discussion, see, e.g., Cummins (1996), Lycan (2000), Ramsey (2017), and Shea (2013).

[41] For some exceptions, see, e.g., Palermos (2018), Carter (2013), Palermos and Pritchard (2013),

[42] The kind of presupposition here is pragmatic presupposition (in the sense of, e.g., Stalnaker 1973) – in that both sides of the disputes about the nature of knowledge act as though it is in the common ground between them that, e.g., beliefs are in the head, memory processes are in the head, etc.

[43] For some prominent recent defenders of cognitive internalism against challenges from embedded and extended cognition camps, see Adams and Aizawa (2008; 2010).

and retrieval), cognitive internalism is usually read as maintaining that, necessarily, cognitive processes (e.g., memory storage and retrieval) play out entirely inside the head[44]; alternatively such processes are materially realised exclusively by physical processes in the brain. When framed as a view about *states* of cognition (e.g., beliefs), the view implies that your beliefs are literally in your head, in the sense that the physical subvenient bases of your beliefs are all and only intracranial subvenient bases.[45]

Here is perhaps the most straightforward picture of how cognitive internalism is so easily 'smuggled in' – and uncritically so – as a presupposition in epistemology. Propositional *knowledge* – of central interest in epistemology – is assumed *ex ante* to *entail* belief, truth, and justification, as per the traditional JTB analysis.[46] The project of analysing knowledge – which dominated the second half of 20th century epistemology – aimed to work out how these three (and perhaps other) conditions relate to each other when one has knowledge. *Belief* relates to the other two conditions – at least, in a way that matters for analysing knowledge – in so far as beliefs are propositional attitudes with a representational (i.e., mind to world) direction of fit. That is, after all, what really matters in the analysis of knowledge, because it is exactly *this kind of a thing* that is capable of being true and justified and thus, as the thought goes, capable of being known. And furthermore, in at least *paradigmatic* cases of propositional knowledge (think of simple perceptual knowledge – viz., your knowledge that there is a hand in front of you, which you generate and then retain in memory), it seems plain enough that brainpower is going to be both both necessary and sufficient to (i) generate the (occurrent) propositional attitude with a representational direction of fit (i.e., *that* there is a hand in front of you right now) and then to (ii) store it, as a dispositional belief,

[44] See, e.g., Carter et al. (2014), Wheeler (2018), Carter et al. (2016), Kiverstein (2018), Palermos (2018), and Palermos and Pritchard (2013), and Pritchard (2010).

[45] The thesis applies not only to beliefs, but also to, e.g., desires and emotions. Thus, a view on which emotions supervene partly on something extracranial, including on partially extracranially driven appraisal processes (see, e.g., Carter et al. 2016; Kruger and Szanto 2016) is incompatible with cognitive internalism.

[46] For some detailed overviews of this project, see Shope (2017) and Ichikawa and Steup (2018). For criticism, see Williamson (2002, Ch. 1)

in memory.

Thus, the 'map' to unearthing the cognitive internalist presupposition in mainstream thinking about knowledge is accordingly a pretty direct one, with two key 'links' in the chain: the first that gets us from knowledge to belief (via the 'entailment thesis' that knowledge entails belief) and the second that gets us from belief to cognitive internalism (where the latter is the assumed picture about how the former is realised and maintained).

Let's focus now on the *first link* in this chain that burrows us down to the cognitive internalist dogma (we'll return to the second link in the next section). Does knowledge *really* entail belief? If so, what is the best way to interpret this claim?

1.3.1 Knowledge and belief

While the idea that knowledge entails belief is widely assumed[47], it is rarely argued for positively (apart from being defended against objections[48]), with two notable exceptions being G. E. Moore (1962) and Keith Lehrer (1968). Moore famously tried to show that knowledge entails belief via a (albeit somewhat odd) linguistic test, and Lehrer (1968) opted for a proof aimed at showing that knowledge formally entails belief. Neither is promising.

According to Moore:

> There certainly is a common use of belief in which 'I believe' entails 'I don't know for certain'. Is there another in which 'I know for certain' entails 'I believe'? One reason why it seems so is because 'I thought I knew' entails 'I believed' (1962, 115).

[47] In particular, we find this assumption in the decades of critical response to Gettier (1963). See, e.g., Shope (2017).

[48] Defences of the knowledge-belief entailment against objections have largely focused on responses in the late 1960s and 1970s to Radford's (1966) 'unconfident examinee' case, which we discuss later in this section.

It does seem plausible that speaker who says 'I thought I knew' that *p* is committed in some way to accepting that they believed that *p*. But let's simply *grant* for the sake of argument that patterns like the one Moore mentions constitute linguistic evidence that knowledge entails belief (either occurrent or dispositional). Even on this charitable assumption, there is, as Carolyn Black (1971) has observed, equally compelling linguistic data that would seem to support the very *opposite* conclusion. Take for example, this case: 'I say that my books are in my office. You ask 'Do you believe that your books are in your office?' I say '*No*. I *know* that my books are in my office' (Black 1971, 155–6, my italics). The felicitousness of *this* kind of exchange is a problem for arguments that attempt to establish that knowledge entails belief (of any sort) simply on the basis of our patterns of using the words 'knows' and 'believes'.

So what about Lehrer's (1968) proof? Here is the proof, which he takes to be sufficient to establish to a doubter that knowledge entails belief.

1. If *S* does not believe that *P*, then *S* does not believe that he knows that *P*;
2. If *S* does not believe that he knows that *P*, then, even though *S* correctly says that *P* and knows that he has said that *P*, S does not know that he correctly says that *P*.;
3. If, even though *S* correctly says that *P* and knows that he has said that *P*, S does not know that he correctly says that *P*, then *S* does not know that *P*;
4. (Therefore) If *S* does not believe that *P*, then *S* does not know that *P*. (1968, 498)

There are problems with both premises (1) and (3). The problem with (1) is that it is either false or *at best* questionbegging, given what Lehrer was attempting to do here. Just consider that the kind of opponent Lehrer is out to convince might very well think that "S knows that *p*" is compatible with the antecedent of (1). But then, (1) comes out false if *S* doesn't believe that *p because S knows that p*, given that, on that supposition, it's possible that *S* will believe *that S* knows that *p*. But even if this problem with (1) could be dealt with, there are independent problems with

(3): just suppose your friend tells you they don't know all the lines of a certain poem by William Blake, but then (after telling you this) they proceed to recite the poem perfectly; this seems like a plausible case where – even though they don't know *that* they have correctly recited it – they nonetheless know the lines.[49] They had them mastered better than they had thought. This assessment of the case, however, is incompatible with (3).

Interestingly, we don't find many other attempts[50] to positively establish the widely held assumption that knowledge entails belief (and, as it turns out, Lehrer himself abandoned his own proof later[51], opting instead for a view on which knowledge entails not belief but *acceptance*.)

Instead, what much of the literature on the knowledge-belief 'entailment thesis' concerns is whether outlying attempts to *challenge* the assumption are sound. The most widely discussed case on this score – also one that involves a kind of 'knowledge-with-lack-of-confidence' structure – is due to Colin Radford (1966):

> UNCONFIDENT EXAMINEE: Kate is taking a history test. She had studied carefully and has been doing well on all the questions so far. She has now reached the final question, which reads "What year did Queen Elizabeth die?" As Kate reads this question she feels relief, since she had expected this question and memorized the answer. But before Kate can pause to recall the date, the teacher interrupts and announces that there is only one minute left. Now Kate panics. Her grip tightens around her pen. Her mind goes blank, and nothing comes to her. She feels that she can only guess. So, feeling shaken and dejected, she writes "1603"—which is of course exactly the right answer.

[49] For a similar case, see Black (1971, 157).

[50] While Armstrong's (1969) paper 'Does Knowledge Entail Belief' is ostensibly a defence of the claim, it is less an attempt to establish the thesis than it is to defend it against cases like those from Radford (1966).

[51] See Lehrer ([1990] 2018).

As David Rose and Jonathan Schaffer (2013) put it, 'The case of *Unconfident examinee* represents the leading challenge to the orthodox idea that knowledge entails belief' (2013, S20). Apart from this classic case from the mid 1960s – and the extensive critical response to it (on both sides), which fizzled out in the 1980s – the most notable recent lines of argument against the idea that knowledge entails belief, all in the past 10 years, are due to Blake Myers-Schulz and Eric Schwitzgebel (2013), Katalin Farkas (2015), and Susanna Schellenberg (2017a). Myers-Schulz and Schwitzgebel present experimental evidence[52] that the intuition in UNCONFIDENT EXAMINEEE that Kate has knowledge without belief (that Queen Elizabeth died in 1603) is robust, and on this basis, purport to give a 'second wind' to the old counterexample to the orthodox presumption that knowledge entails belief.[53] Farkas, on the other hand, use cases of *extended cognition* (e.g., cases where one offloads one's memory tasks to a notebook or an iPhone) as plausible cases where one has knowledge without belief, albeit, knowledge stored externally. Finally, Schellenberg's tack is to cast doubt on whether the entailment thesis holds specifically in cases of perceptual knowledge, where (arguably) one knows simply via *seeing* that something is so, and regardless of whether one forms a belief.

Just as we've seen that Moore and Lehrer didn't plausibly *demonstrate* that knowledge entails belief (either occurred or dispositional), there is also a good case to be made that *none* of the above attempts aimed at establishing that knowledge *doesn't* entail belief succeeds, at least in so far as none of these strategies plausibly demonstrates that knowledge does *not* entail dispositional belief. This point turns out to be relevant to the wider transition from 'knowledge to belief, and then from belief to cogntive internalism', given that occurrent belief rather than dispositional belief is more *prima facie* plausibly wed to a cognitive internalist picture of the mind.

[52] For some additional experimental evidence that is meant to vindicate the idea that UNCONFIDENT EXAMINEE counts against the entailment thesis, see Murray et al. (2013).

[53] As they note: "A majority of respondents ascribed knowledge [...]]while only a minority ascribed belief" (Myers-Schulz and Schwitzgebel 2013).

Regarding the UNCONFIDENT EXAMINEE case: The pressure against the entailment thesis is really the strongest when we contrast (i) the observation that Kate's lack of confidence in the proposition that the Queen died in 1603 doesn't seem to preclude her from knowing it, and indeed, manifesting that knowledge unconfidently, with (ii) the thought that Kate must believe and thus *consciously endorse* the content that the Queen died in 1603 at some time if she is to know it at that time. The force of UNCONFIDENT EXAMINEE against the entailment thesis lies in the fact that it leads us to embrace (i), and then on that basis reject (ii).

But importantly, a rejection of (ii) is *compatible* with the thesis that knowledge entails belief, so long as 'belief' is understood in a dispositional sense, where dispositional beliefs are merely *available to mind for endorsement*[54] even when the content of a dispositional belief is not consciously endorsed.[55] Rose and Schaffer (2013) support this rationale on the basis of two considerations. First, Kate's memory trace[56] (viz., that Queen Elizabeth died in 1603) is not destroyed. Second, her guess is no accident.[57] On the second point, they write:

> Indeed it seems as if her memory trace must still be not just present but actually operating in the background to guide her actions, even if she is unable in the moment to appreciate the fact. Putting these two reasons together—to

[54] For some explicit discussions of dispositional belief, its relationship with occurrent belief, see Armstrong (1973), Lycan (1986), and Audi (1994). For an overview, see Schwitzgebel (2019, sec. 2.1).

[55] Though, perhaps – as Murray et al. (2013) – maintain, one must have assented to the proposition at some point in the past. We'll take this issue up when discussing dispositional belief and the extended cognition thesis in Chapter Two.

[56] Memory traces (sometimes referred to as 'engrams' in psychology) are taken to be the means by which we store memories in the brain. For a recent overview of work on memory traces, see de Brigand (2014). For philosophical discussion of memory traces in the epistemology of memory, see Bernecker (2010).

[57] The idea that one might know via reliable guessing, even when one lacks confidence, is given an explicit defence in Ernest Sosa's (2015) virtue epistemology. In particular, see Sosa's case of the eye examination (in his 2015, Ch. 3).

> the extent that it is useful to operate with the picture of a "belief box" in which various propositions are stored—we find it natural to think of Kate as having the proposition that Queen Elizabeth died in 1603 lodged in her belief box throughout. She stored it there during her studies and is still unconsciously guided by it when she "guesses." Indeed we find it natural to imagine that—perhaps later that very day—Kate will recover from her panic and recall the information readily enough. She has the information stored in mind. She is merely temporarily blocked from accessing it normally (2013, S24–5).

It looks, then, as though we should deny that Kate has a dispositional belief only if we are prepared to say that her temporary block is permanent rather than temporary. But *even if it were* permanent, note that the kind of block she has just prevents her from accessing the information stored in mind *normally*. It doesn't prevent her from accessing it *at all* for the reason that this information stored in memory continues to guide her actions. Of course, *were it* to somehow be blocked off from even doing *that*, then we might then deny her the dispositional belief, on account that it is unaccessably stored in memory. However, on that kind of a scenario, there would then be no pressure to attribute to her knowledge. If she guessed correctly, it would be by sheer luck.

The above considerations cast Myers-Schulz and Schwitzgebel's (2013) experimental results in a different light. From the fact that folk are more likely to attribute knowledge than belief in UNCONFIDENT EXAMINEE we have no good reason to reject the entailment thesis – at least not without a clearer sense of which sense of the polysemous 'belief' the participants took themselves to be withholding while at the same time attributing knowledge. Interestingly, as more recent experimental studies indicate.[58], when the same experiments are run while eliciting the dispo-

[58] These are the results reported by Rose and Schaffer (2013), who replicated the Myers-Schulz and Schwitzgebel experiments while more explicitly eliciting the dispositional reading of 'belief'.

sitional reading of belief more so than it was elicited in the original experiments, people's intuitions no longer disproportionately attribute knowledge rather than belief.

The take-away lesson from the UNCONFIDENT EXAMINEE case seems to be this: the case (i) purports to show that it's not the case that knowledge entails belief; (ii) it plausibly *does* demonstrate that knowledge doesn't entail occurrent belief; but (iii) it *doesn't* succeed in showing that knowledge *doesn't* entail dispositional belief – on the contrary, we would plausibly be less likely to attribute knowledge in the case were dispositional belief *not* present.

Although Farkas's argument against the knowledge-belief entailment thesis is ostensibly very different from the line of argument that proceeds from the THE UNCONFIDENT EXAMINEE case, an appreciation of Farkas's wider argument shows that it ultimately slots into the very same kind of (i, ii, iii) structure.

Her argument takes as its basis a case of cognitive offloading from memory to notebook. The case – involving the characters 'Otto' and 'Inga' – was originally used by Andy Clark and David Chalmers (1998) as an argument *against* cognitive internalism, and in favour of the idea that cognition can extend beyond the boundaries of the skull and skin. The Otto and Inga case – and the thesis of 'extended cognition' more generally – will be discussed in detail in Chapter 2 (and introduced in more detail in §1.32). For our purposes now, though, let's just focus squarely on how Farkas thinks the case supports a rejection of the orthodox idea that knowledge entails belief.

The key first step for Farkas is to take a queue from Edward Craig's (1991) thinking about the *purpose* of the concept of knowledge, an understanding of which Craig thinks would help to illuminate what falls in its extension.[59] According to Craig: "[k]nowledge is not a given phenomenon,

[59] This 'Craigian' idea that the nature of knowledge is something we can fruitfully illuminate by first inquiring into what the concept of knowledge is for – viz., what the fucntion of the concept of knowledge is – has enioyed some more recent support under the heading of 'function-first' epistemology. See, e.g., McKenna (2013) and Hannon

but something that we delineate by operating with a concept which we create in answer to certain needs, or in pursuit of certain ideals' (1991, 2) On Craig's view, we 'create' the concept of knowledge in order to meet the need we have to *flag reliable informants*. And so, the on the Craigian view, the *function* of the concept of knowledge is to flag reliable informants, and relatedly, an appreciation of this function as the function it is should guide our thinking about what falls within the extension of the concept of 'knowledge'.

Now, with these Craigian ideas assumed, Farkas encourages us to think about the case of Otto and Inga:

> OTTO AND INGA: Inga would like to go to the Museum of Modern Art (MoMA); she recalls that the MoMA is on 53rd street, and she sets off accordingly. Otto suffers from severe memory loss and therefore he keeps all important information recorded in a notebook which he carries with him all the time. When he decides to go to MoMA, he looks up the whereabouts of the museum, finds it's on 53rd street, and then he sets off. Many people agree that Inga had had the belief that the Museum of Modern Art was on 53rd street even before the issue came up in connection to her current visit. But Clark and Chalmers claim that if Inga has the belief, so does Otto, even before he looked up the information in his notebook. Otto has reliable, constant and easy access to the contents of his notebook, and he endorses the contents of his notebook automatically. This, according to Clark and Chalmers, is enough to qualify him as having the belief (2015, 190).

Farkas's own idiosyncratic take on this case fits neither with traditional thinking (on which Otto's neither believes *nor* knows that MoMA is on 53nd street in virtue of storing this information as he does in his notebook but not in his head); but *nor* does Farkas's assessment line up with the point Clark and Chalmers originally used the case to make, which is that –

(2018). Cf., Gerken (2015).

as they see it – Otto's memory (and thus, his dispositional beliefs stored in memory) lies partly in the notebook, external to his brain. Farkas thinks – and we needn't get in to the details just yet, but we'll return to them – that we should agree with the traditionalist that the cognitive differences between Otto and Inga are substantial enough that, when it comes to attributing 'belief', we should do so dianalogously to Inga but *not* to Otto. On the other hand, however, she thinks we should part ways with the traditionalist – and simply be guided by Craig – when it comes to whether to attribute *knowledge* to Otto. Recall again the Craigian idea that the point of the concept of knowledge is to track reliable informants, and just consider in this light how we use 'knowledge' to track such informants in cases of, e.g., seeking phone numbers. As Farkas writes, in 'some everyday contexts, it is very natural to attribute knowledge to subjects who are in Otto-type situations. You ask me if I know NN's phone number, and I say "sure", reaching for my smartphone' (Farkas 2015, 190).

Putting this all together, Farkas thinks we have compelling reason to think Otto *knows but doesn't believe* that MoMA is on 53rd street, and a *fortiori*, that the knowledge-belief entailment thesis is false. Now, I've suggested at the outset that I think Farkas's argument ends up slotting into the (i,ii,iii) structure that characterised the purported argument against the entailment thesis from UNCONFIDENT EXAMINEE. I now want to explain why.

First, consider that one tempting spot to challenge Farkas's reasoning is her claim that Otto and Inga are different enough that we should *not* attribute dispositional belief across the cases symmetrically. Why not? Why *aren't* Clark and Chalmers right about this, as opposed to the traditionalist? Fortunately, there is a way press back against Farkas without fully opening that can of worms (we'll circle back to it), which is to suggest that *by her own lights* we ought to attribute Otto a dispositional belief. The reasoning here is that attributing such a dispositional belief is the most promising way for Farkas to vindicate her claim that Otto has (extended) *knowledge*. Farkas's rationale for attributing Otto extended knowledge, after all, is meant to be guided by the Craigian idea that we should use 'knowledge' to track reliable informants. The presumption here (which we should grant) is that Otto is such a reliable informant; ask him where MoMA is, he can reliably tell you (via a process that involves consulting his notebook rather than biomemory). Now, *what is it that grounds Otto's reliability about where MoMA is*? It's hardly a brute fact that he's reliable – on the contrary, he's a reliable informant because he *reliably stores the information* (just like Inga does); his information is correct, easily available for endorsement, etc. Indeed, it thus looks quite a bit like the thesis that Otto knows where MoMA is (in virtue of what's in his notebook) would be explained (even granting the Craigian story) by his having something that looks an awful lot like a dispositional belief.

Now, a traditionalist has at this juncture might try to dig their heels in as a matter of principle: 'Cognitive internalism is true and so, necessarily, all cognition plays out in the head; therefore, Otto simply *can't* have a dispositional belief externally stored." But – crucially – it looks like this kind of a principled reason is already out the window for Farkas, who explicitly allows *knowledge* outside the head. Farkas's line that Otto's case features knowledge without belief accordingly occupies a curious area of dialectical space: her claim that Otto has knowledge (that MoMA is on 53rd street) itself seems best *explained* by his having a dispositional belief, in virtue of how he stores the information he does, not in biomemory, but in the notebook. Farkas of course, denies that he has a dispositional belief (by appealing to cognitive internalist thinking); *but* that denial would itself be principled denial only if Farkas were to *also* deny that he has extended knowledge (which she of course does not deny).

Putting this all together, then, it looks as though – as with UNCONFIDENT EXAMINEEE – the case of Otto and Inga (at least, as Farkas is using it) exhibits (i,ii,iii) structure; it is a case that Farkas (i) purports to use to show that it's not the case that knowledge entails belief; (ii) the case plausibly *does* demonstrate that knowledge doesn't entail occurrent belief (given that Otto clearly lacks such a belief, no less than the unconfident examinee does); but (iii) it *doesn't* succeed in showing that knowledge doesn't entail dispositional belief, and if anything, only serves to positively reinforce this idea.

Let's round out our discussion of the knowledge-belief entailment thesis with a brief look at Susanna Schellenberg's domain-specific dismissal of the idea that knowledge entails belief. The line she advances is 'domain specific' because it is meant to apply exclusively to perception, and thus to perceptual knowledge. According to Schellenberg's view of perceptual knowledge, *capacitivism*, a subject (S) has perceptual knowledge that p by *seeing* that p, which requires that S employ 'a capacity to single out what she purports to single out' (2017b, 318) and S's mental state (whereby S sees that p) must have 'the content it has in virtue of S having successfully employed her capacity to single out what she purports to single out' (2017b, 318).

As the reader will have noticed, 'belief' does not feature in the above story. This, Schellenberg thinks, is just as it should be. She writes:

> Orthodoxy has it that one cannot know that p without believing that p. Capacitivism is neutral on whether there is any such belief condition on knowledge. This is attractive, since arguably, we know that p simply in virtue of seeing that p. By contrast, we do not believe that p simply in virtue of seeing that p. After all, I can see that p without forming any beliefs (2017b, 318).

Of course, even if Schellenberg is right, she will have been right about a story of perceptual knowledge acquisition. What about perceptual knowledge *retention?* Suppose you see that p at t_1. At t_2, you are no longer thinking about p. But, if someone asks you at t_2 what p looked like, you remember and can tell them. But this would turn out to be mysterious if at t_2 you didn't retain this information about p in a way that was then later available to mind for endorsement.[60] But that's just the mark of a dispositional belief. To the extent that Schellenberg's capacitivism is a correct story of perceptual knowledge acquisition, this story looks to be compatible with the version of the entailment thesis that has seemed most plausible so far – viz., that knowledge requires at

[60] I'm using the simplified idea of 'available to mind for endorsement' from Rose and Schaffer (2013, secs. 1.3, S22).

least dispositional belief.

Recall now that the 'map' to unearthing the cognitive internalist dogma in mainstream thinking about knowledge had two key 'links' in the chain, one from knowledge to belief (Link 1), the other from belief to cognitive internalism:

This section – critically examining Link 1 – reveals that the most charitable way to unpack Link 1 is as:

- **Link $1_{\textbf{dispositional}}$:** propositional knowledge → (entails) dispositional belief

rather than

- (!) **Link $1_{\textbf{occurrent}}$:** propositional knowledge → (entails) occurrent belief

However, from Link $1_{\text{dispositional}}$, we most plausibly get to cognitive internalism only by way of

- (!) **Link $2_{\textbf{dispositional}}$:** dispositional belief → (is best explained by) a cognitive internalist picture of the mind

rather than:

- **Link $2_{\textbf{occurrent}}$:** occurrent belief → (is best explained by) a cognitive internalist picture of the mind

But this is where the overarching story – from mainstream thinking about knowledge to the cognitive internalist assumption that tacitly underlies it – begins to show some real cracks. Just consider that, whereas Link $2_{\text{occurrent}}$ is *prima facie* very plausible (if not obvious to many), Link $2_{\text{dispositional}}$ really isn't.

The reason Link $2_{\text{occurrent}}$ seems platitudinous is that occurrent belief is usually taken to involve *consciously* entertaining (and subsequently endorsing) a proposition; and a biological brain is plausibly (though this point is of course debatable) necessary *and* sufficient for this kind of conscious activity; accordingly, it is *prima facie* plausible that cognition *of the sort* *that is realised exclusively as the cognitive internalist countenances* is what furnishes us with whatever occurrent beliefs we have. Crucially, however, a biological brain is – though obviously sufficient – *not necessary* for realising the kind of thing that hosting a dispositional belief is generally taken to involve, which is the *storing* of information that is *available* to us for conscious endorsement. If anything, the ubiquity of cognitive offloading (§1.2) suggests that even though biological brains suffice for storing information available for conscious endorsement, they're obviously not necessary because we use them *for this very purpose* increasingly less – especially when it comes to practical information of the sort we rely on to structure our lives. It is, then, at best *prima facie* plausible that cognition of the sort that is realised exclusively as the cognitive internalist countenances furnishes us with only some of our dispositional beliefs. But this means, then, that the phenomenon of dispositional beliefs is best explained by a picture of the mind that allows for storage of information available for conscious endorsement to sometimes be handled intracranially, sometimes (and increasingly often) not.

At this juncture, the proponent of cognitive internalism might simply double down as follows: "even if the sense in which knowledge entails belief is best understood as Link $1_{\text{dispositional}}$ *rather* than Link $1_{\text{occurrent}}$, and indeed even if it *looks* as though we can make sense of many of the dispositional beliefs we have without assuming anything like cognitive internalism, it remains that cognitive internalism stands up as an independently and overwhelmingly plausible 'pillar' in the philosophy of mind; it establishes the bounds of cognition in a way that aligns with centuries of philosophical thinking, and we are better placed simply accepting the implications of cognitive internalism wherever they lead us, *even where* they don't align so well with our other commitments (at least, when these other commitments lack the kind of 'bedrock' status that cognitive internalism enjoys). And so, despite initial appearances to the contrary, we should not accept but resist the temptation to think that the process of storing information available for endorsement in notebook or iPhone (rather than in biomemory) is a genuine *cognitive* process, and thus, we should resist attributing 'beliefs' and 'knowledge" on the basis of such storage.

Does the proponent of cognitive internalism here have a point? This really depends on whether cognitive internalism is (or deserves to be) the kind of 'pillar' in our theorising that the above reasoning suggests. As it turns out, pillars fall, and lately, old 'internalist pillars' in particular have been falling right and left.

1.3.2 The Falling Pillars of Cartesianism

Until relatively recently, the study of knowledge was – following a tradition inherited from Descartes[61] – a thoroughly 'internalist' enterprise in three key ways.

First, it used to be taken for granted that the content of our thoughts is determined entirely by the inner workings of the mind – viz., *content internalism*.[62] On this way of thinking, your intentional attitudes (e.g., your beliefs and other attitudes that are *about* things) are about the things that they are about (rather than about other things) in virtue of your psychological states and nothing else. Any two people in the same psychological states, then, must be thinking *about* the very same thing. For those (like Descartes) who are aligned with this kind of thinking, it's easy to see how 'rigorous philosophical inquiry must proceed via an inside-to-out strategy'; and of course, as was apparent in the *Meditations*, from *this* kind of methodological starting point, the challenge of (non-circularly[63]) defeating the sceptic becomes especially difficult.[64]

[61] The typical reference point here is the *Meditations*, however, Descartes' internalist picture of the mind and the way it represents the world is not limited to his epistemology; it is also central to his wider philosophy of mind. See, e.g., Cottingham (2002).

[62] For some representative discussions of cognitive internalism, see Loar et al. (1988), Kriegel (2013), and Fodor (1987).

[63] As Descartes suggested, even from a content internalist starting point, one can 'transition' from knowledge of one's mind to knowledge of the world if one is entitled to the claim that there is a non-deceiving God. However, a famous objection to Descartes is that it is not clear how one can get to this conclusion non-circularly. For discussion, see, e.g., Markie (1992).

[64] Arguably, as some epistemic externalists (e.g., Sosa 1997) have pointed out, analogous problems arise for indirect realist strategies in the epistemology of perception (e.g., Moore's) which purport to vindicate external world perceptual knowledge as based on

Even so, content internalism is not itself an epistemological thesis (even if it has some epistemological ramifications); it's a thesis about how the content of our thoughts and words are individuated. An importantly different kind of internalism – also inherited from Descartes and widely assumed until around the 1970s[65] – is *epistemic internalism*.[66] Epistemic internalism is not a thesis about what our thoughts and words refer to, but about what kinds of things *justify* our beliefs in a way that matters for knowledge. It is in principle compatible with either content internalism or content externalism.[67] What the epistemic internalist maintains is that epistemic justification is *solely* determined by factors that are internal[68] to a person.[69] Such factors include, e.g., what mental states one is in, what is accessible to one via reflection alone, etc. A simple reason why this kind of view (a centrepiece of Descartes' epistemology, but with origins as early as the *Theatetus*) has plausibly enjoyed the support it has is that we tend to think of the kind of justification that matters for knowledge as being associated with reasons and evidence, and the matter of what reasons and evidence one has seems – on the face of things – to be determined by fac-

inference from information just about the qualitative character of our experiences.

[65] It's important not to run together the longstanding endorsement of an internalist picture of epistemic justification with the related, but separate, issue of whether this picture of epistemic justification has a longstanding place in a justified-true-belief analysis of propositional knowledge. As Dutant (2015) has called received thinking about the place of the JTB analysis in the history of epistemology since Descartes into doubt, this doubt doesn't apply to the largely internalist way in which epistemologists have (until the rise of externalism in the 1960s and 70s) thought and talked about knowledge-relevant justification.

[66] For a sample of epistemic internalist positions in epistemology, see Alston (1988), Chisholm (1973), Conee and Feldman (2004), and Huemer (2006).

[67] Though, for some critical discussion on this point, see Chase (2001), Pritchard and Kallestrup (2004), and Carter et al. (2014).

[68] The 'internal' in internalism is usually taken to be something like 'internal to one's psychology' or to one's 'mental states'. And *those* are things that are almost invariably understood as brainbound. That said, it is at least in principle possible to envision a more raeical kind of epistemic internalism, paired with a more inclusive conception of what one's psychology and mental life can consist of. For a discussion of this more radical kind of spin on epistemic internalism, see Carter and Palermos (2015).

[69] See, e.g., Poston (2014) and Madison (2017).

tors internal to one (e.g., what your mental states are).[70]

Rounding out the three internalist 'pillars' of Cartesian epistemology is our old friend *cognitive* internalism on which what is claimed to be 'internal' to a thinker is not the content of their thoughts (content internalism) or what matters for justifying their beliefs (epistemic internalism), but rather the *material realisers* of her cognising, including whatever thoughts and beliefs she has, justified or not.

The suggestion – canvassed in the previous section on behalf of the traditionalist – that cognitive internalism is some kind of unalterable 'pillar' that mustn't be dislodged is really not very compelling in the context of appreciating that – of these three internalist 'pillars of Cartesianism' – the *first two have already fallen*, and both within just the past 50 years.

Content externalists in the 1970s[71] and 1980s[72] have shown how our environments play a crucial role in individuating meaning and mental content, and to such an extent that content internalists are nearly extinct in 2021. As Juhani Yli-Vakkuri and John Hawthorne (2018) put it – in a recent monograph purporting to be the final nail in the coffin of this kind of internalist thinking – 'entanglement of our minds with the external world runs so deep that no internal component of mentality can easily be cordoned off'. With the exception of Hawthorne and Yli-Vakkuri's purported final takedown, content externalism is now so popular it is rarely taken to need any additional argument. Essentially, philosophical thinking has 'flipped' almost completely since the mid 1970s, and on a position fundamental to our grip on the very nature of thought.

What about *epistemic internalism*, then? It has slowly but steadily (since the 1960s) been heading the way of content internalism. According to results from a PhilPapers Survey published by David Chalmers and David Bourget in (2014), only about a quarter of 931 philosophers surveyed (246 / 931 (26.4%)) self-identify as epistemic internalists. This is so even

[70] A good example of this kind of assimilation of 'reasons' and 'evidence' talk with epistemic internalism is found in Chisholm (1977).

[71] (Putnam 1975).

[72] (Burge 1986).

though internalism captured the default position in epistemological theory from Plato, to both rationalists (Decartes) and empiricists (Locke and Hume)[73] all the way up to Gettier (1963). While debates between epistemic internalists and externalists remain contentious, one thing that is clear is that epistemic internalism is no longer the default view but rather the exception.[74]

So is *cognitive internalism* the only 'Cartesian pillar' that should be thought of as 'safe' from the externalist wave – and as such, permanently fixed? The short answer is 'no' for the reason that this final internalist pillar has at least partially (arguably: *mostly*) fallen *already*, as the past 20 years of the philosophy of cognitive science suggests. It's just that – put simply – this news hasn't quite spread to mainstream epistemology.

The cracks in cognitive internalism started quite small.[75] Forget iPhones and the like for a moment, and just think about your hands, and how you move them around, gesturing as you talk. This kind of gesturing, not only facilitates communication, but it also helps language *processing* (McNeill 1992). Likewise, consider the baseball outfielder (McBeath, Shaffer, and Kaiser 1995) trying to catch a fly ball, by running in a direction that makes the ball appear to follow a straight line. In doing this, the outfielder is solving a complex problem not just by perception, but by a kind of 'perception-action coupling' – viz., by using perceptual information to

[73] For clear presentation of Hume's internalist foundationalism, see, e.g., Sosa (1980). For Locke (e.g., *Essay IV*, xvii, 24) epistemic internalism was a feature of his wider assimilation of epistemic justification with doing one's epistemic duty.

[74] Key to epistemic internalism's downfall is arguably the sheer strength of the thesis itself – viz., its contention that *everything* that matters for justification must be internal to one's psychology, or available to one by reflection alone. As externalist epistemologists such as Alvin Goldman (1979) have emphasised, such views can't countenance the insight that the reliability of a belief forming process is among those things that seem to matter. But, the externalist is not similarly restricted; the externalist can consistently allow that some of what matters for epistemic justification is determined by factors internal to one's psychology, though not everything.

[75] Outside of analytic philosophy, the idea that cognition might be embodied was already gaining some traction in 19th century continental phenomenology of perception. For an overview of the history of embodied cognition, see Gallagher (2014).

guide movement and then using movement to hold the perceptual information constant.

The above are just some representative example cases – others (many of which have appeared just since the 1990s[76]) involve visual consciousness[77], concepts[78], memory[79], moral cognition[80], etc. – which have been taken to favour the view that cognition is best understood as not only taking place in the brain, but more widely, as *embodied*. Wilson and Foglia (2017) articulate the core of the 'embodied cognition' thesis as follows:

> Many features of cognition are embodied in that they are deeply dependent upon characteristics of the physical body of an agent, such that the agent's beyond-the-brain body plays a significant causal role, or a physically constitutive role, in that agent's cognitive processing.

What the evidence for embodied cognition suggests is that cognition is not merely (as traditionalists would have it) 'sandwiched between, while segregated from' perception and action. The dependence of the former on the latter simply too deep to separate in the clean way the traditionalist/internalist would want.[81]

Getting down to brass tax: if *any* part of an agent's *non-brain* body has *ever* played a physically constitutive role in cognitive processing, then strictly speaking, cognitive internalism is false.[82] And as more evidence has come in that validates this very idea, embodied cognition has increasingly taken

[76] For an overview, see Gibbs (2005a) and Wilson and Foglia (2017, sec. 5).

[77] See, e.g., Noë (2005) and Hurley (1998).

[78] (Lakoff 2012).

[79] (Sutton 2006).

[80] (Haidt, Koller, and Dias 1993).

[81] This 'insepararbility' idea is sharpened further by those friends of embodied cognition who go a step further to think of the body and mind as a kind of dynamical system (e.g, Chemero 2011; Beer 1995; Palermos 2016, 2014).

[82] See Gibbs (2005b). Some cognitive scientists have called into question whether the empirical evidence supports what proponents of embodied cognition take it to support. For a response to some of these rejoinders, see Miracchi (2021).

over as the ‘default’ position in cognitive science. As Fred Adams (2010) – a dyed-in-the-wool traditionalist – concedes: “The view that cognition is embodied [...] is rapidly gaining prominence in the world of cognitive science, *and is aiming for dominance* (2010, 619). According to Lawrence Shapiro (2014), embodied cognition is”now one of the foremost areas of study and research in philosophy of mind, philosophy of psychology and cognitive science.”

It is hard to see how cognitive internalism should deserve any kind of sacrosanct status when the tide in cognitive science is now generally against it. But if *that’s* right, then isn’t it just a clear mistake for epistemologists to cling tacitly to cognitive internalism?

Maybe – in a sense – not. Consider this line of argument: “Let’s assume cognition is embodied – granted! Even so, this is a far cry from suggesting that you can have *beliefs* in your *phone*. Your phone is *not* part of your brain *or* your biological body!” This kind of rejoinder suggests that perhaps the best candidate for a plausibly sacrosanct thesis in the neighbourhood of cognitive internalism isn’t strict cognitive internalism after all, but rather the more permissive cognitive *bio-internalism* – the thesis that cognition is essentially *biologically* realised.

However, even if we shift the goal posts of sanctity from cognitive internalism to *cognitive biointernalism*, we still fail to capture anything properly sacrosanct. Two straightforward challenges on this score come from cognitive neuroscience over the past 5 years alone: (i) the 2015 creation of the first artificial neuron (Simon et al. 2015), and (ii) the first successful case (in 2019) of creating artificial memories from scratch and implanting them in mice, where the artificial memories guided behaviour indistinguishably from non-implanted memories (Vetere et al. 2019). Note that in neither of these cases is cognition realised entirely as the biointernalist would have it.

A larger elephant in the room, however, comes from the philosophy of cognitive science, where researchers are increasingly open and explicit in their denial of even *cognitive biointernalism*. It is here where it will be useful to circle back to the case of Otto and Inga, originally due to Clark and

Chalmers (1998). Both the cognitive internalist *and* the more permissive cognitive biointernalist are going to diagnose Inga and Otto asymmetrically when it comes to whether they count (respectively) as *remembering* – prior to accessing this information from storage – that the Museum of Modern Art is on 53rd street. In Inga's case, we attribute to her a paradigmatic dispositional belief, in virtue of storing this (previously endorsed) information in biomemory. In Otto's case, we – according to the cognitive internalist *and* biointernalist – deny this dispositional belief attribution, simply because the information is not stored in biomemory; it's stored somewhere else.

But how important should *this* be, really? Proponents of extended cognition[83] think that giving this kind of theoretical weight to the material constitution and location of our memory storage is outdated and unprincipled – and as David Chalmers[84] puts it a form of – *bioprejudice*. A more egalitarian approach to the bounds of cognition would have us focus – when deciding whether to include something as part of a cognitive process – less on what it's made of and where it is, and instead on what it does. If something is *doing* the same thing as something that's part of a cognitive process, then why not - in the spirit of parity of treatment - rule it in?

This is the central (then-)radical idea from Clark and Chalmers' (1998) "The Extended Mind", which is summed up (1998, 8) in their 'parity principle':

> **Parity Principle**: [I]f, as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is part of the cognitive process (1998, 8).

For those willing to reason in accordance with this principle, it will follow that cognition is not *merely* embodied, but also that it can in some

[83] (See, e.g., Clark and Chalmers 1998; Clark 2010, 2003, 2008, 2015; Carter, Gordon, and Palermos 2016; Menary 2006, 2010a; Carter and Palermos 2015; Palermos 2016, 2014; Palermos and Pritchard 2013; Rowlands 2010; Hutchins 1995).

[84] See, e.g., Chalmers' forward to Clark's (2008).

circumstances be *extended* - viz., in the sense that not just extracranial but *extraorganismic* things in one's environment (e.g., notebooks, smartphones, tactile vision substitution systems[85], eyeborgs[86], etc.) can (in certain circumstances) partially *constitute* that agent's cognitive system. On this way of thinking, we not only can, but should – viz., with reference to the Parity Principle – treat Inga and Otto symmetrically in terms of memory.[87] Not only Inga, but also Otto, remembers – prior to accessing this information from storage – that the Museum of Modern Art is on 53rd street. Additionally, in Inga's case, we attribute a dispositional belief (that the Museum of Modern art is on 53rd street), in virtue of her storing this information in biomemory. And by parity of reasoning, in Otto's case, we attribute an (extended) dispositional belief with this same content, in virtue of Otto's storing the same information in (extended) memory.

Of course, for the champion of extended cognition, not *everything* that one causally interacts with while engaging with a cognitive task is going to get 'ruled in' as part of one's 'extended' cognition. Far from it. One of the key research problems in the contemporary literature on extended cognition is how exactly to distinguish cases like Otto, where the parity principle is plausibly satisfied, from cases where we should think it is not – e.g., as when one consults a phone book, or just happens to use a device for a one-off task.[88] We will be looking more closely at these issues in the next chapter, and thinking about how a plausible answer might help us to envision what extended *knowledge generating* cognitive abilities might look like (Chapters 2 and 3).

[85] (Bach-y-Rita and Kercel 2003). See also Palermos (2014) for discussion.

[86] See Pearlman (2015) for an overview of the case of Neil Harbisson's eyeborg technology, and Carter and Palermos (2016) for discussion of its significance in the wider extended cognition debate.

[87] For some recent representative work that discusses different ways to embrace kind of 'extended cognition' diagnosis of the Otto/Inga case, see, e.g., Clark (2012), Kiverstein (2018), (ed.), Carter and Kallestrup (2016), and Menary (2010b).

[88] As Allen-Hermanson (2013, 793) puts it: 'If a notebook counts as part of one's mind, then why not the yellow pages, the internet, or even parts of the natural world that supply information and support cognition?" For related worries, see Rupert (2004). For a replies to this 'cognitive-bloat' problem, see Carter and Kallestrup (2020), Clark (2010), and Palermos (2014).

But for now, it will be helpful to zoom back out and take a bird's eye view at where we've gotten to by this point in the chapter: (§1.3.1) The widespread tacit commitment of epistemologists – for the purposes of doing epistemology – to an internalist picture of the mind isn't justified; and (§1.2) there are, however, risks to remaining so committed – viz., the risk of a widening mismatch between epistemological theory and practice. Even so the epistemologist (like anyone else) should accept an internalist picture of the mind, and whatever is implied by it, if the *status* of this particular picture of the mind is sacrosanct, or deserves to be treated as a kind of theoretical pillar – one such that we should alter what comes into conflict with it, rather than to alter the pillar itself. *However* (§1.3.2) we've now seen in this section that this is hardly the case – and on the contrary – that the tide in recent cognitive science is moving against not only cognitive internalism (e.g., embodied cognition) but even against cognitive biointernalism (e.g., extended cognition).

The combined results to this point offer a presumptive case for thinking that when we – as epistemologists – face new questions posed by the influx of digital storage and generation of information, there is really no need to think we must answer them in a way that first assumes the kind of intracranialism about the mind that is incompatible with (literal) digital knowledge. Instead, a more promising approach may be to *leave such assumptions behind*, and – following trends in cognitive science – take seriously the the idea that not all of the subject matter of epistemology is intracranial. That is, it is worth working out – with the same seriousness with which we approach traditional epistemology's questions – how to decide when information stored *outside* the head (e.g., in our smartphones, apps, digital calendars, etc.) rises to the level of digital knowledge, and when it falls short, and the vast cluster of questions, under the banner of 'digital epistemology', in this neighbourhood.

A thoroughly 21st century epistemology should prioritise at least some of these questions, just as – for example – 20th century epistemology prioritised its own questions. But *which* questions of 'digital epistemology' are the important ones? The next section will canvass a proposed shortlist – viz., questions that we should want answers to (i) *from the perspective* *where we care about getting things right*, representing the world accurately rather than inaccurately, etc.; and (ii) *given that the means by which we represent the world are increasingly digitally offloaded and outsourced*.

1.4 Desiderata for a 21st Century Epistemology: A Shortlist

In addition to traditional questions mainstream epistemologists are already focusing on – questions that have tacitly been taken in epistemology to apply in the main to brainbound cognition – we should (for broadly the same kinds of reasons) also want to know the answers to a selection of pressing *new* and digitally-oriented epistemology questions, questions that become more pressing the more we continue our trends of offloading and outsourcing.

Here is, in no particular order, a shortlist:

Conversion question

What is needed to 'convert' mere correct digitally stored information into digital knowledge? To the extent that we prize knowledge over mere true opinion[89] (cf., Plato's *Meno*), we should – by parity of reasoning – value whatever makes the (structurally analogous) difference in the digital case. An understanding of just what this is will help us better articulate the 'gold standard' in digital epistemology, and to distinguish it from the rest of the accurate, reliable information we navigate on a regular basis.

Ability question

Can countenancing digital knowledge be reconciled with the idea that knowledge arises from 'ability' and if so, how? A platitude about knowledge is that it arises from ability[90] – this presumably applies, *mutatis mutandis*,

[89] See, e.g., Carter et al. (2018), Kvanvig (2003), and Pritchard (2009).

[90] For some notable expressions of this idea, see, e.g., Greco (2009), Pritchard (2012), Sosa (2010), Turri (2011), Miracchi (2015), Kelp (2018), and Carter (2016).

in the traditional case as well as in the digital case. Getting a grip on *how* digital knowledge arises from abilities will help us to better understand, by extension, how to gain more digital knowledge.

Environmental luck question

What does knowledge-undermining environmental epistemic luck look when transposed to the digital case? A familiar idea in traditional epistemology is that 'being in an epistemically bad environment' is enough to make a justified belief *unsafe* and so unknown.[91] While this idea is well-understood in the case of physical environments (e.g., an environment with holograms or facades around) it's less clear what constitutes a bad 'virtual environment'. Understanding this better – e.g., the digital analogoues of 'fake barns', etc. – will help us know when digital knowledge is undermined by luck.[92]

Anti-Sceptical question

To what extent can digital knowledge be vindicated against sceptical challenges? Standard sceptical arguments threaten to imperil (brain-bound) propositional knowledge, typically by exploiting the idea that the obtaining of certain 'deception' scenarios would be indistinguishable from the ordinary situation we find ourselves in.[93] Whether familiar anti-sceptical strategies (appealed to in the case of brain-bound cognition) are effective equally well in the case of sceptical threats to *digital* knowledge remains to be seen.

[91] For a canonical expression of this idea, see Pritchard's (2005), and in particular, the distinction between environmental and intervening knowledge-undermining (i.e., veritic) epistemic luck.

[92] For an early discussion of this kind of issue, see Carter (2013) – and, more recently – (2017).

[93] Such arguments typically appeal to epistemic closure or underdetermination principles. For discussion, see Pritchard (2016).

Defeat question

In what ways might digital knowledge stand distinctively subject to defeat? Defeaters in traditional (brainbound) epistemology are mostly well-understood.[94] However, an epistemology open to *digital* knowledge immediately invites a revision to the standard picture – particular concerning how to model mental state or psychological defeaters. These defeaters defeat *in virtue of* being possessed and counting against (viz., by rebutting or undercutting) the justification one has for believing the target proposition. Because the possession conditions for digital knowledge are more expansive than in brainbound epistemology, so likewise will be the potential sources of defeat. Getting a grip on how the mechanisms of defeat work in the digital case will allow us to better understand how we can avoid losing the digital knowledge we have.

Delegation/verification question

What is the epistemically optimal way for us to decide which tasks to outsource entirely to intelligent machines, and to what extent is further outsourcing appropriate (or not) to verify the results of that same outsourced cognition? For better or worse, we nowadays can't compete with, e.g., Google Deep-Mind when it comes to certain kinds of problem solving and discovery in science and mathematics (see §1.2); in light of this, we are increasing under pressure to rely on machines to make (at least certain kinds of) discoveries at these frontiers; we face new puzzles, however, when machines exceed human capacities not only at the level of discovery but also at the level of verification. Understanding the limits of appropriate epistemic dependence on intelligent machines will be valuable for better managing our 'hybrid' inquiries – viz., inquiries pursued by humans but 'executed' by machines.

[94] See, e.g., Brown and Simion (2021), Pollock and Cruz (1999), Bergmann (1997), and Piazza (forthcoming).

Big data and epistemic rights question

How is the EU 'Right to an Explanation' supposed to be respected in cases where we are the subject of a purely algorithmic decision-making process? According to the 2018 General Data Protection Regulations (GDPR), EU citizens have a right to an explanation (Art. 22, 13-15, Recital 71) about how purely automated decisions are made when such decisions directly affect our welfare.[95] The problem, in short, is that there is a fundamental mismatch between (i) the kind of mathematical models and decision trees used to generate the knowledge about us on which automated decisions are made (including decisions about the kinds of personalised information that is send to our apps and phones); and (ii) human styles of reasoning and interpretation. Given this mismatch, simply (e.g.) giving someone access to the underlying code explaining the automated decision will not thereby satisfy (a plausible interpretation) of a right to such an explanation. Understanding what *is* sufficient is a digitally-sourced epistemological problem with wider societal payoffs.

The reader will see that the *Conversion Question*, the *Ability Question*, the *Environmental Luck Question*, the *Anti-Sceptical Question*, and the *Defeat Question* all concern digital knowledge by way of digital information storage. These are, in effect, questions posed by cognitive outsourcing. The *Delegation/Verification Question* and the *Big Data and Epistemic Rights Question* concern digital knowledge by way of digital knowledge generation. These are questions posed by cognitive (complete) outsourcing, rather than by cognitive offloading.

The remainder of the book will address all of these questions. In some cases, the answers developed will be detailed and defended, in other cases, only provisional answers will be sketched, and roadblocks to providing fuller answers will be noted.

The 'shortlist' of questions I've given priority to here – of course – is hardly exhaustive of what a digitally-minded epistemologist will be

[95] See here Goodman and Flaxman (2017); cf., however, Wachter et al. (2017) for criticism.

curious about. Each of them, though, gets at something that, as I've indicated above, I take to be epistemically important. As these questions are relatively new, the reader is very welcome – and encouraged – to consider how my own answers in the chapters that follow might come up short, and to try to do better.

1.5 Concluding remarks

This opening chapter has aimed to introduce the reader to – as well as to *motivate* – digital epistemology as a research programme alongside traditional epistemology, and to sketch an agenda that sets the scene for the questions we'll be investigating in the rest of the book. The motivation for the project came in two parts – one positive, the other negative. The positive motivation (§1.2) involved a brief tour of some of the increasingly digitised ways we nowadays store and generate information, and in a way that is not very well suited to theorising about through the lens of traditional brainbound epistemology. The 'negative' motivation was to cast doubt upon a certain traditional way of thinking about epistemology's relationship to cognition which, if held in its clutches, would leave the idea of digitally stored or digitally generated knowledge looking radical or incoherent.

What we've seen (§1.3) is that epistemologists have no good reason to cling to ideas about cognition that are increasingly outdated in cognitive science, and this is especially so given that a natural motivation for relying on this picture – viz., that it is needed to make sense of knowledge-apt belief – is incorrect. The upshot of this negative motivation is that the epistemologist needn't face the questions posed by digital information storage and generation in a way that is artificially constrained, when it comes to our truth-directed (epistemic) evaluations of this storage and generation. By asking whether digital information can be knowledge, we can feel free to ask this literally, and to ask what else is entailed by it.

The reader already primed to think about knowledge-apt cognition unshackled from a tacit commitment to cognitive internalism might not have needed all the convincing I've attempted to do here. Some dyed-inthe wool traditional epistemologists might have needed more. Even so, I hope to have at least laid out a case for thinking that there's some (radically) different work for 21st century epistemologists to grapple with, and for those not persuaded that we should take the idea of digital knowledge literally (as I am suggesting we do), then the remaining chapters should hopefully still be of interest – as they show *how* new problems stemming from cognitive offloading and outsourcing might be addressed, and in doing so will invite improvement.

Chapters 2-4 will address the 'offloading question' – viz., the *Conversion Question*, the *Ability Question*, the *Environmental Luck Question*, the *Anti-Sceptical Question*, and the *Defeat Question*, and 5-6 will focus on the 'outsourcing questions' viz., the *Delegation/verification question* and the *Big data and epistemic rights question*.