4 Distributed traces and the causal theory of constructive memory

John Sutton and Gerard O'Brien

4.1 Memory as causal

A student remembers her surprise birthday party (Selwood, Harris, Barnier, & Sutton 2020, p. 415). In her apartment after work, she found balloons, then found her friends hiding in the living room; there was cake, and a crown with her name on it, and she put on 'something nicer'; someone spilled wine on the carpet, its 'first red stain'. Then 'all of a sudden our neighbour was in our living room. The music stopped and everyone froze. Especially me. I sank down on the sofa, and was so embarrassed'. She had trouble getting people out, but 'everything was fine': they all went to the pub, and later ate pie on the way home.

Though elicited in the peculiar context of an experiment (in this case, a study of how remembering events on your own differs from recalling them collaboratively), this is an otherwise unexceptional report of a personally significant past event. That event was complex and structured: it occurred within a delimited time frame but extended over a number of component episodes. The student experienced it actively, interpreting what was happening in her own way, attending to select, salient aspects of the event. For her now, later, to remember it is to weave a complex mental tapestry of perceptual, affective, and conceptual threads: among others, sets of dynamic visual images of locations and scenes, and of detailed visual experiences of people, objects, and actions; related sets of olfactory, auditory, gustatory, and embodied images, of the smell, the chatter and the music, the cake and the drinks, the bodily feelings of pleasure, or of sinking in embarrassment; and an array of affects and concepts, images, and attitudes relating to her feeling and thoughts at the time, and to her take on the social and emotional relations among her friends.

This person, it's natural to think, remembers her surprise party partly *because* she experienced it. While remembering it now is an activity prompted by an experimenter's request and influenced by many other factors, the recalled episode has a more-or-less integrated place in her past as a result of *causal* connections of certain kinds between it and the present remembrance. Significant personal events are woven in to more or less coherent autobiographical narratives that form and maintain our sense of the causal connectedness of events and actions in time (J. Campbell 1997; Ismael 2016).

DOI: 10.4324/9781003002277-7

It's also natural to think that the memory is in some sense *embodied*, or *carried* with the person over time. It may be highly context-sensitive, in that what and how she remembers may change in light of later events, or in reports to different audiences, and in that the factual and emotional details, the accuracy, and the qualitative experience of remembering may vary over repeated retellings (Temler, Barnier, Sutton, & McIlwain 2020). But the student's capacity to recall these episodes is transportable, typically *not* tied to one context alone. Though there is no guarantee in any particular case, she takes her access to *some* such events in her personal past with her, even if her path through life takes her far from their location and from any direct reminders (Sutton 2009).

While remembering is a situated activity in the present, it also has vital *diachronic* aspects. It typically involves *past* events, and is one ingredient among others that make us creatures *with* a past to which, in remembering, we fallibly lay claim (S. Campbell 2004). The forms of causation in play here, across multiple experiences, memories, and emotions, and over the weave of a life, are neither simple nor easy to track. This is not the collision of isolated billiard balls, but rather causation as sedimentation, where causal connections are multiple, indirect, and context-dependent (Deutscher 1989, p. 61; Sutton 1998, p. 308).

4.2 The causal theory of memory

Such considerations motivate the causal theory of memory (CTM; Martin & Deutscher 1966). CTM aims to illuminate them by invoking causal processes operative in remembering a past event, and memory traces which are in some sense *about* (aspects of) the event from which they derive.

The causal theory of memory is intended as an objective, third-person account, to catch the causal basis of memory's significance. Martin and Deutscher respected the diversity and richness of our language and practices around remembering, while refusing to treat the individual subject's perspective as the final authority (Deutscher 1989; Sutton & Windhorst 2009). CTM seeks objective marks or criteria by which to identify remembering. Such criteria, it is hoped, distinguish genuine memory of a past event from other phenomena: from, for example, merely appearing to remember it, from (merely) imagining it, or from knowing about it from a source other than personal experience (such as relearning it, after previously forgetting it).

In the case of the student's memory of the surprise party, for example, we can ask how traces of the spilling of the wine or of the student's embarrassment operate causally within the complex processes driving her active remembering, alongside the experimenter's request for such a narrative. We can also ask how closely the remembered episode matches the student's original experiences, if for some reason we need to assess its status as a genuine memory.

These are not questions about *subjective* differences between remembering and (say) imagining, about how in psychological practice we come (fallibly) to *take* ourselves sometimes to be remembering, sometimes to be imagining

(Michaelian 2016, pp. 71, 120). CTM allows for genuine error: I may well take myself to be remembering, when in fact I am not.

While psychologists and cognitive neuroscientists working on memory have typically acknowledged its diachronic nature (Tulving 2007, p. 66), few show specific interest in the CTM. Their projects are empirical, explanatory, or descriptive: they study the nature, mechanisms, processes, and operations of human memory across interacting social, behavioural, phenomenological, cognitive, and neural dimensions. Many philosophers, of course, share these interests and aims, participating alongside scientists in what Michaelian calls 'the naturalistic project of describing memory as a psychologically real process' (2016, p. 69). CTM is often rightly modified to fit better with memory science, and we are about to add to decades of debate on the implications of the resulting constraints.

Before we get to this core business, though, we note in advance that we do not claim that an understanding of human memory is exhausted by reference to the neurocomputational processes in question. It is entirely compatible with (though strictly independent of) all we say here to see them, rather, as the bio-internal wings or components of broader systems spanning brain, body, and world. In particular, we suggest in our final section that these descriptive facts about memory neither dictate nor exhaust the content of the causal theory. We focus there on 'external' forms of context-sensitivity, to add to the rich 'internal' context-sensitivity which is our primary concern. Future work that treats these forms of context-sensitivity together, we will suggest briefly, may engender more realistic assessment of relations between CTM and the cognitive neurosciences of memory.

4.3 The causal theory of constructive memory

Our work in the 1980s and 1990s was firmly in that naturalistic camp, assessing in different ways the philosophical implications of connectionism, where the superpositional 'storage' of multiple memories over the same resources gives rise to *distributed* traces, with many 'representations' in any one 'representing' (Haugeland 1991; McClelland & Rumelhart 1986; O'Brien 1993; Sutton 1998; van Gelder 1991). While CTM was not our primary concern, it was clear that connectionism would require some standard notions of mental content and mental causation to be either rejected or substantially liberalised. We adopted the latter, revisionary option: O'Brien argued against the elimination of content in connectionism (O'Brien 1991; Ramsey, Stich, & Garon 1991), and Sutton (1998) defended a connectionist account of contentful memory traces against a variety of critics (pp. 298–316). These were ongoing discussions: an explicit debate with Deutscher over the extent to which 'the causal analysis is in some tension with the extent to which remembering is a constructive activity' (Sutton & Windhorst 2009, p. 79, with a helpful response in Deutscher 2009, pp. 97–98) suggests that we were not 'slow to recognize that the distributed conception of traces may be in tension with the contentful conception of traces' (Michaelian & Sant'Anna 2021, p. S323). We do accept that 'there is a pressing need for further work' on the causal theory of constructive memory (Michaelian & Robins 2018, p. 21), and so we aim here to combat claims that 'distributed memory traces are incompatible with the CTM' (Robins 2016, p. 2994) and that 'the widely-adopted distributed conception of traces … [leads] inexorably to the contentless conception' (Michaelian & Sant'Anna 2021, p. S319).

In what follows, we briefly rehearse relevant features of connectionism and assess its application to the forms of memory in question. We identify two important, related aspects of these forms of memory which are sometimes neglected, but which lie at the heart of connectionist approaches. Our primary focus (in Section 4.6) is in applying a novel account of content for distributed representations to the debate on memory. We also sketch an approach to explaining the context-sensitive causal roles of such distributed traces. Finally, as promised, we pan back, considering the causal theory of constructive memory in relation to the development and pragmatics of practices of remembering.

We focus on challenges to the causal theory of constructive memory that arise from consideration of specific features of distributed traces, rather than those motivated only by general concerns about mental content or mental causation. We adopt Michaelian's starting point, an investigation of human remembering 'as it unfolds in the real world', as opposed to deploying an 'analytic methodology' that seeks 'an account of remembering immune to all possible counter-examples' (2016, pp. 3–4). We do not seek sharp necessary and sufficient conditions that apply across all contexts. Clear cases of remembering, and cases which are clearly not remembering, can legitimately be identified even if other cases are uncertain, the subject of reasonable disagreement. Rather than watertight 'analysis', then, the empirically informed view we defend embraces some vagueness at the edges, not as 'a tolerable defect' (Michaelian 2016, p. 91) but as a pointer to memory's deep context-sensitivity.

There are good, independent reasons to retain both the causal theory of memory and the idea that the traces it invokes are distributed. So the conjunction of these views is worth defending. The challenges posed by holding both – by defending the causal theory of constructive memory, or CTCM – are real puzzles about memory and its place in our lives, not mere artefacts of theory. These are difficult topics, which require precisely navigated integrations of challenging and changing fields in neuroscience, cognitive and developmental psychology, and many areas of philosophy. There are principled reasons why it's difficult to pin down what counts as an 'appropriate causal connection' between past experience and present recall, both in general and in particular cases. The point that claims to truth in memory are often desperately hard to assess, in theory and in practice, should not be surprising, and should not be a reason to give up on them. Both causal approaches to memory and distributed traces are valuable: identifying possible tensions between them should motivate us not to jettison commitment to one or to both, but to find clearer and more creative responses.

4.4 Connectionism and episodic memory

As a neurally inspired rival to the classical computational theory of mind (Fodor 1975), connectionism captures structural and temporal properties of the brain's neural networks in virtue of the way it deploys transient *activation patterns* and enduring sets of *connection weights* (O'Brien & Opie 1999; for an overview, see Buckner & Garson 2019).

Each unit in a connectionist network has an activation level (modelled on a neuron's spiking frequency) that is communicated to other units in the network via modifiable, weighted connections (modelled on synapses). From moment to moment, each unit sums the weighted activation it receives and generates a new activation level that is some threshold function of its current activity and that input. This is how a network responds to its inputs, generating a stable pattern of activity across its constituent units. Altering the network's connection weights alters the activation patterns it produces. Consequently, a network can learn to generate a range of target patterns in response to a range of inputs. These stable patterns of activation, generated rapidly in response to the flux of input impinging on individual networks, are taken by connectionists to constitute a transient form of information coding, often referred to as *activation pattern representation*. Activation patterns are vehicles of *explicit* representation: there is a *one-to-one* relationship between a specific activation pattern and an element of the network's representational domain.

While activation patterns are transient features of connectionist networks, a trained network, in virtue of the particular configuration of its connection weights, has a longer-term capacity to generate a set of target activation patterns, in response to cueing inputs. This second form of information coding, referred to as *connection weight representation*, is the basis of long-term memory in connectionist systems. Such long-term storage of information, by contrast with activation pattern representation, is *superpositional* in nature, since each connection weight contributes to the 'storage' of every stable activation pattern the network can generate. This is the connectionist implementation of *distributed* representation. The information stored in a network is not encoded in a physically discrete manner. Instead, a single appropriately configured network encodes a set of contents in a way that grounds its capacity to produce a set of activation patterns: there is a *one-to-many* relationship between this complex of connection weights and the elements that compose its representational domain (Clark 1993; Churchland 1995).

Much connectionist research with individual networks is conducted at the level of relatively low-level perceptual and categorical tasks such as colour categorisation and face recognition. Such tasks are modelled by networks of a small number of units and layers (relative to the brain, that is). The episodic form of memory exemplified by the surprise birthday example, in contrast, is a high-level cognitive achievement, implicating multiple sensory and executive pathways, each of which involves a myriad of neural circuits across the brain. Social and contextual features of the retrieval context may iteratively influence ongoing neural processes: as I narrate a past experience, someone else's responses can shape the content, form, and course of my remembering, in continuous reciprocal interactive causation. The relevant neurocognitive architectures are multi-level, integrating diverse cognitive domains across nested networks of interacting networks (De Brigard 2014a, 2014b). Despite other theoretical differences, cognitive neuroscientists agree that memory 'systems are much more interactive than we once thought' (Moscovitch et al. 2016, p. 124). The forms of representation and computation at the heart of connectionist neurophilosophy do ground the high-level neurocomputational architecture of episodic memory. But contemporary connectionism on its own does not offer an exhaustive theory of memory and cognition. It needs to be supplemented, first and at least, by incorporation into a fuller cognitive neuroscience.[1]

Yet leading connectionists from the outset did intend to generate 'a distributed model of human learning and memory' (McClelland & Rumelhart 1986), and questions about integration across levels continue to drive research in this field.[2] For our purposes, connectionism remains the best mechanistic demonstration of how the computational processing of representations gives rise to properties of generalisation, integration, and context-sensitive pattern-transformation which are also apparent in the dynamics of human memory (Kumaran & McClelland 2012; McClelland 1995; see Section 4.5).[3]

Let's return to our student's remembrance of the surprise birthday party. In the context of our experiment, this memory likely arose through 'generative' retrieval, a top-down, cue-driven deliberate process, rather than effortless 'direct' retrieval (Harris, O'Connor, & Sutton 2015). We did not image the student's brain in this study, but the characteristic time course and spatiotemporal dynamics of such retrieval processes are fairly well understood (Daselaar et al. 2008; Greenberg & Rubin 2003). On any contemporary neurocognitive view, this episodic memory implicates processing across an enormously complex web of interconnected neural networks, including sensory, emotional, and higher cognitive pathways, together with circuits in the hippocampus and the prefrontal cortex. At this global level, activation spreads as we remember events, rising and falling across interconnected networks of the networks that operate together as we identify, relive, feel, and perhaps narrate such structured, personally significant past experiences in real time.[4]

Fast changes in the stabilisation of all these networks cascade across the brain, generating myriad local activation patterns over specific component networks as the experience takes its course. From a connectionist perspective, such rapid sequences of localised activation pattern stabilisations within each of these networks are the (often minimal or fragmentary) components of representational content, which combine to generate the complex cascades of molecular representational states that constitute ongoing experience. These more global states, in virtue of the activation patterns from which they are composed, are explicit representing vehicles with discrete representational contents. As, for example, memory of the cake and the crown gives way to memories of the spilled wine and of embarrassment at the neighbour's arrival, stabilised patterns in gustatory and visual cortex are replaced by further stabilisations in emotional and kinaesthetic networks, such sequences generating the dynamic quality of the remembering experience across a few seconds.

As this is *constructive* memory, the activation patterns generated during this experience may differ markedly from those that arose during the party itself: from such a diachronic perspective, they will likely be both selective (discarding many details of the original experience), and generative (adding extra details). Such changes arise in constructive processes at any stage along the way, with extra content integrated from other sources, or generated from traces of other experiences. For the causal theorist of constructive memory, the visual details the student now retrieves need not match her visual experience at the time, and can even go beyond it to some extent: but with its diachronic focus, the causal theory expects *some* similarity here, such that the remembered details should 'not go too far beyond' those that were experienced (Michaelian 2016, p. 92).

The ongoing sequence of activation patterns takes the form it does, according to connectionism, ultimately in virtue of the configurations of connection weights at the level of the constituent networks. These configurations were so modified during the birthday party (and afterwards, in ongoing consolidation and reconsolidation) that now, in remembering, they enable the reconstruction of the activation patterns that contributed representational contents to the original experiences, even if these patterns are in certain respects now partial, impoverished, or altered. These configurations of connection weights constitute the distributed memory traces that mediate between experiences and their retrieval. They are also the source of connectionism's greatest strength as an approach to memory, and of the critics' concerns about its compatibility with causal theories. We consider each point in turn.

4.5 Dispositional memory and causal holism

Connectionist networks differ markedly from their classical computational counterparts in encoding information superpositionally over the long term. We argue that this is a great strength of connectionist computation: since *all* of the information encoded across a network is causally implicated *every time* the network processes an input, there is no need to process individual items of information separately. For connectionists, this 'causal holism' has the potential to overcome the vicious problems of computational intractability that stymied classical AI (O'Brien & Opie 1999, 2009). We focus here on its additional and distinctive philosophical pay-off. Connectionism, as the heart of the causal theory of constructive memory, catches two important, related aspects of human remembering: the difference between occurrent and dispositional memory, and memory's constructive or generative nature. These are features familiar in both ordinary and scientific views of memory, not exclusive to any one discourse or project. While they are by no means universally appreciated, neither are they new discoveries. A theory of memory that respects or – better – offers detailed accounts of these features is to be preferred to one that says little about them.

When the student told us about the surprise birthday party, her remembering was an activity at a particular time: an *occurrent* memory. But before and again after this exercise of her memory, she still – in another sense – remembers the party, even when she is not actively, presently thinking about it at all: this enduring capacity of hers is *dispositional* memory. This distinction is easily grasped, and long acknowledged – it was clearly articulated by Aristotle and by Locke (Sutton 1998, 2020a). Although the student's memories of these events may be partial or incorrect in various ways, and may of course ultimately become inaccessible, she doesn't simply *forget* them when she is remembering something else, or when she is swimming, or sleeping. Our best sciences of memory should acknowledge and explain both forms of memory, and their relations.

It is natural to characterise dispositional remembering in two slightly more substantive ways. Not only, we reasonably think, do memories endure when non-occurrent, but they continue to matter or make a difference: they remain part of our history and of who we are, and they have ongoing causal efficacy even when not explicitly in mind. This is not to posit one integrated global memory trace that fixes or binds *all* events of the surprise party in some single, unitary form. It is merely to claim that those component traces do endure, in some form, fragmentary or dispersed or superposed as that may be: they are neither impotent nor lost, not dissipated entirely between experience and recall. It makes a difference to the student's ongoing mental life that she lived through those events and remembers them, however imperfectly. This was a vital plank in the connectionist resistance to classical cognitivism: mental causation does *not* (as Fodor thought) require explicit representation.

Further, such enduring dispositional memories operate holistically. It's not that each event in my personal past is retained separately, sitting passively in cold storage until accessed at will, pulled out again just as it was first experienced, to make a single causal contribution to ongoing processing. Rather, we are well aware of the dynamics and context-sensitivity of memory's ongoing operation. This is not a point about memory's frailty and fragility, its errors and confusions. It's that we are all accustomed to changes in the significance, implications, and content of what we recall, sometimes outside awareness or control, sometimes in line with our changing epistemic, emotional, or evaluative perspectives on our past (Goldie 2012, pp. 26–55). Experiences we recall more or less reliably need not show up in ongoing mental life in static form: they contribute both to other activities of remembering, and to an array of other psychological operations, even when not explicit. Much of what we remember is updated as we have other related experiences, and much of what we remember integrates into or guides our ongoing cognition and action even when it's not occurrently active. As well as – sometimes – successfully recalling specific past events, human minds also tend easily to link new experiences with relevant memories, and to generalise across memories with similar content.

Such causal holism falls naturally out of the connectionist picture of distributed memory traces we outlined earlier (O'Brien 1998, p. 82). Because superpositional 'storage' – which is not distinct from ongoing processing, as it is in classical computation – creates such 'composite' memories, it has long been recognised as a detailed mechanistic implementation of a non-archival model of memory, and of the alternative view of memory processes as creative, selective, generative, and dynamic.[5] So far, so good – the context-sensitivity and content-addressability of connectionist processing drives flexible generalisation, the capacity to extract the central tendencies of a set of experiences, and other apparent characteristics of human mental life (McClelland & Rumelhart 1986, p. 193; Clark 1989, p. 99). On this measure, distributed representations as postulated in connectionism clearly outperform old and new localist accounts of engrams or memory traces. These are, we submit, strong considerations in favour of the connectionist account of distributed traces as an approach to memory, before any countervailing challenges are considered.

4.6 Distributed traces, content transmission, and causal history

We now respond to the two central concerns about the compatibility of distributed memory traces with CTCM. The first is that distributed memory traces cannot *transmit content* from experience to remembering in the way that, for the causal theorist, distinguishes cases of remembering from non-memorial forms of retention. The second is that distributed memory traces cannot appropriately mediate the requisite *causal history* between original experiences and their retrieval. Our task now is to explain how connectionist computation, contrary to critics' concerns, has the representational resources to implement memory traces that can transmit content and mediate appropriate causal histories between experience and retrieval.

4.6.1 The challenge to contentful traces

We defend – albeit in significantly revised form – the standard causal theorist's claim that there is some transmission of content between experience and retrieval.[6] If the causal connection is sustained by memory traces, such traces must be *representing vehicles* that convey (at least some) content of experienced episodes to subsequent remembering. The connectionist CTCM seems to have this covered. As we have seen, connectionists talk in terms of information *stored* long-term as connection weight representations. These configurations of connection weights appear to be representing vehicles capable of conveying contents over time in memory, where no such content transmission will be operative in either relearning or merely imagining the events.

But some philosophers of memory are not so sure. They worry that talk of connection weights 'storing' information is misleading. Because connectionists set their caps against the static storage of discrete items typical of classical models, it is easy to think they are rejecting content entirely. Superposition is a form of 'storage', but of a non-conventional form that can be tricky to grasp. As Elman (1993) wrote, in connectionism,

> once a given pattern has been processed and the network has been updated, the data disappear. Their effect is immediate and results in a modification of the knowledge state of the network. The data persist only by virtue of the effect they have on what the network knows.
>
> (p. 89)

So when De Brigard notes that '"storing" is a rather misleading term', and characterises a memory trace as a 'dispositional property to reinstate … the complex hippocampal-neocortical pattern of neural activation' (2014a, p. 411, 2014c, p. 169), some commentators read him as *rejecting* content (Hutto & Peeters 2018, p. 105; Michaelian & Sant'Anna 2021, p.S314; Hutto, Chapter 3, this volume). Such critics are happy to describe connectionist memory traces in terms of the *dispositional* properties of neural networks, but take this to mean that 'strictly speaking, no content is stored' (Michaelian & Sant'Anna 2021, p.S324).[7]

Debate about the representational credentials of configurations of network connection weights is not new. William Ramsey noted that whereas in

> classical models it is typically the case that causally distinct structures encode commands for specific stages of the computation … in trained connectionist models, this type of specificity is not possible. While it might be true that some connection weights contribute to some episodes of processing more than others, there is no level of analysis at which we can say a particular weight encodes a particular command or governs a specific algorithmic step in the computation. Instead, all the system's know-how is superimposed on all the weights with no particular mappings between the two.
>
> (Ramsey 1997, pp. 48–49)

Ramsey is targeting the specifically *distributed* nature of connectionist memory traces: the fact the information is *superpositionally* encoded across a network's connection weights, with a *one-to-many* relationship between this configuration of weights and the elements that compose the network's representational domain. This leads Ramsey to conclude that 'there doesn't appear to be any other level of understanding or explanatory motivation that requires us to view the weights as representations' (1997, p. 51).

If connectionist memory traces are not representing vehicles in good standing, they cannot transmit content from experience to retrieval. In defending a connectionist CTCM, therefore, we need to show that such scepticism about the representational credentials of connection weights is misplaced. Our novel response which follows differs in detail from those offered by other connectionists (Churchland 2012; Haybron 2000; Shea 2007). While it is true that direct application of these debates to the philosophy of memory in particular is relatively recent, critics of content in connectionism need to acknowledge and answer these distinct and detailed views. They should not simply ignore them while asserting that distributed content is incoherent.

4.6.2 Why we need contentful traces

We contest a 'purely' dispositional, contentless interpretation of connectionism by focussing on the *explanatory gap* between the microphysical properties of connectionist networks and their capacity to successfully navigate their task domains. Architecturally identical networks trained up on to the same task domain but from distinct random initial assignments of connection weights come to occupy different points in 'weight space'.[8] As a consequence, each trained network responds to the *same* inputs by generating patterns of activation that occupy *different* points in 'activation space'.[9] At the microphysical or barely neural level of individual weighted connections and unit activation values, therefore, these different networks have nothing in common. From this microphysical perspective, nothing distinguishes those networks capable of successful performance in the task domain from those that are not.

This is what's wrong with the attempt to explain episodic memory entirely 'without a trace'. Consider, for example, the alternative offered by Hutto to accompany his attack on 'the explanatory vacuity' of our accounts of implicit, content-carrying distributed representations (Chapter 3, this volume, note 11). On his view, remembering does involve or require some internal similarity over time. But, he insists, this is not similarity of *content*. Rather, such similarity is to be sought at the level of 'specific neural patterns' alone: it is not a *representational* similarity, not even a psychological similarity. At the micro-neural level to which Hutto directs us, he expects to find 'the re-instantiation' of such 'specific neural patterns'. Such 'recreated neural patterns' will 'suffice' for memory if in 're-occurring' they are 'similar enough to those that occurred during the original experience' (Chapter 3, this volume, p. xx). Hutto repeatedly invokes this kind of neural commonality over time, and it is all he offers as a putative explanatory mechanism for memory. The 'replication' of a 'neural pattern' must be 'sufficiently similar to that which underwrote the original, remembered experience' (Chapter 3, this volume, p. xx).

Given that Hutto claims to acknowledge the extent of dynamic neural redeployment in relevant systems, it is surprising that he is so confident about finding such similarity, about the genuine re-instantiation or replication of 'specific neural patterns' across time and context. It is an empirical matter, but we believe there is little reason to expect any such 're-creation' or 're-instantiation' of specific patterns identified at the microphysical or neural level alone. We might also expect to hear more from Hutto about what these 'specific neural patterns' might be; about how they change over time; and about what this 'similarity' consists in, or how it is even in principle to be identified. Such a thin alternative does not successfully fulfil Hutto's 'radical' wish to eliminate traces and memory content.

Like other theorists of distributed connectionist content, we aim in what follows to offer, in contrast, full and detailed accounts of the kind of commonality we *are* likely to find, across distinct occasions (for example) on which the student remembers the same surprise party. What, then, *explains* the successful performance of networks that have nothing in common at the microphysical or barely neural level? To answer, and to defend the contentful conception of traces, we have to ascend to a higher, emergent level of description, and to show why reference to the *representational* capacities of connectionist networks is ineliminable.

4.6.3 How we get contentful traces

Successful networks differ from their unsuccessful counterparts in that they embody, within their configuration of connection weights, sufficiently accurate *structural models* of the task domain. These structural models, multiply realisable in the microphysical substrate of connectionist networks, are acquired during a network's training regime. Once in place, they govern a network's response to any input and enable it to relax into a stable activation pattern that corresponds with the region of the embodied representational landscape that constitutes the response to the input. We now demonstrate that connection weight representing vehicles earn their explanatory keep by explaining how connectionist networks compute (O'Brien & Opie 2006).

It is a well-known feature of connectionism that a network trained on a corpus of inputs constructs an activation pattern landscape partitioned into separable regions corresponding to salient categorical distinctions between the elements that compose its task domain (Clark 1993; Churchland 1995). What is not always appreciated about these activation patterns, however, is that collectively, they *structurally resemble* aspects of the task domain over which the network has been trained.[10] Indeed, this structural resemblance relation anchors the representational interpretation of activation patterns (O'Brien & Opie 2004).[11] With this representational interpretation of activation patterns in hand, we can see what different connectionist networks trained on the same task domain have in common, despite their microphysical differences: they each realise the same activation pattern representational landscape.

The deeper commonality between networks trained up on the same corpus of inputs – the fact that their connection weights embody a structural model of the task domain – requires some teasing out. Key players in network processing are what O'Brien and Opie call *fan-ins* (2006). A fan-in is the vector of weights modulating the effect of incoming activity on a particular hidden unit. Any feedforward network has one fan-in per hidden unit, each being to a row of the network's hidden layer weight matrix. Fan-ins effect the transformation of the network's input space into its hidden unit activation space. Specifically, each fan-in determines how one hidden unit responds to input, by way of a product of input activation and fan-in values. This product is then modified by the hidden unit's activation function to produce the value along a single coordinate in activation space. A network's fan-ins thus interface directly with the structure of the vectors coded at the input layer, and ultimately determine the structure of activation space. Most importantly for our purposes, investigations of familiar feedforward networks trained to perform such tasks as face recognition and colour categorisation (Laakso & Cottrell 2000) reveal that the configurations of connection weights that compose fan-ins structurally resemble aspects of the network's task domain (for details, see O'Brien & Opie 2006).

This second relation of systemic or structural resemblance secures an interpretation of connection weights as representing vehicles. Furthermore, this representational interpretation of network connection weights explains the successful performance of microphysically divergent networks operating in the same task domain. Despite their microphysical differences, each of these networks in the course of its training regime has sculpted a configuration of connection weights that structurally resembles, and hence represents, the task domain. Because the connection weights are ultimately responsible for the patterns of activity that networks generate in response to their inputs, this also explains why each of these networks achieves the same representational landscape in activation space: the structural resemblance embodied in its connection weights is causally responsible for the structural resemblance embodied in its activation space. These commonalities between the networks are invisible at the microphysical level of description. They are revealed only when we ascend to the abstract level at which these structural resemblance relations reside. Hence, only at this emergent level is there an illuminating explanation of their performance.

The foregoing, we think, is a strong and detailed response to those critics who charge that the distributed memory traces invoked by connectionist versions of CTCM are incapable of transmitting content between original and remembering experiences. Our discussion of the representational credentials and computational dynamics of connectionist systems derives from analysis of relatively simple networks performing straightforward categorical tasks (O'Brien & Opie 2006), but there is no reason to think the lessons don't generalise to more complex systems and cognitive capacities. Network connection weights are contentful representing vehicles anchored in structural resemblance relations with their task domains. As such, they are capable of transmitting contents between the activation patterns that originally established these relations of resemblance and the subsequent activation patterns they are instrumental in recreating. Again – explaining episodic memory in terms of 'specific neural patterns' alone, with no invocation of content, does not work. Far from discarding the notion of a representational trace, we *require* reference not only to weighted connections, but to the contentful representing vehicles which they constitute.

4.6.4 Distributed traces and causal history

Commentators raise a second concern about the compatibility of distributed memory traces with CTCM. What makes CTCM a *causal* theory of memory is its commitment to an appropriate causal connection between experience and remembering, mediated by memory traces. Not any old causal connection is enough: this causal connection should distinguish memories from one another, and remembering from relearning. Thus 'the causal chain leading back to the experience must be distinguishable from other causal chains' (Michaelian & Robins, 2018, p. 17). Perhaps the connectionist rendering of CTCM violates this requirement, precisely because it holds that memory traces encode information in a superpositional fashion:

> The traditional conception of traces involves fixed, explicit contents carried by distinct local vehicles … Proponents of distributed conceptions challenge this matrix of ideas, arguing that we should give up at least some of the features of the traditional conception…. If [the connectionist view] is right, we may be able to refer to memory traces in a loose sense, since a specific experience will result in a specific modification of connections in the network, but … there are no traces in the sense of *distinct vehicles* carrying *distinct contents*…. [It thus] remains unclear how … distributed causal theorists would have us understand the nature of the causal connection between retrieved memories and experiences.
>
> (Michaelian & Robins 2018, p. 21, original emphasis)

Like humans, connectionist models are good at generalising and integrating information. But perhaps there is an unacceptable cost. Does connectionism rule out a capacity to remember specific items or events? For Robins, its characteristic blending effects come 'at the expense of retaining the specifics of any particular past event': because 'the effect of any particular pattern will wash out over time, … distributed traces do not have individually distinguishable causal histories' (2016, pp. 3005, 3008). But if the CTM requires 'the possibility of tracing the unique causal influence of a particular past event up until the time that it is remembered', Robins claims, 'memory must be structured so as to retain discrete traces for each past experience a person is capable of recalling' (2016, p. 3011). This line of thought, if successful, would rule out connectionism, with its distributed traces, as a model of human memory: 'distributed network accounts of memory traces do not provide a way to track the causal history of memories for particular past events' (Robins 2016, p. 3009).

We postpone an equivalently full response to this concern to another occasion, partly for reasons of space: we will be content if our treatment of distributed content finds favour. We also acknowledge the difficulty of meeting this challenge, and of understanding better the complex causal nexus in which distributed traces are involved, and the senses in which connectionist causation is a matter of degree (Haybron 2000, p. 367). We offer here only a few promissory remarks on how these issues might be approached.

First, while we agree that a theory must allow for humans to remember specific past events, even if often partially and imperfectly, such specificity should not be over-emphasised. Episodic memory very often does not involve or require uniquely distinguishable causal pathways back to sharply delimited past events, because very often its content is not singular: repeated, recurrent, generic, or phasic events are very often remembered (Andonovski 2020; Mac Cumhaill 2020; Schechtman 1994; Sutton 1998). In such cases, the pressure on

CTCM to provide a way of singling out a specific causal connection running back to one event alone is released.

Secondly, we accept that the connectionist approach must significantly liberalise the way we understand how specific past events are singled out in memory, when they are. In refusing to posit discrete enduring traces for each past experience, we are in a sense treating representations of individual events 'like the town of Brigadoon' (Damasio 1995, pp. 103–104; Sutton 2004, p. 513), coming into being again only on the spot and rarely. The causal residues of many past events are smeared into all ongoing processing in a network, such that the ongoing presence of the before in the after is composite. This is not to rule out all determinacy in the relevant causal pathways: experiences alter connection weights at particular times, modifying the causal powers of a network in a determinate fashion, such that the network is now capable of responding to further input in new, different, or specific ways. But in any or most episodes of remembering, many such causal pathways, back to many past events, will be partly contributing to the occurrent processing. Perhaps there are resources in treatments of complex causation in metaphysics and philosophy of science to support the idea of distinctions between these various causal pathways running from experience to remembering, so as to provide a way of singling out a specific causal contribution running back to one event. But if not, we suggest, connectionists can simply bite the bullet here and show that CTCM can survive even this.

So, thirdly, we can challenge the assumption that a causal theory of memory requires a unique and distinguishable causal connection running through from a particular experience to retrieval. Instead, we might propose that memory requires *some* causal connection between them (though not necessarily a unique and distinguishable one), and some appropriate relation of content between the experiences then and now. In other words, CTCM would not rely on a unique causal connection to ensure that a memory is tied to a particular event. Instead, it would rely on a (liberalised) causal connection plus similarity of content.

In some cases this will result in a degree of indeterminacy as to whether a subject is remembering one particular event among a range of similar events. But is this a problem? This is what memory is like. The causal connection postulated by CTM does not have to be useable in practice, the causal history not necessarily itself recoverable. What matters in practice, rather, in distinguishing memory of one event from another, or remembering from imagining, or personal memory from testimony, will sometimes be given not by tracing these internal causal pathways, but by way of features of the external context. This returns us to the point we flagged briefly earlier, about the scope of CTM in relation to the broader pragmatic contexts of remembering.

4.7 Causal theories in context: Pragmatics and development

The amounts and forms of similarity between patterns activated at experience and at retrieval vary substantially. Causal connections between experiences, memory traces, and acts of remembering are complex and variable. It may turn out that distinguishably unique causal pathways cannot be identified in the midst of such complex internal computational processing across long periods. We conclude by suggesting that this outcome would signal not that the causal theory of memory has failed, but merely that it is incomplete.

At their most ambitious, causal theorists hope to pin down criteria to distinguish genuine remembering from many other cognitive phenomena. The idea is that true remembering is intuitively different from mere imagining, and from falsely 'remembering' or merely seeming to remember. Remembering on the basis of personal experience differs from knowing something because it has been relearned, or on the basis of testimony. It is also different from thinking – even in an episodic way – about future events or counterfactual events.

Part of the task in making progress towards this goal is to latch on to our best sciences of memory, to understand better the neurocognitive mechanisms and operations in play over the many phases of these complex processes. We have sought to contribute to that naturalistic project of adjusting the causal theory for better fit with the cognitive neurosciences. But the concerns of the causal theory, we can now see, are also broader than this. The causal theorist is supplementing the naturalistic project with something extra, doing something beyond drawing lessons from empirical science. We conclude with some initial thoughts on this.

There are many distinctive versions of CTM (for examples from early in the recent revival of philosophy of memory, see Bernecker 2010; Debus 2010; Naylor 2012). Differences and disagreements between them are not due to misunderstanding of, or lack of access to, extra facts about neurocognitive processes or the paths of internal causal transmission. In assessing candidate criteria for genuine memory against thought experiments or real-world cases, philosophers disagree on what counts as, for example, 'a causal link of the right kind' between past experience and present recall, evaluating intuitions about what feels 'not quite right' (Debus 2017, pp. 67–68). Likewise when we turn from disagreement over theory, criteria, or general intuitions to disagreement over particular cases. In practice, in applying the norms or standards by which certain phenomena and not others are accepted as genuine cases of remembering, people (individually and collectively) do not typically treat further neurocognitive facts about a particular case as vital or decisive for assessing a memory claim. Such facts may be relevant in one way or another, but they will be assessed alongside 'external' facts about the history and situation, considerations of plausibility, and the many other context-dependent features we factor into our attempts to resolve uncertainty or disagreement about whether someone is really remembering an event.

In socio-cognitive practice, both these standards themselves and the thresholds at which they are applied change, and they operate differently in different settings. The criteria we apply to claims to remember are partly pragmatic: different thresholds and standards apply when narrating a past event in casual conversation among friends, compared to bearing witness in a court of law. Although complex cases or situations can stretch and trouble our shared practices, much everyday memory talk is effortlessly sensitive to context, as people adjust the standards by which they assess claims about the past to fit distinctive settings and functions, from courtroom to dinner table, or in conversations with their partners, therapists, or bosses. Such normative, epistemic, or pragmatic considerations, highlighted again in recent philosophy of memory (Andonovski 2021; Craver 2020), were often previously discussed as forms of deep context-sensitivity (Campbell 2006; Deutscher 1989; Neisser 1982; Sutton 2003).

Even within specific cultural and social contexts, there can be reasonable disagreement on whether a particular memory is appropriate connected to relevant past events, or is sufficiently similar to earlier experience. This is because of this variability in the intuitions and judgements we have learned to apply, about what counts as a memory. In this domain, there is room for informed judgement or decision. In relation to distinctions between genuine memory and other different cognitive phenomena, people have learned to apply norms or standards. These norms or standards may vary somewhat and have grey areas at their edges, and they can be difficult to access and render explicit. Through enculturation in development, we come to apply standards, norms, or conditions, both in assessing others' memory claims, and importantly also in deploying our own cognitive and metacognitive capacities so as to discriminate, interpret, and communicate our cognitive processes as memories or as something else (Craver 2020, p. 267; Fivush 2019; Jablonka 2017; Sutton 2020b).

The internal neurocognitive processes on which causal theorists – including ourselves – have typically focussed thus do not, we suggest, exhaust the relevant fields of evidence. Getting more facts about either the general operation of such mechanisms, or specific computational processes in a single case, will not always settle the issues of which causal connections matter, and what similarity is required over time. But this does not mean that there is no real distinction between remembering and (for example) imagining: assessments or applications of causal criteria in social and psychological practice are not random or subjective, but are intersubjectively constrained, by both descriptive 'external' or situational facts and by (changing) contexts. The involvement of pragmatic or social factors in assessing memory claims does not render such assessments relative. The existence of uncertain cases, where it's not clear whether the causal connections between past and present are appropriate, or whether there is sufficient similarity between what was experienced and what is remembered, should not encourage us to give up: given the complexity and variability of both neurocognitive and socio-cognitive processes, we should expect to be challenged in our assessments of particular cases.

Future causal theorists can expand their scope to attend to such factors, even if they must thereby give up on any dream of finding watertight criteria to apply across all cases and contexts. But our core business here has concerned the internal, neurocognitive components of these larger systems in which memory processes are embedded. In applying one connectionist account of how distributed traces have content and causal efficacy, we defend the idea that memory has special diachronic aspects, that – as highlighted by the causal theory of memory – memory makes a claim on the past.

Acknowledgements

Many thanks to the editors for helpful comments on an earlier draft, and for their extraordinary efforts in promoting the philosophy of memory. We are also grateful to Max Deutscher, Jon Opie, and Carl Windhorst.

Notes

[1] Some connectionists, seeking decisive moves away from classical cognitivism, came to reject individualism and develop '4E' alternatives. But they did not thereby abandon connectionism. Rather, for Clark (1997; 2013, pp. 17–19), Rowlands (1999, 2010, pp. 41–52), and Sutton (2009), 'connectionism's alternative accounts of cognitive neurodynamics, though already enough to confirm that remembering is a constructive, furiously active process, had to be *supplemented* both by more direct critique of the individualism of classical cognitivism, and by stronger, more integrative, and practice-oriented views of the situated or distributed mind' (Sutton 2015, p. 414, emphasis added). As confirmed by the fact that the current authors hold opposing views about individualism, connectionism is compatible with both individualism and anti-individualism: questions about mental representation are orthogonal to questions about the location of cognition (Sutton 2015). There is no tension in combining connectionism and the extended mind, as Hutto (Chapter 3, this volume) claims, because on the natural and influential 'second-wave' view, external representations *complement* and 'need not mimic or replicate the formats, dynamics, or functions' of traces in the brain (Sutton 2010, p. 194).

[2] Notably, a longstanding computational concern that superposition might lead to catastrophic interference – where new learning overwrites existing memories – led to theories of complementary learning systems in the mainstream cognitive neuroscience of episodic memory (McClelland, McNaughton, & O'Reilly 1995). Recent innovations in deep learning are driving new developments on this front, some focussed more at systems level, others on different ways of modifying connections within individual networks (Hasselmo 2017; Kirkpatrick et al. 2017; Kumaran, Hassabis, & McClelland 2016; McClelland, McNaughton, & Lampinen 2020; Shea 2022).

[3] Localist alternatives to distributed representation do survive in the neuroscience of memory. While we don't have space to discuss current theories that posit discrete and stable engrams in (for example) rats' fear memory, we note that what arguably remain dominant views still acknowledge substantial instability or 'representational drift' across distributed representations (Rule, O'Leary, & Harvey 2019).

[4] Specifically, over periods of 10–20 seconds, in response to a question or cue, distinctive activation trajectories mark out explicit search and retrieval processes (involving the hippocampus), identification and metamemory monitoring processes (involving prefrontal and parietal cortex), sensory processes as the remembered scene is maintained and developed (including, in this case, a range of sensory cortical areas with visual cortex and precuneus likely dominant), and emotional dynamics (involving complex interacting networks across amygdala, somatosensory cortex, and more) (Rubin 2006, 2019).

[5] The notions of distributed traces and superpositional storage are not uniquely tied to modern connectionist models, but operate at an abstract level. So we can identify them in radically different scientific contexts, such as the neurophilosophy of fleeting 'animal spirits' which animated Descartes' rich account of corporeal memory (Sutton 2000), and which elicited widespread early modern criticism for depicting memory and mind as sites of 'a great deal of preposterous confusion' (Henry More, in Sutton 1998, p. 129). Critics of constructive memory were often horrified at those who,

like Descartes, seemed to be recommending an 'assimilation of imagination and memory' (Foti 1986, p. 636; Sutton 1998, pp. 50–113).

[6] Some non-standard views retain causal connections by way of traces, but suggest that these are 'contentless' traces (Werning 2020; see Michaelian & Sant'Anna 2021 and Hutto, this volume, for discussion). Though our view of traces is highly liberalised relative to some, we reject this final step. Our response below to those who deny traces altogether applies equally to 'contentless' conceptions of traces. For relevant prior discussions, see Matthen (2010); Vosgerau (2010).

[7] For Perrin, connectionism is 'discarding the very notion of a representational trace at the neural level': there are only weighted connections between neural nodes (Perrin 2018, p. 40). For Hutto, episodic memory 'involves and depends upon' not transmitted content, but only 'the re-instantiation of specific neural patterns' (Chapter 3, this volume, p. xx).

[8] The weight space of a network is a Euclidean vector space in which each of the network's connection strengths (corresponding to the strengths of the synapses between the neurons in a real neural network) is represented as the position along a distinct coordinate axis. The dimensionality of this space corresponds to the number of connections in the network. We can picture training a network as a trajectory through weight space, and different final positions in the space as alternative ways of successfully dealing with the task demands.

[9] The activation space of a network is a Euclidean vector space in which the level of activity of each of the network's processing units (corresponding to individual neurons in a real neural network) is represented as the position along a distinct coordinate axis. The dimensionality of this space corresponds to the number of processing units in the network. A network responds to an input by stabilising at a particular point in this activation space.

[10] One system *structurally resembles* another when the *physical* relations among the objects that comprise the first preserve some aspects of the relational organisation of the objects that comprise the second (O'Brien & Opie 2004). Activation space is a *mathematical* space used by theorists to portray the set of activation patterns a network generates over its hidden layer. Activation patterns themselves are physical objects (patterns of neural firing, if realised in a brain), and thus distance relations in activation space codify *physical* relations among activation states. The set of activation patterns generated across any trained-up connectionist network constitutes a system of representing vehicles whose physical relations sustain a structural relation with respect to the task domain over which the network has been trained.

[11] This structural resemblance story is not intended as a general theory of mental representation: such resemblance relations are insufficient to ground representation on their own. Instead, structural resemblance is a theory of *content determination*, which in turn plugs into a broader account of mental representation. Representation is a *triadic* relation implicating *representing vehicles*, *represented objects*, and *interpretations* within a cognitive system (O'Brien 2015; von Eckardt 1993). Content determination concerns the relationship between representing objects and representing vehicles, such that the latter are capable of disposing cognitive systems to behave appropriately towards the former (O'Brien & Opie 2004).

References

Andonovski, N. (2020). Singularism about episodic memory. *Review of Philosophy and Psychology*, *11*, 335–365.

Andonovski, N. (2021). Memory as triage: Facing up to the hard question of memory. *Review of Philosophy and Psychology*, *12*(2), 227–256.

Bernecker, S. (2010). *Memory: A Philosophical Study*. Oxford University Press.

Buckner, C., & Garson, J. (2019). Connectionism. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/fall2019/entries/connectionism/.

Campbell, J. (1997). The structure of time in autobiographical memory. *European Journal of Philosophy*, *5*(2), 105–118.

Campbell, S. (2004). Models of mind and memory activities. In P. DesAutels & M. U. Walker (Eds.), *Moral Psychology: Feminist Ethics and Social Theory* (pp. 119–137). Rowman & Littlefield.

Campbell, S. (2006). Our faithfulness to the past: Reconstructing memory value. *Philosophical Psychology*, *19*(3), 361–380.

Churchland, P. M. (1995) *The Engine of Reason, the Seat of the Soul*. MIT Press.

Churchland, P.M. (2012). *Plato's Camera*. MIT Press.

Clark, A. (1989). *Microcognition*. MIT Press.

Clark, A. (1993). *Associative Engines: Connectionism, Concepts, and Representational Change*. MIT Press.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. MIT Press.

Craver, C. F. (2020). Remembering: Epistemic and empirical. *Review of Philosophy and Psychology*, *11*(2), 261–281.

Damasio, A. (1995). *Descartes' Error: Emotion, Reason, and the Human Brain*. Picador.

Daselaar, S. M., Rice, H. J., Greenberg, D. L., Cabeza, R., LaBar, K. S., & Rubin, D. C. (2008). The spatiotemporal dynamics of autobiographical memory: Neural correlates of recall, emotional intensity, and reliving. *Cerebral Cortex*, *18*(1), 217–229.

De Brigard, F. (2014a). The nature of memory traces. *Philosophy Compass*, *9*(6), 402–414.

De Brigard, F. (2014b). The anatomy of amnesia. *Scientific American Mind*, *25*(3), 39–43.

De Brigard, F. (2014c). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese*, *191*(2), 155–185.

Debus, D. (2010). Accounting for epistemic relevance: A new problem for the causal theory of memory. *American Philosophical Quarterly*, *47*(1), 17–29.

Debus, D. (2017). Memory causation. In S. Bernecker & K. Michaelian (Eds.), *The Routledge Handbook of Philosophy of Memory* (pp. 63–75). Routledge.

Deutscher, M. (1989). Remembering 'remembering'. In J. Heil (Ed.), *Cause, Mind, and Reality* (pp. 53–72). Springer.

Deutscher, M. (2009). In response. *Crossroads: An Interdisciplinary Journal for the Study of History, Philosophy, Religion, and Classics* *4*(1), 92–98.

Elman, J. (1993). Learning and development in neural networks: The importance of starting small. *Cognition 48*, 71–99.

Fivush, R. (2019). *Family Narratives and the Development of an Autobiographical Self: Social and Cultural Perspectives on Autobiographical Memory*. Routledge.

Fodor, J. A. (1975) *The Language of Thought*. MIT Press.

Foti, V. (1986). The Cartesian imagination. *Philosophy and Phenomenological Research*, *46*, 631–642.

Goldie, P. (2012). *The Mess Inside: Narrative, Emotion, and the Mind*. Oxford University Press.

Greenberg, D. L., & Rubin, D. C. (2003). The neuropsychology of autobiographical memory. *Cortex*, *39* (4–5), 687–728.

Harris, C. B., O'Connor, A. R., & Sutton, J. (2015). Cue generation and memory construction in direct and generative autobiographical memory retrieval. *Consciousness and Cognition*, *33*, 204–216.

Hasselmo, M. E. (2017). Avoiding catastrophic forgetting. *Trends in Cognitive Sciences*, *21*(6), 407–408.

Haugeland, J. (1991). Representational genera. In W. Ramsey, S. P. Stich, & D. E. Rumelhart (Eds.), *Philosophy and Connectionist Theory* (pp. 61–78). Erlbaum.

Haybron, D. M. (2000). The causal and explanatory role of information stored in connectionist networks. *Minds and Machines*, *10*(3), 361–380.

Hutto, D. D., & Peeters, A. (2018). The roots of remembering: Radically enactive recollecting. In K. Michaelian, D. Debus, & D. Perrin (Eds.). *New Directions in the Philosophy of Memory* (pp. 97–118). Routledge.

Ismael, J. (2016). *How Physics Makes Us Free*. Oxford University Press.

Jablonka, E. (2017). Collective narratives, false memories, and the origins of autobiographical memory. *Biology & Philosophy*, *32*(6), 839–853.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., … & Hassabis, D. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*(13), 3521–3526.

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, *20*(7), 512–534.

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573–616.

Laakso, A., & Cottrell, G. (2000) Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology*, *13*, 47–76.

Mac Cumhaill, C. (2020). Still life, a mirror: phasic memory and re-encounters with artworks. *Review of Philosophy and Psychology*, *11*, 423–446,

Martin, C. B., & Deutscher, M. (1966). Remembering. *Philosophical Review*, *75*(2), 161–196.

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

McClelland, J. L. (1995). Constructive memory and memory distortions: A parallel-distributed processing approach. In D. L. Schacter (Ed.), *Memory Distortions: How Minds, Brains, and Societies Reconstruct the Past* (pp. 69–90). Harvard University Press.

McClelland, J. L., McNaughton, B. L., & Lampinen, A. K. (2020). Integration of new information in memory: New insights from a complementary learning systems perspective. *Philosophical Transactions of the Royal Society B*, *375*(1799), 20190637.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419.

McClelland, J. L., & Rumelhart, D. E. (1986). A distributed model of human learning and memory. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 2, pp. 170–215). MIT Press.

Michaelian, K. (2016). *Mental Time Travel: Episodic Memory and Our Knowledge of the Personal Past*. MIT Press.

Michaelian, K., & Robins, S. (2018). Beyond the causal theory? Fifty years after Martin and Deutscher. In K. Michaelian, D. Debus, & D. Perrin (Eds.). *New Directions in the Philosophy of Memory* (pp. 13–32). Routledge.

Michaelian, K., & Sant'Anna, A. (2021). Memory without content? Radical enactivism and (post) causal theories of memory. *Synthese*, *198*, S307–S335.

Moscovitch, M., Cabeza, R., Winocur, G., & Nadel, L. (2016). Episodic memory and beyond: The hippocampus and neocortex in transformation. *Annual Review of Psychology*, *67*, 105–134.

Naylor, A. (2012). Belief from the past. *European Journal of Philosophy*, *20*(4), 598–620.

Neisser, U. (1982). Memory: What are the important questions? In U. Neisser (Ed.), *Memory Observed: Remembering in Natural Contexts* (pp. 3–19). Worth Publishers.

O'Brien, G. (2015). How does mind matter? Solving the content causation problem. In T. Metzinger & J. M. Windt (Eds.), *Open MIND: 28(T)*. MIND Group.

O'Brien, G., & Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences*, *22*, 127–148.

O'Brien, G., & Opie, J. (2004). Notes toward a structuralist theory of mental representation. In H. Clapin (Ed.), *Representation in Mind* (pp. 1–20). Elsevier.

O'Brien, G., & Opie, J. (2006). How do connectionist networks compute? *Cognitive Processing*, 7(1), 30–41.

O'Brien, G., & Opie, J. (2009). The role of representation in computation. *Cognitive Processing*, *10*(1), 53–62.

O'Brien, G. J. (1991). Is connectionism commonsense? *Philosophical Psychology*, *4*(2), 165–178.

O'Brien, G. J. (1993). The connectionist vindication of folk psychology. In S. M. Christensen & D. R. Turner (Eds.), *Folk Psychology and the Philosophy of Mind* (pp. 368–387). Erlbaum.

O'Brien, G. J. (1998). Being there: Putting philosopher, researcher and student together again (review of Clark, *Being There*). *Metascience (new series)*, 7, 78–83.

Perrin, D. (2018). A case for procedural causality in episodic recollection. In K. Michaelian, D. Debus, & D. Perrin (Eds.). *New Directions in the Philosophy of Memory* (pp. 33–51). Routledge.

Ramsey, W. (1997). Do connectionist representations earn their explanatory keep? *Mind & Language*, *12*(1), 34–66.

Ramsey, W., Stich, S., & Garon, J. (1991). Connectionism, eliminativism, and the future of folk psychology. In J. Greenwood (Ed.), *The Future of Folk Psychology* (pp. 93–119). Cambridge University Press.

Robins, S. (2016). Representing the past: Memory traces and the causal theory of memory. *Philosophical Studies*, *173*(11), 2993–3013.

Rowlands, M. (1999). *The Body in Mind: Understanding Cognitive Processes*. Cambridge University Press.

Rowlands, M. (2010). *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. MIT Press.

Rubin, D. C. (2006). The basic-systems model of episodic memory. *Perspectives on Psychological Science*, *1*(4), 277–311.

Rubin, D. C. (2019). Placing autobiographical memory in a general memory organization. In Mace, J. (Ed.). *The Organization and Structure of Autobiographical Memory* (pp. 6–27). Oxford University Press.

Rule, M. E., O'Leary, T., & Harvey, C. D. (2019). Causes and consequences of representational drift. *Current Opinion in Neurobiology*, *58*, 141–147.

Schechtman, M. (1994). The truth about memory. *Philosophical Psychology*, 7(1), 3–18.

Selwood, A., Harris, C. B., Barnier, A. J., & Sutton, J. (2020). Effects of collaboration on the qualities of autobiographical recall in strangers, friends, and siblings: both remembering partner and communication processes matter. *Memory*, *28*(3), 399–416.

Shea, N. (2007). Content and its vehicles in connectionist systems. *Mind & Language*, *22*(3), 246–269.

Shea, N. (2022). Moving beyond content-specific computation in artificial neural networks. *Mind & Language*. https://doi.org/10.1111/mila.12387

Sutton, J. (1998). *Philosophy and Memory Traces: Descartes to Connectionism*. Cambridge University Press.

Sutton, J. (2000). The body and the brain. In Gaukroger, S., Schuster, J., & Sutton, J. (Eds.), *Descartes' Natural Philosophy* (pp. 697–722). Routledge.

Sutton, J. (2003). Truth in memory: The humanities and the cognitive sciences. In McCalman, I., & McGrath, A. (Eds.), *Proof and Truth: The Humanist As Expert* (pp. 145–163). Australian Academy of the Humanities.

Sutton, J. (2004). Representation, levels, and context in integrational linguistics and distributed cognition. *Language Sciences*, *26*(6), 503–524.

Sutton, J. (2009). Remembering. In Aydede, M., & Robbins, P. (Eds.), *The Cambridge Handbook of Situated Cognition* (pp. 217–235). Cambridge University Press.

Sutton, J. (2010). Exograms and interdisciplinarity: History, the extended mind, and the civilizing process. In Menary, R. (Ed.), *The Extended Mind* (pp. 189–225). MIT Press.

Sutton, J. (2015). Remembering as public practice: Wittgenstein, memory, and distributed cognitive ecologies. In Moyal-Sharrock, D., Coliva, A., & Munz, V. (Eds.), *Mind, Language, and Action: Proceedings of the 36th International Wittgenstein Symposium* (pp. 409–443). Walter de Gruyter.

Sutton, J. (2020a). Movements, memory, and mixture: Aristotle, confusion, and the historicity of memory. In Fink, J. L., & Mousavian, S. (Eds.), *The Internal Senses in the Aristotelian Tradition* (pp. 137–155). Springer.

Sutton, J. (2020b). Personal memory, the scaffolded mind, and cognitive change in the Neolithic. In Hodder, I (Ed.), *Consciousness, Creativity and Self at the Dawn of Settled Life* (pp. 209–229). Cambridge University Press.

Sutton, J., & Windhorst, C. (2009). Extended and constructive remembering: Two notes on Martin and Deutscher. *Crossroads: An Interdisciplinary Journal for the Study of History, Philosophy Religion, and Classics*, *4*(1), 79–91.

Temler, M., Barnier, A. J., Sutton, J., & McIlwain, D.J. (2020). Contamination or natural variation? A comparison of contradictions from suggested contagion and intrinsic variation in repeated autobiographical accounts. *Journal of Applied Research in Memory and Cognition*, *9*(1), 108–117.

Tulving, E. (2007). Coding and representation: Searching for a home in the brain. In Roediger, H. L., Dudai, Y. E., & Fitzpatrick, S. M. (Eds.). *Science of Memory: Concepts* (pp. 65–68). Oxford University Press.

van Gelder, T. (1991). What is the 'D' in 'PDP'? A survey of the concept of distribution. In Ramsey, W., Stich, S. P., & Rumelhart, D. E. (Eds.), *Philosophy and Connectionist Theory* (pp. 33–59). Erlbaum.

von Eckardt, B. (1993). *What Is Cognitive Science?* MIT Press.

Vosgerau, G. (2010). Memory and content. *Consciousness & Cognition*, *19*, 838–846.

Werning, M. (2020). Predicting the past from minimal traces: Episodic memory and its distinction from imagination and preservation. *Review of Philosophy and Psychology*, *11*(2), 301–333.