Forthcoming in *European Journal for Philosophy of Science*

Autonoesis and the Galilean science of memory: Explanation, idealization, and the role of crucial data

Nikola Andonovski
Centre for Philosophy of Memory, Université Grenoble Alpes

Abstract:

The Galilean explanatory style is characterized by the search for the underlying structure of phenomena, the positing of "deep" explanatory principles, and a view of the relation between theory and data, on which the search for "crucial data" is of primary importance. In this paper, I trace the dynamics of adopting the Galilean style, focusing on the science of episodic memory. I argue that memory systems, such as episodic and semantic memory, were posited as underlying competences producing the observable phenomena of memory. Considered in idealized isolation from other systems, episodic memory was taken to underlay the ability of individuals to remember events from their personal past. Yet, in reality, memory systems regularly interact, standing in many-to-many relations to actual memory tasks and experiences. Upon this backdrop, I explore a puzzle about the increasing prominence of the notion of autonoetic consciousness in Tulving's theory of episodic memory. I argue that, contrary to widespread belief, the prominence is not best explained by the purported essential link between autonoetic consciousness and episodic memory. Rather, it is explained by the fact that autonoetic consciousness, hypothesized to uniquely accompany episodic retrieval, was considered a source of crucial data, predictable only from theories positing a functionally distinct episodic memory system. However, with the emergence of a new generation of theories, positing wider memory systems for remembering and imagination, the question of the relation between episodic memory and autonoetic consciousness has been reopened. This creates a pressing need for de-idealization, triggering a new search for crucial data.

> Remembering past events is a universally familiar experience.
>
> Tulving (1983, p.1)

> Much of science begins as exploration of common sense, and much of science, if successful, ends if not in rejecting it then at least going far beyond it. The science of memory, although still in its formative years, is no exception to the general rule.
>
> Tulving (2001a, p.1505)

1. Introduction

The core explananda of the memory sciences are the abilities of organisms to acquire, retain, and retrieve information for productive use. The behaviors in which these abilities are manifested are many and diverse: remembering events, facts or objects, but also imagining the possible past or future, and navigating

complex physical and social environments. The guiding idea that has oriented much memory research, in both humans and animals, is the positing of *multiple memory systems* underlying this cognitive and behavioral variety (Squire & Zola-Morgan 1991; Schacter & Tulving 1994; White & McDonald 2002). Memory systems are taken to be functionally integrated structures with clusters of distinguishing properties, including dedicated principles of operation, kinds of representation, and neuroanatomical substrates. Theorists appeal to the activity of systems with such properties to account for core features of the phenomena of interest. The computational properties of hippocampal and striatal systems, for example, are frequently taken to underpin the abilities of organisms to remember unique experiences and form habits, respectively (e.g., McDonald & Hong 2013; Chersi & Burgess 2015). Indeed, as Ferbinteanu (2019) notes, phenomenologically different types of memories are often characterized as resulting *from* the independent operation of specialized memory systems.

This explanatory attitude involves a significant element of idealization. As theorists readily admit, activities like remembering or imagining typically arise from the coordinated interactions of a number of distinct memory systems and mechanisms (Schacter & Tulving 1994; Poldrack & Rodriguez 2004; White et al. 2013). The idealization is nevertheless justified by the pursuit of explanatory depth. Abstracting from the "messy" details of systems interactions, which may vary with context and experimental design, memory researchers can advance bolder hypotheses—moving from generalizations of empirical data to formulation of unifying explanatory principles. Such principles can identify basic computational and structural properties that underlie the variety of memory activities (Anderson 1983; Eichenbaum 2000). They can also articulate general reasons for the existence of multiple, functionally specialized and possibly evolutionarily selected, memory systems (Sherry & Schacter 1987; McClelland et al. 1995).

Practices of idealization extend also to the attitudes researchers take to the relation between theory and data. Given the widespread interaction between memory systems, a lot of the available data—whether from simple introspection, careful phenomenology or sophisticated experimental procedures—will be messy, relating to the tested hypotheses only in an indirect and imperfect way. In such circumstances, complete coverage of empirical data is not a reasonable expectation. Theories of memory—especially bold ones, offering potentially far-reaching insights—are given more empirical leeway and allowed to survive seemingly disconfirming evidence. The twin commitments to idealization and "epistemological tolerance" (Botha 1982) of this kind constitute an explanatory style arguably widespread in the natural sciences. Baptized "the Galilean style" (Weinberg 1976), its importance and applicability to the sciences of the mind has been most consistently, if not always most carefully, defended by Noam Chomsky (1978, 1980, 2002).

The tacit adoption of this Galilean style by memory theorists, while opening us space for theory building, has led to a host of important problems. With an almost exclusive focus on the properties and

independent operations of specialized memory systems, the principles and modes of their interaction have remained poorly understood (Kim & Baxter 2001; McDonald & Hong 2013; Goodroe et al. 2018). More importantly, theorists in the grip of an idealization have often lost track of the importance of such interactions for the production of memory phenomena. Anxious to close the gap between "surface" properties of the phenomena and increasingly abstract theories of them, they have often done so hastily and without a proper appreciation of the complexity of such task. In the pursuit of explanatory depth, de-idealization may warrant as much attention as idealization. Relatedly, interactions between systems haven't always been taken into account when considering the posits and predictions of rival theories. This has resulted in a lack of clarity about the evidential status of theories and the nature of potentially (dis)confirming evidence.

In this paper, I will trace the dynamics of adopting the Galilean style, focusing on the science of episodic memory. In section 2, I introduce the core commitments of the style, illustrating the roles they played in the development of the science. I argue that memory systems, such as episodic and semantic memory, were posited as underlying competences producing the observable phenomena of memory. Believed to operate in a proprietary way, the episodic memory system was theorized about in idealized isolation from its interactions with other systems, with the goal of gaining explanatory traction and theoretical insight. Under the key idealization, episodic memory was taken to underlay the ability of individuals to remember events from their personal past. Yet, the significance and prevalence of systems interactions was clearly recognized. Memory systems regularly interact to produce the phenomena of memory, standing in many-to-many relations to actual memory tasks. On this view, counterintuitively, the episodic memory system is neither sufficient nor necessary for the performance of nominally episodic tasks. In section 3, I introduce the notion of autonoetic consciousness, tracing a puzzle about its increasing prominence in Tulving's theory of episodic memory. I argue that, contrary to widespread belief, the prominence is not best explained by the purported essential link between autonoetic consciousness and episodic memory. Rather, it is explained by the fact that autonoetic consciousness, hypothesized to uniquely accompany episodic retrieval, was considered a source of crucial data, predictable only from theories positing a functionally distinct episodic memory system. In short, data about autonoesis was taken to point to the existence of episodic memory. In section 4, I examine the ways in which empirical developments in the 21st century led to a new generation of Galilean theories, positing a wider cognitive system underlying both remembering and imagination. In this theoretical landscape, the question of the relation between episodic memory and autonoetic consciousness has been reopened. I argue that that investigation of this relation requires better understanding of memory systems interactions. This creates a pressing need for de-idealization, triggering a new search for crucial data. I end by examining the role of commonsense notions and categories in scientific theories of memory.

2. The Galilean Science of Memory

In 2.1., I introduce the Galilean explanatory style, identifying its core commitments and characterizing the role Galilean idealization plays in scientific theorizing. In 2.2., I illustrate how the style was employed in the science of memory. I argue that memory systems were posited as underlying competences, with proprietary principles of operation, producing the observable phenomena of memory. While these systems interact virtually all the time, they can be investigated in idealized isolation in order to gain explanatory traction on a very complex phenomenon.

2.1. The Galilean Explanatory Style

Scientific idealization and the relationship between theory and data are central concerns of the philosophy of science and have been investigated from a variety of perspectives, by theorists with diverse explanatory aims and tools (Cartwright 1983; Wimsatt 2007; Potochnik 2017). The close connection between the two is nevertheless brought into sharp focus in Chomsky's methodological remarks, whose primary aim is to underscore the necessity of developing a natural science of psychology. This task requires the adoption of an explanatory style exemplified by the natural sciences and inaugurated—or so the story goes—by Galileo and his contemporaries:

> The great success of the modern natural sciences can be attributed to the pursuit of explanatory depth which is very frequently taken to outweigh empirical inadequacies. This is the real intellectual revolution of the seventeenth century (Chomsky 1978, p. 10).

When studying the human mind and behavior, there is no reason to abandon this general attitude. "Any serious approach to such topics will attempt, with whatever success, to adopt 'the Galilean style'" (Chomsky 1980, p. 219). A serious approach—of the kind the memory sciences certainly purport to embody—will eschew any "methodological dualism" and seek methodological, if not theoretical, unification with the rest of the natural sciences (Chomsky 2000).

The nature of the Galilean style, and the costs and benefits of adopting it, have received significant attention in the literature. Historical analyses have explored the role it played in 17th century science, cataloguing its features and revealing them at work in prominent episodes of the period (Wisan 1978; Feyerabend 1979). Such work has allowed theorists to assess whether the Galilean style can be defined in terms of a set of distinctive attributes and to examine its applicability to the sciences of the mind (Botha 1982; McMullin 1985).[1] In a recent contribution, Allott, Lohndal, and Rey (2021) provide a particularly

[1] It is worth noting that some authors are skeptical that the Galilean style can be characterized uncontroversially in terms of a small set of distinctive attributes, *if* such characterization is expected to have the required measure of historical credibility (e.g., Botha 1982). This is certainly an important exegetical question. Here, however, I am less concerned with establishing historical

clear characterization, illustrating the ways in which the style is manifested in contemporary linguistic research. Their analysis is useful for our purposes, which have less to do with establishing historical authenticity and more with illustrating the prominence of a set of explanatory attitudes—and the dynamics associated with their adoption—in the memory sciences.

Allott et al. (2021, pp. 518-522) isolate three core commitments of the Galilean style: (i) a conception of science as a quest for underlying explanatory structure; (ii) a distinction between superficial generalizations and deeper explanatory principles; and (iii) a view of the relation between theory and data, on which the search for "crucial data" is of primary importance.

First, scientists aiming to uncover the structures underlying particular phenomena cannot rely solely on the collection of "surface" data, whether through ordinary observation or carefully designed experiments. Most phenomena under scientific investigation are the result of *interaction effects*: they are produced by the interactions of multiple—sometimes, indeed, a great number of—mechanisms and systems (p. 519). This is inarguably the case with the majority of phenomena investigated by the sciences of the mind.[2] With such widespread interaction between mechanisms, scientists seeking to uncover the underlying structure of phenomena are required to idealize. Galilean idealization is characterized by the intentional simplification of a theory or model of such structure with the purpose of gaining explanatory traction (McMullin 1985; Weisberg 2007). It may involve the introduction of distortions or simply the omission of components of the structure so as to focus on the remaining ones—"ignoring some aspects of the world in order to understand others" (Pietroski & Rey 1995, p. 89). Importantly, the omitted components are often known, or at least suspected, to be causally relevant to the production of the target phenomena.[3] A characteristic Galilean strategy involves the study of a mechanism in idealized isolation; e.g., studying the effect of gravity on a body's motion while abstracting away from friction or electric charge. Anchored on such a strategy, Chomsky's program is characterized by the study of a domain-specific cognitive system—a *competence system* or simply *competence*—that underlies human linguistic performance. While this system never works in isolation—all actual phenomena of language use are the result of the interaction of multiple systems—it can be theorized about independently because it is taken to operate in a proprietary way (Chomsky 1980; Allott & Smith 2021).

---

authenticity, and more with illustrating the way in which a cluster of explanatory attitudes is exemplified in the memory sciences. For the purposes of the paper, we may take these attitudes to characterize what Botha calls "a *lax* Galilean style", a mode of inquiry that represents an important methodological tool in the (psychological) sciences, even if it is only loosely connected to Galileo.

[2] Think, e.g., of the variety of systems underlying a simple mental activity like watching the movement of a flock of seagulls. These include shape and color processing systems, motion and object tracking systems, various auditory systems etc. Examples of this kind are easy to generate. See also note 5.

[3] Hence, Galilean idealization is importantly different from what Weisberg (2007) calls "minimal idealization": the practice of constructing theories or models that include only the causal factors that "make a difference" to the occurrence of a target phenomenon. Galilean and minimal idealization differ both in their representational ideals and in the way they are typically justified. For discussion, see Weisberg (2007, pp. 640-649). Thanks to an anonymous referee for prompting me to clarify this point.

Second, and relatedly, Galilean idealization is driven not by the need to generalize over available data but rather by the pursuit of "deep" explanatory principles. While generalizations are useful—they make clearer the phenomena that warrant explanation, paving the way for their potential "stabilization" (Hacking 1983; Feest 2011)—they don't offer insight into their underlying structure. Properly accounting for such structure requires positing explanatory principles that "unify a variety of [empirical] generalizations and ground them in a system that has a certain degree of deductive structure" (Chomsky 1978, p. 16). These principles are often abstract and inferentially removed from the data on which they bear, characterizing properties of mechanistic structure in formal or mathematical terms (Botha 1982; Allott et al. 2021).[4] The formulation of deep unifying principles is as characteristic of Galilean explanations in the special sciences as it is of those in basic physics (Cartwright 1983). Chomsky's formalized grammars, considered independently of language use, provide a notable case in point.[5]

Finally, and perhaps most controversially, idealized Galilean theories should not aim for complete, or even substantial, coverage of empirical data. With much of the data messy, and resulting from intricate interaction effects, even our best theories will fall radically short of full empirical adequacy.[6] In such circumstances, embracing epistemological tolerance is the best course of action. As Donald Hebb put it at the dawn of modern cognitive science, we should try to devise theories that account for aspects of phenomena "we might have some chance of accounting for and not worry if the theorising appears to be inadequate in some respects to cover all known features of the system" (in Delafresnaye 1954, p. 499). These theories will be judged on their explanatory depth *and* their capacity to explain "crucial data" (Allott et al. 2021, pp. 520-522). Crucial data are data that pertain to important features of the target phenomenon—*as theorized* under the relevant idealization—and are predictable from the candidate theory but not from rival theories. As we have seen, the Galilean scientist aims to formulate explanatory principles that provide insight into the functioning of a system, often considered in isolation. These principles, and the nature of the rival explanations, then determine which data is selected as crucial. Crucial data thus adjudicate between competing theories of phenomena. Theories in generative linguistics, to provide an illustrative example, do not aim to account for all data concerning language use but only for data predictable from their but not from rival theories (Chomsky 1980; Rey 2020).

---

[4] In his original characterization, Weinberg (1976) emphasized the pursuit of *mathematical* models as a central property of the Galilean style (see also Koyré 1943). While Chomsky seemingly borrows this commitment, it is an open question to what extent formal theories in psychology (need to) have a mathematical structure in the sense familiar from physics (see Botha 1982, pp. 9-11). For this reason, I do not include mathematization as a distinctive property of the Galilean style.

[5] On Chomsky's view, the actual *use* of language, affected as it is by myriad performance factors, is simply too complex to cover by a single linguistic theory (see, e.g., 2000, Ch. 2). "Like the trajectories of leaves or automobiles, [language use] is a massive interaction effect" (Allott et al. 2021, p. 519).

[6] Chomsky takes this idea even further: "If someone were to descend from heaven with the absolute truth about language or some other cognitive faculty, this theory would doubtless be confronted at once with all sorts of problems and 'counter-examples', if only because we do not yet understand the natural bounds of these particular faculties and because partially understood data are so easily misconstrued" (1980, p. 10). See 4.2. for a recent echo of this sentiment, expressed by a memory scientist.

The focus on crucial data does not entail that unexplained data are simply disregarded. Rather, they are only temporarily set aside as not of primary importance. Indeed, since the main justification for Galilean idealization is pragmatic—simplifying a phenomenon to gain explanatory traction on it—the maturation of a theory should come with systematic *de-idealization* (McMullin 1985; Weisberg 2007). Once the original idealization has afforded understanding of the relevant explanatory principles, simplifying assumptions can be gradually eliminated to correct for the "deviations" from the truth. For Galilean theories of mental systems, this will often involve examination of the ways in which the target system interacts with other systems to produce introspectively and behaviorally identifiable phenomena. As such a theory develops, it will aim to shed light on increasingly wider class of such phenomena, typically in combination with other theories. For these reasons, idealized theories can serve as the bases for continuing research programs (McMullin 1985).

2.2. The Galilean Science of (Declarative) Memory

The modern science of memory began, unsurprisingly, with the accumulation of surface data. After the pioneering work of Broca (1861), Ebbinghaus (1885), and Thorndike (1898), data collected through introspection were supplemented with experimental and neuropsychological data of a kind that will come to dominate the memory sciences.[7] As experimental procedures matured, it gradually became evident that the phenomena of memory are diverse and of varying robustness. Memory, as J.R. Anderson (2007, p. 122) observed, "is involved in almost everything we do, [even if] most of the time we think of ourselves not as remembering". Other animals, it turned out, also manifest myriad memory-based behaviors (McDonald & White 1993; Eichenbaum 1994). Cutting through the clutter of empirical data required Galilean boldness, which the *multiple memory systems* (MMS) approach, emerging as a major research framework in the 1960s and 1970s, was to supply. The animating idea of the approach was nicely summarized in a programmatic paper by Schacter & Tulving (1994):

> [It is] the idea of multiple memory systems…different *neurocognitive (brain/mind) structures* whose physiological workings produce the *introspectively apprehensible and objectively identifiable consequences* of learning and memory (p. 3, emphasis added).

The pursuit of depth is at the heart of the MMS approach: explaining what is introspectively and experimentally accessible in terms of what is not so accessible, the "hidden" workings of memory systems.

Memory systems, then, are neurocognitive structures involved in the acquisition, retention, and retrieval of information. Making this basic characterization more precise has turned out quite tricky, not the

[7] For a good historical overview of the key developments, see Bower (2000).

least because of the uneasy relationship between cognitive and neural approaches to systems individuation (Anderson 2015; De Brigard 2017; Ferbinteanu 2019). At a minimum, however, memory systems can be considered sets of correlated memory processes (Tulving 1985a; Schacter & Tulving 1994). A memory process is a cognitive/neural operation—such as pattern completion or relational binding—carried out in the performance of a memory task.[8] A process may involve a proprietary kind of representation, with characteristic format and/or content, and may have a unique neural substrate. Importantly, while some memory processes may correspond to introspectively identifiable cognitive events (cf. Cowell et al. 2019), many will not. In fact, it is precisely the positing of formal operations, at some inferential remove from the data of introspection, that best exemplifies the search for explanatory depth. We see it most clearly at work in computational models of core memory processes such as consolidation or structural abstraction (e.g., Benna & Fusi 2016; Whittington et al. 2020). By design, constituent processes of a system frequently interact—and are thus probabilistically correlated—in memory performance: e.g., temporal order encoding facilitates retrieval of object or temporal information from each other (MacDonald et al. 2011). Nevertheless, some processes may be constituents of *multiple* systems (cf. Sherry & Schacter 1987).[9]

Crucially, memory systems are *not* posited by "superficial" generalizations of empirical data. Undeniably useful in the process of discovery, such generalizations are mere descriptive tools:

> [O]ne can think of verbal memory, recognition memory, and olfactory memory as different kinds of memory. Distinctions of this sort can help to describe and organize empirical facts. But *these kinds of purely descriptive forms of memory do not constitute memory systems* (Schacter & Tulving 1994, pp. 11-12, emphasis added).

Memory systems, to borrow the Chomskyan idiom, are not generalizations of *performance*, cataloguing the properties of actual memories produced for different behavioral purposes (e.g., in tasks of verbal, recognition or olfactory memory). Rather, they are underlying *competences* alleged to explain core properties of such performance. For that reason, theories of memory systems should not be characterized as taxonomic proposals, offering organizational schemes for the classification of memor*ies* (*pace* Willingham & Goedert 2001). They are rather explanatory accounts of the cognitive structures that produce the, undeniably diverse, memories.[10]

---

[8] If you are worried about defining memory processes in terms of *memory* tasks, you should be, but probably not to death. See Francken et al. (2022) on the likely inevitability of "circular", and iterative, characterizations of tasks, processes, and mechanisms.
[9] Sherry & Schacter (1987) distinguish between a *strong* view, on which component processes of a system interact only with each other, and a *weak* view, on which any of the components can interact with processes outside the system. Following the authors, I adopt the weak view in this paper. This should be obvious in the discussion that follows.
[10] We should be careful here. It shouldn't be controversial that there has been uncertainty in the literature pertaining to whether MMS proposals are explanatory or taxonomic. I agree with Willingham & Goedert (2001) that some theorists, especially in moments of carelessness, have tried to have their cake and eat it too. Yet, as Willingham & Goedert readily admit, most MMS proposals have been offered *as* explanatory theories. Indeed, the debates concerning the individuation of memory systems reflect

This is the context in which the notion of episodic memory—and to some extent semantic memory—entered the repertoire of the cognitive sciences. On the view, which can safely be considered canonical in the MMS literature, the major divide in long-term memory is between declarative and non-declarative memory. Non-declarative memory systems, distributed widely across the brain, support the implicit acquisition of a variety of perceptual, motor, and cognitive skills (Reber et al. 1996; Squire 2004). Declarative memory systems, in contrast, are involved in the encoding, storage, and explicit recall/recollection of information. They are subserved by a multicomponent network centered on the medial temporal lobes (Cohen & Squire 1980; Eichenbaum & Cohen 2001). Introduced as a "pre-theoretical" hypothesis (Tulving 1972), the distinction between episodic and semantic memory was soon developed into a serious proposal about two—functionally distinct, yet interacting—declarative memory systems (Tulving 1983, 1985b, 2002a).[11] Despite persistent criticism, some of which anchored on issues explored below, the distinction has remained a staple of declarative memory research, at least in humans (Renoult et al. 2019; Ranganath 2022).

Surface data about episodic memory come from both everyday experience and experimental tasks. Remembering past events, Tulving told us in the opening lines of his *Elements* (1983, p. 1), is a universally familiar experience, with a characteristic phenomenology and epistemic profile. When remembering their personal past, subjects feel like they are drawing on previous first-hand experience of the remembered events. (The feeling will later be identified as a core feature of the "autonoetic consciousness" accompanying recollection; see 3.2. & 3.3.). They *re*-experience these in memory, in a quasi-perceptual way, "traveling back into the past in their own minds" (*ibid*). They are thus disposed to assert epistemic authority with respect to the events: claiming to know what happened, and how it happened, to them (1983, Ch. 3). Memories of impersonal facts, in contrast, need not be accompanied by such feelings of past experience, mental time travel, or epistemic privilege. Tulving (1972, 1983) linked the phenomenological data to data from tasks gauging subjects' ability to remember experimentally presented material—paradigmatically, to correctly identify previously studied verbal items and the temporal relations between them (Bower 1970; cf. Tulving & Madigan 1970). In a characteristically bold move, he argued for an underlying competence that accounts for *both* and can potentially unify different empirical generalizations—an episodic memory system.[12]

---

this. That said, there has been some residual confusion about the relation between memory competence and performance, which I hope to delineate in this paper. See the main text below.

[11] For an excellent historical treatment of the development of Tulving's thought about episodic memory, and its relation to semantic memory, see Renoult & Rugg (2020).

[12] It should not be underestimated just how bold this proposal was. Tulving aimed to bring together data from disparate domains and—in the pursuit of depth—posit an underlying structure that has serious potential to unify a variety of generalizations. In an early review of *Elements,* Crowder (1986) noticed this, highlighting Tulving's "radical new focus of episodic memory" (p. 566), a focus which Tulving would later characterize as a "threat to the [then] prevailing order" in the memory sciences (2001b, p. 19).

The proposal is anchored on a key idealization: the episodic memory system underlies the ability of individuals to remember events from their personal past and thus complete relevant ("episodic") memory tasks (Tulving 1983). The distinct properties and operations of the system *explain* the phenomena characterized by the available introspective and behavioral data.[13] Arguing for the functional independence of episodic and semantic memory—a declarative system underlying a person's general "knowledge of the world" (1983, p. 9)[14]—Tulving put forward a list of distinguishing features of the two systems. Among other differences, episodic and semantic memory were taken to differ in their prototypical units of stored information (events vs facts), the coding and organization of this information (temporal vs conceptual), the kind of reference they afforded (autobiographical vs cognitive) as well as the consequences of retrieval (episodic retrieval typically modifies the stored information, while semantic retrieval does not) (1983, pp. 32-57; cf. Tulving 1972, pp. 385-395).These differences were primarily of "diagnostic" value, constituting preliminary evidence for the functional distinctness of episodic and semantic memory (1983, p. 36).[15]

Under the proposed idealization, the two systems—taken to be characterized by proprietary principles of operation—could be theorized about independently. Nevertheless, theorists were warned not to lose track of the significance and prevalence of interactions between them. Episodic and semantic memory are "closely interdependent and interact with one another virtually all the time" (pp. 64-65). Such interactions may be cooperative or competitive and can involve "transfer" of information. This simple point has important consequences for the relationship between memory systems and experimental tasks. While carefully designed tasks may differentially tap underlying systems, *interaction effects abound* (Tulving 1983, Ch. 4, see also Tulving 1991, 2002a; Wheeler et al. 1997). Hence, different memory systems will likely contribute to the performance of even simple memory tasks:

> A participant's performance on a so-called episodic memory task, such as recall or recognition of words encountered in a studied list, depends not only on episodic memory but also on other kinds of memory, such as semantic memory (Wheeler et al. 1997, p. 332).

The independent activity of the episodic system, to put the point differently, will typically not be *sufficient* for a successful performance of a nominally "episodic" task, such as list recall. Perhaps more surprisingly—

[13] At different points in this paper, I adopt the idiom associated with the so-called *ontic* conception of explanation, according to which it is real mind-independent entities—such as neurocognitive systems—that constitute explanations (Salmon 1984). This is primarily for convenience. As far as I am aware, the relevant claims can survive translation to a "representationalist" idiom, associated with the idea that explanations are constituted by explanatory texts of some kind (e.g., sentences, models, diagrams). Tulving, to my knowledge, had no firm view on this issue concerning scientific explanation.

[14] Tulving's conception of semantic memory evolved significantly from 1972, through 1983, to the 2000s. See Renoult & Rugg (2020) for the most important developments.

[15] In the text that follows, I abstract away from some features of Tulving's evolving view of the relation between the episodic and semantic systems; e.g. their position in a class-inclusion hierarchy (1985b) or their process-specific relations, posited by the SPI model (1995). Again, this is mostly for convenience and should not affect the main arguments of the paper.

given that the episodic system was introduced to explain task performance—Tulving and colleagues argued that such activity is also not *necessary*, anchoring their argument on data from amnesic patients:

> It has been shown that even patients with dense amnesia…can nevertheless recall words from studied lists in response to relevant cues [references omitted]. If nonepisodic memory processes are sufficient to allow such patients to perform more or less successfully on what nominally are episodic memory tasks, it means that *the episodic memory system is not necessary for so-called episodic memory tasks…* If so, the same should be true of healthy people (Wheeler et al. 1997, p. 332, emphasis added).[16]

In fact, the relation between systems and tasks is *many-to-many* (Tulving 1991; Schacter & Tulving 1994). While a given memory system is operative in the performance of a variety of tasks, the successful performance of one such task depends—to varying degrees, to be sure—on the activity of multiple systems. Hence, "all tasks are multiply determined" (Tulving 2002a, p. 5). Figure 1 illustrates the many-to-many relation.

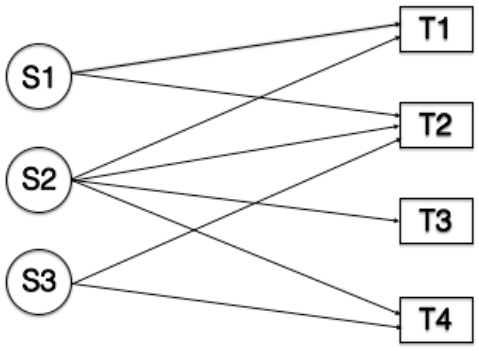

Figure 1. Many-to-Many Relation
The relation between memory systems and tasks is many-to-many. Memory systems causally contribute to the performance of multiple tasks. Moreover, the successful performance of a task typically depends on the activity of multiple systems. Circles represent systems, rectangles represent tasks, and lines connecting them represent relations of causal contribution.

It is difficult to overestimate the importance of this point. As Renoult & Rugg (2020) point out, the idea that memory tasks tap multiple underlying systems—that they are not "process pure"—is widely held and at the basis of some important developments in the sciences of memory.[17] It highlights the general

[16] Admittedly, this point is presented somewhat anachronistically here. In *Elements,* Tulving adopted an "initial" characterization of episodic tasks, anchored on the necessity of the involvement of the episodic memory system (1983, p. 55). As the quote in the main text indicates, empirical results will push Tulving to the rejection of this idea in the 1990s.
[17] For example, the *autobiographical interview*, a widely used memory test, is based on the idea that semantic and episodic elements are regularly mixed in "normal" recall and have to be teased out by special scoring protocols (Levine et al. 2002).

problem of operationalizing cognitive capacities (Francken et al. 2022) and guards against uncritical reliance on data from a single class of experiments (see 3.1.).

At the same time, it provides significant constraints on the use and relevance of *phenomenological data.* The primary reason for this is that particular memories—which, in the context of episodic memory, may be identified with introspectable memory experiences (see, e.g., Martin 2001; Barkasi & Rosen 2020)—will also typically be multiply determined.[18] This, Tulving thought, can be hinted at by reflection on the nature of some prototypical memories—e.g., the memory of a telephone call in which you were informed of an interesting fact. In fact, "*most* experienced and remembered events [will] have factual contents…greatly influenced by semantic memory" (Tulving 1983, p. 55, emphasis added). Episodic information, on the other hand, will often be used in remembering general facts. So, while theorists can talk of episodic and semantic memor*ies*—perhaps as experiences that predominantly engage the episodic or semantic system—they should proceed with caution. Very few, if indeed any, memories are "pure" products of the episodic or semantic system. It may be as difficult to find such memories "as it is to find sodium and chlorine as free elements in nature, although their compound, NaCl, is found in abundance" (*ibid*).[19] This point may in fact generalize, with a variety of systems potentially influencing the structure and content of memory experiences. As a result, we cannot infer the properties of systemic representations *from* the introspected content of memory experiences, at least not without relying on auxiliary assumptions subject to experimental scrutiny (Andonovski 2020; Pan 2022).

The Galilean nature of the MMS approach is clearly exemplified in the pursuit of depth and the decisive move beyond simple empirical generalizations. Memory systems are *explanatory*, not descriptive, posits. They are introduced to shed light into the underlying structure of memory phenomena and to unify a variety of empirical generalizations, from both everyday phenomenology and laboratory experiments. Indeed, this is why the relation between systems and tasks—or between systems and memory experiences, for that matter—is many-to-many. Since MMS theories aim to do more than simply describe features of experience or performance, it is almost inevitable that all tasks will be multiply determined. On such theories, it is possible that the independent activity of a system, posited to explain successful performance of a class of tasks, will end up being neither necessary nor sufficient for it.

[18] Indeed, one can think of particular memories as *solutions* to various cognitive tasks (Andonovski 2021).

[19] An important corollary of this point is that taxonomic proposals intent on classifying *all* memories as, e.g., either episodic or semantic will have a characteristically hard time. Moreover, if they do manage to accomplish this, the resultant taxonomies will likely *not reflect the functioning of the underlying systems*. Tulving (2002a, p. 4) indeed explicitly characterizes the task of unambiguously identifying a particular memory as being either episodic or semantic as "uninteresting", believing it to "lead nowhere". Words of caution for those keen on characterizing the developments in the science of episodic memory as reflecting a search for a good taxonomic criterion of this kind.

In this context, it is not difficult to see why we should be epistemologically tolerant. Some memory phenomena will simply be interaction effects, produced by multiple underlying systems, whose involvement may vary with context and task demands. Theories that aim to gain explanatory leverage by zeroing in on the independent operations of systems cannot be expected to explain all such effects, especially in the early stages of theory building. Hence, they should be given more empirical leeway and not be abandoned in the face of seemingly conflicting evidence. But how much more leeway? And what kinds of data *can* be marshalled as evidence for and against such theories? What, indeed, are the crucial data? In the next sections, I will attempt to cast some light on these difficult questions, exploring the ways in which they have informed the debate about the existence of multiple declarative systems. My focus will be on the notion of *autonoesis,* which, as I aim to illustrate, played a key role in theories of episodic memory. In section 3, I will argue that, for Tulving, data about autonoesis was advanced as crucial data, predictable only from theories of multiple declarative systems. In section 4, I will trace the emergence of a new generation of theories that inherited Tulving's concepts yet developed them in a variety of different ways. I will argue that, contrary to appearances, many of the concerns that motivated Tulving's Galilean theorizing—as well as the problems it encountered—are still with us.

3. Autonoesis and Crucial Data

In 3.1., I discuss the role epistemological tolerance and crucial data play in Galilean theories of memory. I illustrate the difficulty of obtaining such data when dealing with complex, interacting systems with overlapping processes. In 3.2., I introduce the notion of autonoesis, tracing a puzzle about its increasing prominence in Tulving's theory. At first glance, the characterization of autonoesis as an essential, or definitional, feature of episodic memory seems incompatible with the Galilean approach. In 3.3., I connect the discussions, showing that this incompatibility is only apparent. I argue that the prominence of autonoesis is not best explained by its purported essential link to episodic memory but rather in epistemic terms. Hypothesized to uniquely accompany episodic retrieval, autonoesis was considered a source of crucial data, predictable only from theories positing a functionally distinct episodic memory system.

3.1. Declarative Memory: Epistemological Tolerance and the Search for Crucial Data

The primary motivation for positing multiple memory systems stems from the behavioral and phenomenological variety of memory performance. In declarative memory, as we have seen, this is manifested in the seemingly disjoint clusters of features accompanying the remembering of personal events and impersonal facts. It is thus somewhat ironic that such variety can create problematic effects of interaction. If the mapping of memory systems to tasks is many-to-many, and all tasks are multiply determined, then (un)successful performance can be explained in a number of ways—at least in principle.

Consider, for example, subjects who perform averagely on a nominally episodic task, despite significant hippocampal damage (as, e.g., in Robin et al. 2019). This result may be construed as providing some evidence against theories positing a distinct episodic system with hippocampally-supported storage. Yet, it may also be the case that the relevant task can be systematically performed in other, ostensibly "compensatory", ways—e.g., by employing semantic or procedural operations. If this is the case, then successful performance of the task will *not* constitute evidence against episodic impairment. See Figure 2 for illustration of this point.

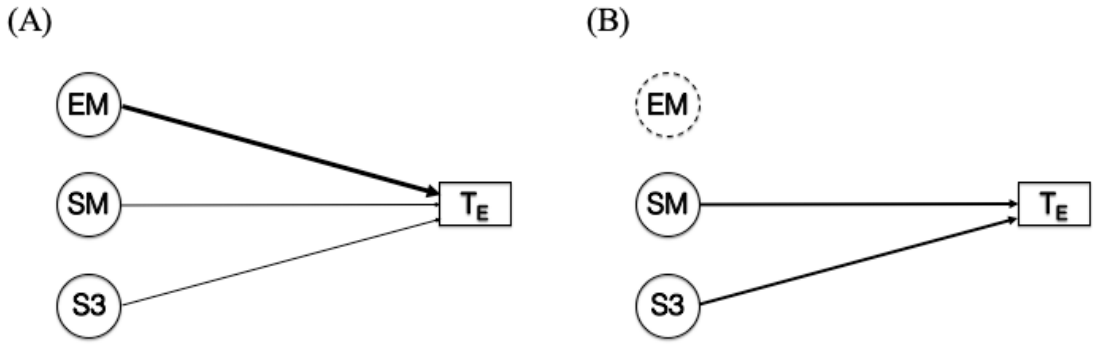

Figure 2. Compensatory Interaction
(A) In regular circumstances, the nominally 'episodic' task T_E taps episodic memory (EM) most strongly, yet also taps semantic memory (SM) and a third system (S3). The thickness of the line represents the strength of the causal contribution. (B) In circumstances in which EM functioning is impaired (indicated by the dashed line), EM doesn't contribute to performance of T_E. Nevertheless, SM and S3 "pick up the slack", with their joint causal contribution sufficient for successful performance of the task. For an account of "competitive" and "cooperative" interactions, see Kim & Baxter (2001).

This is, of course, only an illustration of principle. Not all theoretical possibilities are, or should be, given equal weight or attention. Yet, the problem runs deeper. With our knowledge of systems interactions in its infancy, theorists cannot predict the nature and likelihood of compensatory effects. This may lead to the accumulation of data *seemingly* at odds with theories of memory systems.[20] The embrace of epistemological tolerance in this predicament does not mean that theorists should not take such data seriously. It does mean, however, that they should not abandon bold theories that appear empirically inadequate in some respects; at least not at the first few signs of trouble. It also means that they should, as the theories mature, devise serious alternative hypotheses about the interaction effects that may explain (away) the problematic data. In ideal circumstances, the development of theories will be accompanied by progressively refined approaches to testing the rival explanations. On the ground, however, such refinement

[20] As we will see in 3.3, the accumulation of what Tulving considered only *seemingly* disconfirming evidence—pertaining to the ability of amnesiacs to complete nominally "episodic" tasks—was a major impetus behind the shift of focus to autonoesis.

will often be arduous, with many starts and stops and occasional dead ends (Kuhn 1962; Shadish et al. 2002; Francken et al. 2022).

The Galilean theorist pairs this epistemological tolerance with a focus on crucial data: data that pertain to important properties of the target phenomenon—as theorized under the relevant idealization—and are predictable from the candidate theory but not from rival theories. In our context, the relevant debate concerned the existence of multiple memory systems. Limiting our attention to declarative memory, competing theories disagreed about the existence of a functionally dissociable *episodic* system (Tulving 1983, 1985b; Roediger et al. 1990; Moscovitch 1992; Cabeza & Moscovitch 2013).[21] For Tulving, crucial data in this context was thus data that pointed to the existence of an episodic system with the relevant, theoretically postulated, properties. Sensitive to the methodological issues highlighted above, Tulving and colleagues were quite aware of the difficulty of obtaining data of this kind. Given that memory systems have "fuzzy boundaries, have overlapping constituent processes, and interact with one another in intricate ways" (Schacter & Tulving 1994, p. 18), a theorist cannot solely rely on data from a single class of experimental tasks. Rather, they should look for *converging* data: "dissociations of different kinds, observed with different tasks, in different populations, and using different techniques" (ibid; cf. Tulving 1987; Roediger et al. 1990). These can be manifested in differential performance on classes of tasks alleged to primarily tap the episodic or semantic system (functional dissociations), in uncorrelated performance on "episodic" and "semantic" tasks (stochastic independence), and in selective impairments in relevant populations (neuropsychological dissociations). As the sciences of memory mature, these will be supplemented with neural data as well as integrative proposals about the functional problems memory systems evolved to solve (Sherry & Schacter 1987; Schacter 2022).

This is the context in which the notion of autonoesis was introduced to Tulving's evolving theory of episodic memory. In the next section, I briefly examine its introduction and development, tracing a puzzle about its role and increasing prominence.[22]

3.2. The Curious Case of Autonoesis: Between Competence and Performance

The notion of *autonoetic consciousness*—autonoesis, for short—was present, albeit in an innominate and less conspicuous form, in Tulving's early accounts of episodic memory. In 1972, Tulving characterized the recollection of personal events as intimately linked with a kind of immediate experience of

[21] In many ways, this paper's focus on declarative memory is for expositional convenience. As we will see below, debates in the sciences of memory concerned the existence of (multiple) memory systems *tout court.* See also note 34.

[22] As it will become clear in the main text, Tulving often talked of autonoesis, mental time travel, and chronestesia in the same breath. All of these notions became prominent and were meant to illuminate a cluster of—not fully independent—properties characterizing the experiences of remembering and imagination. The focus on autonoesis here is primarily for the purpose of gaining some clarity and explanatory leverage. Arguably a similar story can be told by focusing on mental time travel.

autobiographical reference (pp. 389-390). By 1983, this experience marked one of the many—Tulving lists 28 (twenty-eight!)—diagnostic features of episodic memory:

> Remembered events are felt by rememberers to be personal experiences that belong to the autogenous past… [T]hey tend to have a definite affective tone that is uniquely and unmistakably one of the salient attributes of recollective experiences (1983, p. 48).

Rememberers, in other words, experience remembered events as events that have happened *to them* in the past.[23] The personal nature and the accompanying feeling of veridicality—roughly translatable to "I know that I experienced this then-and-there"—were taken to be distinctive features of event recollection, being characteristically absent from memories of facts. Hence, they were important properties of the explanandum phenomenon, to be accounted for by appeal to the operations of the episodic memory system.

This is precisely what Tulving aimed to do. On the view presented in *Elements* (1983), the episodic system was taken to process information about previously experienced events, with the information registered in the system in a characteristically direct way—in a non-symbolic, non-conceptual form (pp. 41-44). Since information could be retrieved from this system only if previously entered into its store *and* the system had very limited inferential capabilities, episodic memory was taken to afford "immediate, or first-hand knowledge" of personally experienced events (p. 41). Here, crucially, "episodic memory" refers to the *system*: the underlying competence posited to explain properties of memory performance, phenomenological (the experience of possessing first-hand knowledge) as well as behavioral (its utilization). A specialized episodic system, the guiding idea was, enabled the ability of individuals to remember events from their personal past in the distinctive way presented above. Tulving's view was, of course, designed to face the tribunal of experimental evidence. This means that the guiding idea could turn out to be false: the recollection of personally experienced events may, after all, not be supported by a *distinct* episodic memory system. Galilean boldness requires falsifiability, at least in principle.

Starting with a tone-shifting article in 1985, this explanatory stance was supplemented with an increasingly sharpened focus on the experiential character of event recollection. In the aptly named "Memory and Consciousness" (1985b), Tulving lamented the "neglect of consciousness" in the study of human memory, a state of affairs he found curious given that "to remember an event means to be consciously aware now of something that happened on an earlier occasion" (p.1). Aiming to remedy this problem, he catalogued the kinds of consciousness purportedly characteristic of different forms of memory.

[23] Tulving was, of course, building on a long and complex history of philosophical and psychological characterizations of the experience of personal recollection, a history which I omit for reasons of brevity.

With a nod to Husserl, he labeled the kind of consciousness characteristic of episodic memory "autonoetic" (i.e., self-knowing).[24] Autonoetic consciousness, he told us, was:

> *…necessary for the remembering of personally experienced events*. When a person remembers such an event, he is aware of the event as a veridical part of his own existence. It is autonoetic consciousness that *confers the special phenomenal flavour* to the remembering of past events, the flavour that distinguishes remembering from other kinds of awareness, such as those characterizing perceiving, thinking, imagining, or dreaming (1985b, p. 3, emphases added).

Autonoesis thus provides a clear phenomenal marker of remembering, which Tulving took to be a contemporary variant of James' (1890) "warmth and intimacy". It is no longer one of the, presumably reasonably many, salient attributes of recollective experience. It is *the* attribute—the distinguishing, flavor-conferring component—characterized as necessary for personal remembering. In subsequent articles, Tulving did not shy away from these ideas. In 1989, he started an important methodological article with a direct identification of autonoesis with the experience of warmth and intimacy, marking it as the central feature of memory, a claim echoed in subsequent articles (Tulving 1991, 1993). By 1997, autonoesis had become a core property of episodic memory and "the major defining difference between episodic and semantic memory" (Wheeler et al. 1997, p. 350; cf. Tulving 2002a, 2005), a thesis anchoring experimental and neuroimaging work from the period. Indeed, Tulving and his colleagues made clear that this choice of words was not accidental: they had "defined episodic memory in terms of its dependence on autonoetic awareness", making the relation between the two "as much a matter of definition as a matter of empirical facts" (1997, p. 343).[25]

What should we make of this development? At first glance, the characterization of autonoesis does not seem to sit well with Tulving's Galilean commitments. One of these commitments, recall, was to the multiple determination of memory performance. The structure and content of typical memory experiences are determined by the operations of multiple underlying systems; the episodic and semantic system, at a minimum. As a result, the properties of these systems cannot be directly inferred from the introspected character of the experiences. How, then, can Tulving justify characterizing autonoetic consciousness as a distinguishing feature of episodic memory? And, more puzzlingly, how can the relation between the two be a matter of *definition*, if "episodic memory" was taken to refer to a neurocognitive structure with specific

[24] He reserved "noetic" (knowing) for the *im*personal consciousness linked with semantic memory, and—somewhat curiously, given received wisdom—"anoetic" (non-knowing) for the alleged consciousness associated with procedural memory.
[25] Tulving's sharpened focus on autonoesis was, unsurprisingly, accompanied by a progressively stronger insistence that episodic memory is *uniquely* human (see Tulving 2005 for the final verdict). That said, the claim that episodic memory and autonoesis are uniquely human appears in the very first paragraph of *Elements* (1983), yet another reason to think that Tulving's views had not really changed as much in those twenty-odd years as it is widely believed.

physiological—and presumably empirically discoverable—properties? Would phenomenologically inspired definitions of other physiological structures (digestive, endocrine, renal etc.) be welcome? Or was Tulving lapsing into a form of methodological dualism, embracing standards only appropriate for the study of human minds (Chomsky 2000)?

There is a familiar strategy for resolving this puzzle, which has gained prominence in recent years. According to it, Tulving was not in fact attempting to define properties of the neurocognitive system. Rather, he was only identifying a distinct kind of recollective experience characterizing episodic memor*ies.* He was, to put the point in our Chomskyan idiom, dealing with memory performance, not competence. Perrin et al. (2020, p. 1, first emphasis added) articulate this view well:

> Tulving (e.g., 1985)...argued that what it is to episodically remember *is* to have a mental image of a past event accompanied by [autonoetic] consciousness… In episodic memory, one does not, in other words, merely *know* that the represented event occurred in one's past; one in some sense *relives it*: to episodically remember is "to consciously re-experience past experience" [Tulving 2002a, p. 6].[26]

This reading makes Tulving's characterization of the relation between episodic memory and autonoesis as partly "a matter of definition" less mysterious: he was simply identifying an experiential kind. It also aligns his project with the views of a growing number of theorists aiming to characterize the distinctive phenomenal character of episodic remembering (e.g., Hoerl 2001; Dokic 2014; Fernández 2019; Mahr & Csibra 2018). Finally, the view seems to receive support from Tulving's (1989) avowed rejection of what he called "the doctrine of concordance", positing a general correspondence between kinds of cognitive processes and kinds of experience. Here, one may think, is a case in point: an experiential kind that need not correspond to a distinct cognitive operation.[27]

We should resist this interpretation. This is not only because Tulving continually, and quite explicitly, linked autonoesis to the operations of the episodic memory system.[28] It is also because by setting aside—or even downplaying—this commitment, we would miss the important ways in which it informs the contemporary debate. In the next section, I will argue that Tulving was in fact presenting an empirical hypothesis about the link between autonoesis and episodic retrieval. In its support, he offered crucial data,

[26] It is probably worth noting that Perrin et al. (2020) are actually *mis*quoting Tulving here. In the relevant passage, Tulving tells us that episodic memory is the only *memory system* "that allows people to consciously re-experience past experiences" (2002a, p. 6). The construction "to episodically remember" does not appear in the 2002 article and, in fact, rarely does in Tulving's articles. As I argue in the text, the slide from competence to performance may be a bit more pernicious than philosophers take it to be.

[27] Tulving (1989), in fact, only rejected the *a priori* acceptance of concordance across all domains. The key point was that the doctrine, if accepted independently of, and prior to, empirical investigation, can obscure important differences between kinds of processes. See his (1999) on why processing theorists—allegedly—make this mistake. See also 4.3.

[28] For the detail-oriented: see, e.g., 1985a, pp. 385-388; 1985b, pp. 2-3; 1987, pp. 72-73; 1993, pp. 68-69; 2002a, pp. 5-6; 2005, p. 9; Wheeler et al. 1997, pp. 332-333.

purportedly predictable only by theories positing a distinct episodic system. So, what may look like Tulving abandoning his Galilean commitments is actually him doubling down. My purpose is not merely exegetical. I aim to illustrate the dynamics of adopting the Galilean explanatory style in an, often rapidly, evolving science. Idealization can yield far-reaching theoretical and experimental insights. Yet, as we will see in section 4, it can also obscure the necessity of systematic de-idealization for the purposes of experiment design and theory confirmation.

3.3. Autonoetic Retrieval and Crucial Data

By linking episodic memory and autonoesis, Tulving put forward an empirical hypothesis about the episodic memory system. On the hypothesis, retrieval of information from the system is necessarily accompanied by autonoetic consciousness. This is likely in virtue of the system's distinctive properties: its proprietary representations and/or computational operations. Despite what the occasional rhetoric may suggest, this hypothesis is subject to experimental scrutiny, correction and amendment, and potentially falsification. In its support, Tulving presented phenomenological, neuropsychological, and behavioral data, including results from a novel experimental procedure—the remember-know paradigm—that would become an important tool in the study of memory and consciousness. The data was advanced as crucial data, predictable only from theories positing a functionally dissociable episodic memory system. Data about autonoesis, to simplify only slightly, pointed to the existence of episodic memory. In this section, I examine these claims in order.

The evidence for the proposed interpretation is overwhelming. Tulving (1983), as we saw above, sought to account for the nature of recollective experience by appealing to the characteristic operations of the episodic system. A key element of this proposal was the *synergistic ecphory* model, on which the characteristics of recollective experience are jointly determined by episodic and semantic (retrieval) information. In Tulving (1985b), the model was put to use to frame a hypothesis about the correlation between episodic memory and the newly named autonoetic consciousness:

> Overt memory performance can be supported by different combinations of episodic trace information and semantic retrieval information… From the hypothesized correlation between episodic memory and autonoetic consciousness it follows that the kind of conscious awareness that characterizes an act of recollection varies with the nature of the "mix" of trace and cue information (p. 7).

The larger the proportion of episodic information in the mix, the greater the degree of autonoetic consciousness accompanying a recollective experience was expected to be (pp. 7-10). "Pure" episodic

memories, which may indeed be as rare as free chlorine, will be maximally autonoetic.[29] The model did not only link autonoesis and episodic retrieval, but made specific, and potentially experimentally testable, predictions. As Tulving's theories matured, the hypothesized link continued to play a major role, despite an important change in emphasis.[30] It informed discussions of episodic memory as an underlying competence enabling the binding of multifeature representations in autonoetic recollective experience (Tulving 1987, 1991, 1993). It also—as indeed was made quite clear—framed the interpretation of data from the pioneering neuroimaging (PET) studies aiming to identify the brain regions involved in episodic retrieval (Nyberg et al. 1995; Kapur et al. 1995).

This is the appropriate context for understanding the appeal to definitions. When Wheeler, Stuss, and Tulving (1997) defined episodic memory in terms of autonoesis, they were not aiming to identify the properties of an experiential kind independently of experiment. Quite the contrary, they were framing a hypothesis about the episodic system, which afforded productive experimental investigation of its properties. As Colaço (2022) convincingly argues, definitions often play the role of hypotheses in the process of scientific discovery (cf. Feest 2010). They pick out phenomena of interest, allowing researchers to test inferences about them, including ones pertaining to the clustering of relevant properties. The adoption of a definition as a hypothesis about how certain properties cluster exemplifies the on-going practice of categorization of imperfectly known phenomena—a scientific kinding-in-progress. A careful look at the relevant passages in Wheeler et al. (1997) makes it clear that they should be interpreted along these lines. Immediately upon characterizing the link between episodic memory and autonoesis as definitional, the authors identified promising experimental tools for its study. They went on to catalog clinical and developmental evidence about the relation between episodic memory and self-awareness—a phenomenon they took to be closely related to autonoesis—and about the role of the frontal lobes in the production of the phenomenology of recollection (pp. 343-348). Defining episodic memory in terms of autonoesis allowed Wheeler et al., as it did future researchers, to test which properties of these phenomena cluster. Tulving's repeated characterization of the link as a working *hypothesis,* whose main function was to guide research, provides further support for this interpretation (e.g., 1985b, p. 7; 1987, p. 76; 1993, p. 68).[31]

[29] This model leaves some important questions open. E.g. what are the degrees of autonoetic consciousness constituted by and how are they manifested? Should we think of them in terms of the strength or vividness of the relevant feelings? While questions of this kind are certainly pressing, Tulving's hypothesis was that specific experimental procedures will reflect the underlying nature of the experiences. Thus, in the newly devised remember-know experimental paradigm, the degree of autonoetic consciousness was taken to be reflected in the strength of the disposition to judge that a certain event/item is *remembered* and not just known.

[30] See 4.1.

[31] Cf. Tulving (2005): "Let us begin with a thumbnail sketch, or definition, of episodic memory. Because definitions do play a role in the study of nature, even in today's dominant Zeitgeist of "exploratory" science, and *because definitions have a habit of changing, it is helpful to identify definitions* in a way that sets them apart from others in their class" (p.9, emphasis added).

The focus on autonoetic consciousness—at least in the period between 1985 and 1997—is not to be explained by the purported fact that it is essential for episodic memory. If episodic memory has an essence, on the Galilean picture, it is constituted by a number of systemic properties, with autonoetic retrieval only one of them. It is rather to be explained in epistemic terms.[32] As a *measurable* (Tulving 1985b, pp. 6-10) and *distinguishing* feature of episodic memory—hypothesized to be entirely lacking in semantic retrieval (1985b, pp. 2-3; Wheeler et al. pp. 348-349)—autonoesis afforded a promisingly direct route to its study. To properly appreciate this point, recall Tulving's methodological postulates. Since memory tasks are multiply determined, theorists must keep track of potential interactions when inferring properties of a system from experimental results. Maximizing the prospects for good inferences requires careful alignment of the two—i.e., operationalization via a set of tasks believed to closely track relevant properties of the systems. On the view presented in *Elements* (1983), the hypothesized nature and organization of episodic memory information provided good candidate properties for experimental tracking. This was manifested in the prominence of list learning experiments, in which subjects were tasked with identifying previously studied items and the spatiotemporal relations between them. Yet, the growing awareness of the prevalence of systems interactions would complicate the picture. Experimental evidence, some of which alluded to above, suggested that subjects may perform reasonably well on such tasks relying solely on *non*-episodic processes (e.g., Hayman et al. 1993). By 1997, results of this kind had pretty much convinced Wheeler, Stuss, and Tulving that "semantic memory can handle any propositional fact about the world, including facts that directly involve the rememberer" (p. 349). On this view, successful utilization of information about personally experienced events—or their spatiotemporal structure—does *not* necessarily require the involvement of the episodic memory system.[33]

With this point in mind, we can have a closer look at the role the pertinent data played in the development and justification of the theory. Data about autonoesis, a kind of consciousness hypothesized

[32] Again, the claim is *not* that autonoesis isn't a part of episodic memory's essence. It is only that the prominence of the notion of autonoesis is not best explained by its relation to the alleged essence of episodic memory. It is rather explained by the fact that its *distinguishing* character was considered a source of crucial data (see the main text below). On the Galilean picture, a thing's essence is constituted by more than just its epistemically distinguishing features. That said, it should not be denied that Tulving does occasionally express peculiar views on these issues. (For example, his responses in an interview with Gazzaniga (1991, pp. 90-92) are particularly puzzling in this regard.)

[33] In the literature, this shift is often characterized as signaling a change in Tulving's conception of episodic memory: from a system that stores *what-where-when* information to a system to a system characterized in terms of the subjective experience of autonoesis (see, e.g., Cheng et al. 2016). As I aim to illustrate in the main text, we should be careful about how we frame this point. In an important sense, Tulving's conception of episodic memory did *not* change in the relevant period: in both Tulving (1983) and Wheeler et al. (1997), episodic memory is characterized by a proprietary information store and a retrieval process accompanied by autonoetic consciousness. What changed rather was Tulving's conception of *semantic* memory: from a system that did not store what-where-when information of the relevant ("episodic") kind to a system that did, or at least could (cf. Renoult & Rugg 2020). As a consequence, the utilization of such information—e.g., in an experimental task—was no longer seen as a reliable indicator of the involvement of episodic memory, triggering a strong shift in emphasis toward autonoesis, which *was* seen as such an indicator. This is precisely why data about autonoesis came to play the role of crucial data in Tulving's developing theory. (See the main text below.) I am grateful to an anonymous reviewer for prompting me to clarify this point.

to *uniquely* accompany retrieval from episodic memory, were advanced as crucial data. Predictable from Tulving's theory, but not from rival theories, they were thought to play a key role in adjudication among them. The competing theories, recall, contested the existence of a distinct episodic system, either by positing a unitary declarative system or by eschewing talk of systems altogether. A prominent class of *processing* theories, for example, sought to explain dissociations between performance on explicit and implicit memory tasks—roughly: between tasks that required awareness of memory at retrieval and tasks that did not—by appealing to different modes of processing. While explanatory strategies varied, a notable one involved the distinction between perceptual and conceptual processing (Blaxton 1989, Roediger et al. 1990; Roediger & McDermott 1993). Their differential engagement in explicit and implicit tasks—as well as the interactions between such engagement at encoding and retrieval (cf. Bransford et al. 1979)—were taken to account for the pattern of dissociations, obviating the need for positing functionally dissociable memory systems. In such theoretical landscape, data about autonoesis would seemingly play an important role. If, as Tulving argued, autonoesis uniquely accompanies episodic retrieval, then data about it would be crucial—potentially adjudicating in favor of theories that posit functionally distinct episodic and semantic systems.[34]

Data about autonoesis come from a number of sources, but three played the most prominent roles in the debate. First, there are the phenomenological data, pointing to the distinctive conscious experience accompanying personal recollection. While these do not afford a direct inference to the properties of episodic memory, they constitute one of the converging evidential threads. The second source is the remember-know experimental paradigm, introduced by Tulving (1985b) and further refined by Gardiner (1988, 2001). Data from the paradigm, designed to gauge the kind of awareness associated with retrieval, showed robust correlations between the nature of the memory task (e.g., free or category recall) and the tendency of participants to judge that they *remembered* a previously studied item, rather than simply *knowing* that they had studied it. The data was taken to provide evidence for the correlation between episodic retrieval, thought to be strongly tapped by free recall tasks, and autonoetic consciousness (1985b, pp. 7-9).[35] The final source is unquestionably the most famous. Tulving (1985b) introduced K.C. (there simply N.N.), a patient with a profound amnesia for personal events, caused by a diffuse brain damage, including an almost complete loss of hippocampal tissue (cf. Rosenbaum et al. 2005). K.C. had relatively intact general knowledge of the world yet could not "recall a single event or incident from [his] past", which

[34] On one reading of the dialectic, there were actually *two* issues debated in parallel: the first concerning the very existence of memory systems (as opposed to, e.g., processes), the second the existence of *multiple* (declarative) memory systems. Part of the difficulty of presenting the dialectic in a non-biased way is due to the fact that rival theorists understand it differently. For Tulving (e.g., 1999, p. 12), only the second issue is worth seriously debating, with the opposition between processes and systems "a false belief". For processing theorists, in contrast, the first question takes precedence (see, e.g., Roediger et al. 1999). For expositional convenience, we are forced to discuss the two issues together. In any case, *Tulving's* key claim was that processing theorists could not account for the dissociations between "autonoetic" and "noetic" tasks, even if they could account for the dissociations between explicit and implicit ones. See below.

[35] Results from this paradigm have been interpreted in a number of distinct ways after Tulving. For an overview, see Dunn (2004).

was reflected in his extremely low scores on recognition and cued-recall tests (1985b, p. 4). Strikingly, this impairment was paired with a—heterophenomenologically reported—disturbance of autonoetic consciousness: "[K.C.]'s knowledge of his own past seem[ed] to have the same impersonal experiential quality as his knowledge of the rest of the world" (*ibid*). Namely, K.C. could report (some) facts about his life—e.g., where he spent his summers as a teenager or where he went to school—but could not "mentally travel" back to re-experience any particular event in an autonoetic way. Correlated impairments of personal recollection and, what was typically characterized as, autonoesis were subsequently reported in a number of amnesiacs (e.g., Levine et al. 1998; Klein et al. 2002; Kwan et al. 2010).

Hence, converging data pointed to the systematic correlation between personal recollection and autonoetic consciousness. In non-clinical populations, phenomenological reports were supplanted with behavioral assessments of subjects' remembering judgments, taken to illustrate such correlation. In clinical populations, the selective impairment of the capacity to remember details from the personal past correlated with a disturbance of autonoesis. A theory positing a distinct episodic system could account for these data in a straightforward way. Autonoesis correlates with personal recollection—but *not* with recall of general facts—because recollection involves the retrieval of information from a specialized, functionally dissociable, episodic memory system. Autonoesis is *not* an interaction effect: it necessarily accompanies episodic retrieval (Tulving 1985b, pp. 7-10; 1987, pp. 75-76; 1991, pp. 68-70). In contrast, theories that do not posit (multiple) declarative systems would have difficulties accounting for the data. They would have to explain—in a non-ad hoc way—why, given that there is no episodic store, the retrieval of only certain kind of information is accompanied by autonoetic consciousness. Moreover, they would have to account for the *selective* impairment of personal recollection and autonoesis in subjects with relatively intact general knowledge. Tulving considered the prospects of such "unitarian" theories—which he took processing theories to ultimately be (Tulving 1999)—to adequately respond to these challenges pretty dim. Despite the variety of explanatory resources, they could not account for a simple fact: that "one and the same behavioral response [in a recall or recognition test] could reflect either of two different states of conscious awareness of the past" (Tulving 2002a, p. 5). By distinguishing autonoetic and noetic consciousness—and linking them to the functioning of the episodic and semantic system—Tulving's theory could thus straightforwardly account for crucial data in ways that rival theories could not.[36]

We should be careful with this point. Since all *actual* tasks are multiply determined, the correlation between autonoesis and personal memories will be imperfect. Performance on tasks that tap the episodic system more strongly—perhaps by requiring a larger proportion of episodic information in an ecphoric

[36] To be very clear, I do *not* endorse this conclusion. I am rather attributing it to Tulving, aiming to illustrate the role autonoesis data played in the development of the theory. As we'll see in 4.2, processing theories are alive and reasonably well.

"mix"—will be associated with higher degrees of autonoesis. So, to illustrate the point more clearly, we should consider an idealized scenario. Ex hypothesi, an *ideal* "autonoetic" task would be a task the successful performance of which requires a maximally autonoetic memory—i.e., a memory with the highest degree of autonoesis.[37] On Tulving's theory, successful performance on such a task would depend solely on the operations of the episodic system. So, the theory would predict perfect correlation between task performance and episodic memory functioning (e.g., availability of episode details). Unitarian theories issue different predictions. If there is no distinct episodic system, then we should not expect perfect correlation between the recollection of personal details and ideal task performance. The data presented above, while not from an ideal task, were taken to better fit the predictions of Tulving's theory. Strong correlation between personal recollection and autonoesis favors *splitting* episodic and semantic memory. See Figure 3 for illustration.

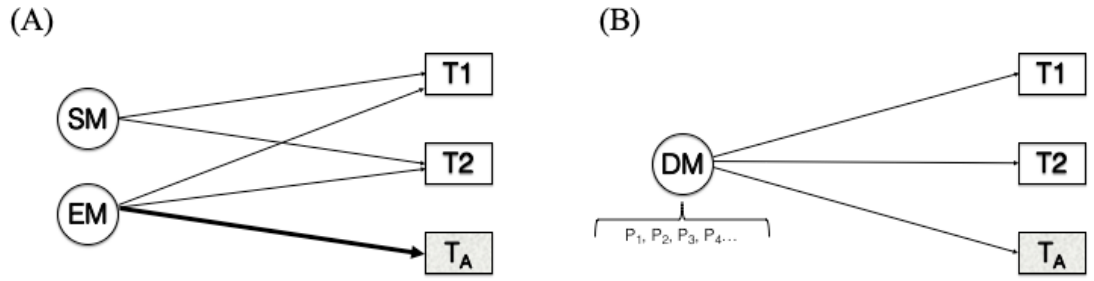

Figure 3. Crucial Data
(A) Tulving's theory predicts perfect correlation between EM function and performance on an idealized task T_A, requiring a maximally autonoetic memory. If EM is impaired and does not contribute to T_A, performance on the task is completely compromised, even when performance on T1 or T2 is partially spared. (B) On 'unitarian' theories, there is no dedicated system, whose proper functioning correlates perfectly with performance on T_A. Performance on T_A is expected to correlate more strongly with performance on T1 and T2. Tulving took the data about autonoesis to better fit the predictions of theories that posit a distinct EM system and to thus constitute *crucial data*. 'P_1','P_2','P_3' & 'P_4' stand for the constituent processes of the DM system.

In its formative period, the science of episodic memory clearly exemplified the Galilean explanatory style. Episodic memory was a neurocognitive system, with distinctive computational and representational properties, hypothesized to underlay the ability of individuals to remember their personal past. The investigation of this system, in idealized isolation from its interactions with other systems, was driven by the pursuit of "deep" explanatory principles that would unify a variety of empirical generalizations about memory. These principles, and the nature of the rival theories, determined which data

[37] Two points are worth highlighting here. First, the notion of an *ideal* autonoetic task is only a conceptual tool, employed to illustrate the different predictions of Tulving's theory and processing (unitarian) theories. Its employment involves no commitment to the claim that such a task is (meta)physically possible. Indeed, given his views, Tulving would likely consider it impossible. Second, by appealing to degrees of autonoesis to characterize "maximally autonoetic" memories, the proposal inherits the problem of clarifying the nature of such degrees. As I indicate in note 29, this is a serious problem for Tulvingian theories.

were selected and considered crucial. Hypothesizing that autonoesis uniquely accompanied episodic retrieval, Tulving considered it a source of crucial data, predictable only from theories positing a functionally distinct episodic memory system. Hence, despite appearances, his focus on autonoesis was also characteristically Galilean, typifying the search for crucial data. In the long run, nevertheless, the Galilean tendency to do away with "surface" taxonomies can clash with the, often unyielding, desire to save the phenomena. In section 4, I will attempt to illustrate how this clash manifests in the contemporary science and philosophy of episodic memory. With rapid developments providing new tools for assessing Tulving's hypotheses, theorists are confronted with the pressing need to *de*-idealize and examine the accommodation—or indeed: elimination—of common-sense notions of memory.

4. Between Systems and Phenomena

Tulving's pioneering treatment of episodic memory formed a basis for a flourishing research program. Theorists working under the umbrella of the multiple systems approach have explored episodic memory from a variety of perspectives, aiming to unify an increasing number of phenomena while gradually eliminating simplifying assumptions. In 4.1., I show how this approach has led to the formation of a new generation of Galilean theories, positing a wider cognitive system underlying both remembering and imagination. I introduce the two major families of such theories, highlighting their disagreement about the relation between episodic memory and autonoesis. In 4.2, I argue that the investigation of this relation, and of the explanatory prospects of the rival theories, requires a better understanding of the modes of interaction between memory systems. This creates a pressing need for de-idealization, triggering a new search for crucial data. At the same time, it enlarges the gap between memory systems and phenomena, opening difficult questions about their relationship. In 4.3., I explore some of these questions, examining the role of commonsense notions and categories in scientific theories of memory.

4.1. Episodic Memory and Autonoesis: Tulving and Beyond

Wheeler, Stuss, and Tulving (1997) concluded their review article with another shift in emphasis. Autonoetic consciousness, which they still considered a "critical" feature of the episodic memory system, was proposed to be expressible in *many* forms of higher cognition. The authors, indeed, re-characterized personal recollection as one form of a more general capacity marked by such consciousness—the capacity to "consider self in the past, present, and future" (p.346). It was a sign of things to come. The next quarter century would bring an explosion of research on episodic memory and autonoesis. Recognizably Tulvingian in its central themes and concerns, the research would be characterized by two important developments. First, steadily accumulating evidence would reveal a surprisingly deep processing connection between recollection and a variety of other, prototypically imaginative, activities. Second, there would be more

systematic examination of the nature of autonoesis and the way it is manifested in different cognitive activities.

A wealth of clinical, neuroimaging, and behavioral data, pointing to the close processing connection between remembering and (future-oriented) imagination, has gradually amassed. Tulving's (1985b) characterization of K.C.'s clinical profile was again agenda-setting. Not only was K.C. unable to remember particular episodes from his life; he was also unable to "imagine anything that he is likely to do on a subsequent occasion" (p. 4). His profound amnesia, it seemed, was as much an impairment of imagination-supporting consciousness as it was of personal memory. Subsequent neuropsychological studies confirmed and extended this result, with hippocampal amnesiacs showing difficulties in imagining novel events and scenarios—located not only in the future, but also in the possible past (Klein et al., 2002; Hassabis et al., 2007; Rosenbaum et al., 2009; De Brigard & Parikh 2019). Neuroimaging studies provided converging evidence for the close relationship, implicating brain networks consistently engaged in both remembering and imagination of possible events (Schacter et al. 2012; Mullally & Maguire 2014; Hodgetts et al. 2016). Behavioral research revealed additional parallels, such as analogous temporal proximity effects and dependence on the capacity for generating mental imagery (D'Argembeau and Van der Linden, 2004, 2006). Finally, studies with children yielded evidence of an early developmental relation between personal remembering and imagination (Quon & Atance 2010; Coughlin et al. 2014).

More systematic, and conceptually rigorous, examination of autonoetic consciousness accompanied these developments. Theorists made a sustained effort to identify its different qualitative components, including self-reference, the sense of subjective time and space, the feeling of veridicality etc. This led to the development of a flurry of related notions, from the catchy mental time travel (Tulving 1983, 1985b, 2002a; Suddendorf & Corballis 1997, 2007), through chronestesia and self-projection (Tulving 2002b; Buckner & Carroll 2007), to the more esoteric palinscopy and micro-time (Tulving & Lepage 2000; Hassabis & Maguire 2007).[38] These were not meant to describe fully independent phenomena, but rather to provide multiple perspectives on the cluster of properties characterizing the experience(s) of remembering and imagination—adjusting the focus with the explanatory demands. The notions influenced philosophical characterizations of the states' phenomenal character (Hoerl 2001; Dokic 2014; Perrin et al. 2020). More prominently, they anchored experimental work aiming to identify the variety of computational operations underlying the phenomenology. Extant work on the remember-know paradigm was supplemented with novel procedures for studying self-reference and self-projection (e.g., Arzy et al. 2008; Anelli et al. 2016), the processing of person, space, and time (e.g., Spreng & Mar 2012; Gauthier and van Wassenhove 2016;

[38] One cannot but admire Tulving's prodigious concept creation. If philosophy is uniquely characterized by the creation of concepts (Deleuze & Guattari 1994), then he is quite the philosopher.

D'Angelo et al. 2023), and the ways in which these processes affect the phenomenology of remembering and imagination (see Miloyan et al. 2019; Miloyan & McFarlane 2019).

Memory scientists have framed these developments in miscellaneous, but generally congruent, ways. The core Galilean insight—and it is now time to switch to present tense—is that personal remembering is subserved by the operations of a cognitive system, which also supports future-oriented and counterfactual imagination. Crucially, this is not simply a reiteration of the commitment to the multiple determination of memory performance. It is a new guiding idealization: a cognitive system *for* remembering and imagination. Yet, what often goes unappreciated is the variety of competing characterizations of the function and structure of this system. For our purposes, we can distinguish two major families of theories with different explanatory priorities and seemingly different theoretical commitments.

The first family of theories—which, for a really unfortunate lack of a better word, we may call *autonoetological*—emphasize the systemic relation between event representation and autonoesis. Their focus is on the computational operations that enable autonoesis and its manifestation in past- and future-oriented cognition: self-reference, self-projection, temporal processing etc. For Tulving (2005), episodic memory is now a neurocognitive system that:

> …makes possible mental time travel through subjective time — past, present, and future. This mental time travel allows one, as an "owner" of episodic memory ("self"), through the medium of autonoetic awareness, to remember one's own previous "thought-about" experiences, as well as to "think about" one's own possible future experiences. (p. 9).

Personal remembering and imagination are deeply connected *because* they are both forms of such autonoetic time travel, enabled by an integrated neurocognitive system. Similar views abound. Suddendorf & Corballis (1997, 2007) likewise identify a system for mental time travel, linking its proper function to the prediction of the future. Buckner & Carroll (2007) talk of a system for self-projection enabling remembering and prospection, an idea refined by Klein (2013, 2016) who takes the autonoetic components of episodic memory to be of primary causal relevance for such projection.

There is, however, a second family of theories that is equally prominent. *Simulation* theories focus on the construction of quasi-perceptual event representations—simulations, models or scenarios—used in a variety of contexts and tasks. Personal remembering and imagination are connected, on these views, because they both prototypically employ such representations. Schacter & Addis (2007, 2009), notably, posit an "episodic construction system" dedicated to the simulation of possible events, a process that involves the flexible recombination of details from previously experienced events. Closely related accounts characterize the function of the system as the construction of atemporal scenes—or dynamic scenarios—

which may be put to different uses, including remembering, imagination, and navigation (Hassabis & Maguire 2007, 2009; Rubin & Umanath 2015; Cheng et al. 2016). Importantly, simulation theories are *not* typically committed to the systemic relation between event representations and autonoesis.[39] While, for example, simulated events can be situated in the subjective past or future, they need not be (Cheng et al. 2016; De Brigard & Gessell 2016; Mahr 2020). Whether, and how often, event simulations are accompanied by forms of autonoetic consciousness is a topic of active research (see 4.2.).[40]

In this novel theoretical landscape, the question of the relation between episodic memory and autonoesis has been gradually reopened. In the next section, we will see why this reopening requires systematic de-idealization and triggers a new search for crucial data.

4.2. Systems and Phenomena: A New Search for Crucial Data

On Tulving's (1985b, 1987) theory, a dedicated episodic memory system underlies the ability of individuals to remember events from their personal past. The system—theorized about in idealized isolation from its interaction with other systems—has a number of characteristic properties, among which autonoetic retrieval stands out as a distinguishing one. Autonoesis is necessarily linked to episodic memory function in that "information about an experienced event can be retrieved only explicitly...with conscious awareness of the earlier experience" (Tulving 1999, pp. 20-21). Tulving's theory, as we have seen, afforded productive investigation of episodic memory, shedding light on the problems facing unitarian theories. If autonoetic and noetic consciousness regularly accompany retrieval from distinct memory systems, then theories that do not posit such systems make at least one distinction too few. Yet, as some prominent critics have pointed out (e.g., Roediger et al. 1990), the explicitly endorsed logic of functional dissociations has to be applied consistently and across the board. Systems theorists need not only establish dissociations *between* proposed memory systems. They also have to look *within* them—e.g., to the possible dissociations between episodic representation and autonoesis. If the two can come apart—and do so regularly and/or frequently—then there may not be an integrated system for autonoetic remembering after all. The systematic exploration of this issue, as we will see, requires a better understanding of the modes of interaction between episodic memory and other cognitive systems, to be achieved by gradual and controlled *de-idealization.* Relatedly, it requires a careful analysis of the nature and explanatory prospects of autonoetological and simulation theories. Such an analysis, I aim to illustrate, raises difficult conceptual and empirical problems, pointing to a pressing need for crucial data.

[39] This does *not* mean that they are committed to the absence of such a connection.
[40] The line dividing simulation and autonoetological theories is likely blurrier than presented here. (E.g. it's not clear how to classify Klein's (2016) idiosyncratic account). Nevertheless, the idealized presentation helps us zero in on the issue examined in 4.2.

Potential evidence of dissociation between episodic representation and autonoesis comes from a variety of sources. One kind that has flown under the radar of Tulvingian theories concerns *implicit*—i.e., nonconscious—retrieval of episodic information. In a noteworthy study, Sheldon & Moscovitch (2010) employed the remember-know procedure to examine the relation between recollection and performance on two implicit memory tasks: lexical decision and word stem completion. They found that remembered words were associated with greater priming effects than were words given a "known" rating (or studied but non-recognized words). The results, the authors argued, illustrate the involvement of episodic retrieval processes in priming. A number of related findings have surfaced in the recent literature. Wimmer & Shohamy (2010) showed that the implicit reactivation of associated memories biases the choice between alternatives that were never directly experienced, a result that spurred important work on the role of episodic memory in decision making (see Wimmer & Büchel 2016, 2021). Ramey et al. (2022) illustrated that implicit retrieval of episode information also modulates the use of schemas in spatial memory decisions, even if more weakly than explicit retrieval does. Results of this kind dovetail nicely with two-stage theories of episodic retrieval—the first rapid and unconscious, the second effortful and conscious (Moscovitch 2008).

Closer to home, we may look at the possible absence—or lack of integration—of autonoetic components in personal remembering. Klein & Nichols (2012) caused a stir by reporting the case of patient R.B. After a head trauma caused by an automobile accident, R.B. was reportedly able to recollect particular episodes from his life in the rich, quasi-perceptual way characteristic of episodic memory, yet without the accompanied feeling that these had happened to *him.* The experiences lacked the special autonoetic flavor of "ownership", despite the fact that R.B. knew that he was, in fact, remembering.[41] Apart from pathological cases of this kind, some theorists have argued that more mundane, everyday memories can lack such autonoetic flavor. A person can arguably entertain an episodic scenario of their favorite team winning a recent football game without the accompanying sense that the event "belongs" to them or was previously experienced first-hand (Bermudez 2017; Millière & Newen 2022). While this sense of ownership is difficult to study experimentally, theorists have recently devised procedures for investigating the degree of integration between event representation and another component of autonoesis: temporal orientation. In a trailblazing study, Mahr et al. (2021) found recall of temporal information to be only weakly predicted by recall of episodic contents. Mahr & Schacter (2022), relatedly, found that subjects were consistently more likely to confuse event simulations with the same temporal orientation that they were simulations that shared their status as memories or imaginations. These results, the authors argue, suggest that temporal orientation may be determined by mechanisms distinct from those of episodic simulation.

[41] It is perhaps worth noting that some theorists have taken the report of Klein & Nichols (2012) with an amount of salt. One of the several reasons for this is the peculiar expressive sophistication of R.B., who tended to characterize his anomalous experiences in familiar theoretical terms, speaking, e.g., of taking "ownership" of memories and of his "working memory loss" (p. 688)

Neuropsychological data also points to possible dissociations, with selective impairments of episodic representation and self-related processing (Arzy et al. 2009; Andelman et al. 2010). In a particularly interesting study, Kurczek and colleagues (2015) report a double dissociation between patients with bilateral hippocampal damage and patients with bilateral medial prefrontal cortex damage. The first group of patients were impaired in their ability to construct detailed event simulations yet showed normal ability to incorporate themselves in narratives of the events. Patients in the second group, in contrast, were able to construct detailed simulations but incorporated themselves in narratives of the events much less frequently than did participants in the control group. The result, pointing to the differential contributions of the hippocampus and the medial prefrontal cortex, sits well with neuroimaging data, which suggest the existence of two distinct "subnetworks" of the midline default network—the first associated with episodic simulation and centered on the medial temporal lobes, the second associated with self-related processing and centered on the ventromedial prefrontal cortex (Andrews-Hanna et al. 2010; Dafni-Merom & Arzy 2020).

Data of this kind seem to point decisively in the direction of disintegration. Yet, things are not so simple. Memory systems are taken to be complex neurocognitive structures, with fuzzy boundaries and overlapping constituent processes. A proper assessment of the data thus requires answers to a galaxy of difficult questions. Crucially, establishing whether autonoetic operations are constituents of the episodic memory system requires specification of the degree of integration necessary for constituency. While pragmatic considerations may be unavoidable here—the explanatory context may dictate how strongly we should expect constituent processes of a system to be correlated—we should be wary of weakly motivated or arbitrary cut offs (cf. Rupert 2009). Related epistemological issues abound. The degree of integration should presumably be established relative to *all* tasks and activities that the relevant processes contribute to. (Otherwise, we would fail to distinguish systemic integration from regular interaction in the performance on a preferred family of tasks.) In circumstances of incomplete knowledge, this is impossible to do without reliance on contingent, and potentially empirically problematic, assumptions. This issue is compounded by the difficulty of establishing the immediate relevance of neuroimaging and network data, especially in light of the apparent diachronic dynamicity of neural mechanisms (De Brigard 2017; Ferbinteanu 2019).

Moreover, autonoetic consciousness itself may turn out to be mechanistically heterogeneous. The phenomenon, as we have seen, is comprised of a number of conceptually separable components: a self-related sense of ownership, a sense of (travel through) subjective time and space, a feeling of veridicality etc. It is possible, if not likely, that these components are underlaid by cognitive processes which *vary* in their degree of integration with core processes of episodic representation. There may not be, in other words, a single answer to the question of whether autonoetic processes are constituents of episodic memory. At

the stage of development, we probably cannot do better than embrace the Galilean approach and idealize away from such heterogeneity. Nonetheless, we shouldn't lose track of the substantive assumptions behind this idealization, assumptions that will eventually have to undergo more extensive empirical and theoretical scrutiny.

For these reasons, the question of the relation between episodic memory and autonoesis has been reopened, with a number of viable explanatory strategies on offer. For autonoetological theories, the organizing idea remains familiar, even if the grip of the idealization has been loosened. Autonoetically flavored recollection and imagination are the paradigmatic—indeed: signature—forms of expression of the episodic system. The best way to account for this is to characterize the system as tightly integrated, with constituent processes underlying autonoesis: a system for autonoetic time travel (Tulving 2005; Suddendorf & Corballis 2007). While component processes of this system may support activities of a different kind—e.g., lexical decision, word stem completion or spatial schema use—this is simply a reflection of the fact that systems have fuzzy boundaries and overlapping constituents. Indeed, given the intricate interactions between cognitive processes, *not* finding evidence for heterogeneous use would be more of a surprise. Similarly, the degree of integration between constituent processes of the system may vary and may even do so with the nature of the task. Hence, data showing such variety do not constitute evidence of functional distinctness; not at least in the absence of a number of ancillary assumptions.

Yet, rival explanations are readily available. On some simulation theories, autonoetically flavored thought can be seen as resulting from the interaction of episodic simulation and distinct, functionally specialized, cognitive systems. Such interaction, which may occur at both input and output stages, calibrates episodic simulations, adopting them for various forms of mental time travel: from personal recollection to episodic future thought (Mahr 2020; Andonovski 2022). Autonoetic consciousness, on this view, is an elaborate *interaction effect.* In a recent paper, Pan (2022) presents one such account, focusing on the production of the autonoetic sense of memory ownership. This sense, he argues, results from the interaction of episodic simulation—delivering only first-order event content—and the mindreading system, which interprets such deliverances as forms of self-knowledge, embedding them in a metarepresentational format. It is this "external" process of interpretation, reliant on the monitoring of a plethora of processing cues, that accounts for the distinctive phenomenology of personal recollection. Similar accounts, focusing on related components of autonoesis, are available in the literature (e.g., Mahr & Csibra 2018; Perrin et al. 2022). In fact, in light of the recent revival of processing views of memory (Henke 2010; Cabeza & Moscovitch 2013), interactionist accounts of autonoesis may even take more radical forms, denying the existence of a specialized episodic representation system (see, e.g., Klein 2016). Such accounts certainly deserve a more

comprehensive treatment, best reserved for a future occasion. Figure 4 illustrates the rival explanatory strategies.

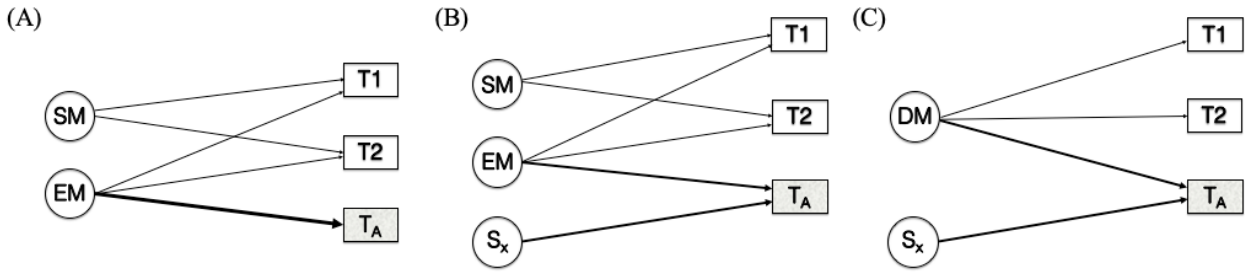

Figure 4. Rival Explanations
(A) "Tulvingian" autonoetological theory: autonoesis is produced solely by the activity of the EM system. As before, TA stands for an idealized ("maximally autonoetic") task. (B) An interactionist simulation theory: autonoesis is produced by the interaction of EM and another specialized system S_X (e.g., mindreading). (C) A radical interactionist account: autonoesis is produced by the interaction of S_X and constituent processes of a general declarative memory system (DM).

This is a familiar Galilean predicament. With accumulating data at significant remove from major theories, substantial empirical coverage is not on the table. Theories can offer in-principle accounts and idealized models—typically with a number of free parameters—of episodic memory, imagination, and autonoetic consciousness. Yet, at present, comprehensive, detailed, and predictively powerful accounts of the phenomena remain a long way away. It is not difficult to be pessimistic about this state of affairs. Widespread concerns about a "theory crisis" in psychology have brought attention to the relative lack of robust phenomena and valid psychological constructs (Eronen & Bringmann 2021), as well as to the rarity of direct theory-testing research (Oberauer & Lewandowsky 2019), issues that undoubtedly plague contemporary memory science. In a peculiar echo of Chomsky (see fn. 6), Popov (2023) worries that even if we were miraculously given the absolute truth about memory, we wouldn't be able to recognize it, not the least because we do not yet understand the relation between the data and our underlying cognitive faculties. This is certainly worth worrying about. Yet, if the Galilean is to be trusted, it is just how things tend to go, especially with messy, complicated subjects, and especially at early stages of research. Faced with these difficulties, we should embrace a properly cautious epistemological tolerance, not worry too much about empirical limitations, and focus on theories that account for key aspects of the phenomena.

We should, in other words, devise and favor theories that account for crucial data. This is much easier said than done. Identifying data that pertain to core features of episodic memory, adjudicate between rival theories, and relate to abstract explanatory principles is a daunting prospect. It may have to be done, nevertheless. If the goal is the discovery of the "organs" of memory, there are few serious alternatives to the formulation of bold hypotheses, aiming to unify different empirical generalizations and to offer

potentially far-reaching theoretical insights. Half a century after "Episodic and Semantic Memory", the Galilean search for deep principles underlying the introspectively apprehensible and objectively identifiable consequences of learning and memory remains *the* template for future work. In the final section, I discuss the accommodation of commonsense notions of memory and remembering, as such work progresses.

4.3. Episodic Memory: Science, Accommodation, and Elimination

Relating the universally familiar experience of remembering past events to an underlying neurocognitive system was the formative act of episodic memory science. At first glance, the act suggested the possibility of a direct reduction of properties of the experience to the operations of the system. To psychologists, it provided a phenomenological anchor for emerging theoretical and experimental work, legitimizing talk of episodic memory *as* the system that allows people to consciously re-experience events from the personal past. It encouraged philosophers to reopen old debates about the nature of memory, shelved since the heyday of psychological and logical behaviorism. Some saw in Tulving's project a restoration of old philosophical categories, taking episodic memory to be the kind of memory philosophers had investigated for centuries before (Martin 2001). Others, freed from the methodological constraints of conceptual analysis, were emboldened to pursue new metaphysical projects, aiming to identify personal recollection with an operation of the newly "discovered" memory system (Michaelian 2016; Cheng & Werning 2016).

If the science of episodic memory began with exploration of common experience, however, it always had its sights far beyond it. The Galilean pursuit of depth is regulated by the ideal of theoretical unification, maximal empirical coverage via the smallest number of basic explanatory principles (Chomsky 1980). It was exemplified in Tulving's radical proposal about a competence underlying *both* familiar recollective experience and the ability of subjects to identify and recall experimentally presented material. It took a particularly noticeable form in theories of mental time travel and episodic simulation, aiming to bring together a variety of phenomenologically disparate cognitive activities, from recollection and imagination all the way to route planning and navigation. The hypothesized explanatory principles—often formulated by reference to newly-formed categories, such as encoding specificity, complementary learning or scene construction—were becoming increasingly distant from commonsense descriptions of the phenomena. This is especially conspicuous on interactionist theories, which see autonoetically flavored remembering as resulting from the interaction of several underlying systems. Even on Tulvingian theories, however, the process seems inevitable and destined to intensify. It is due not only to the avowed multiple determination of performance, but primarily to the programmatic push for unification. As theories aim to cover more disparate phenomena, basic principles become more distant from the descriptions of the

phenomena. As a result, "the explanation of any given phenomenon becomes more inferentially complex" (Collins 2007, p. 629).

A recent proposal by Rubin (2022) illustrates this dynamic nicely. Dissatisfied with standard classifications of memory kinds, which he characterizes as hierarchical and categorical in nature, Rubin advances a dimensional alternative. On the proposal, episodic and semantic memory are situated in a multidimensional conceptual space, defined by a number of continuous dimensions corresponding to relevant underlying processes. While three such dimensions—the involvement of explicit, self-reference, and scene construction processes—are preliminarily chosen, the model can, in principle, be enriched with additional dimensions. Rubin makes much of the alleged virtues of the dimensional approach, as its generativity and the flexibility of accommodating novel categories. Yet for all this, the most conspicuous shift away from the standard model is not from a categorical to a dimensional approach.[42] It is from a classification of memory systems to one of memory *states*—from competence to performance. On Rubin's account, the notion of "episodic memory" survives but it no longer refers to an underlying structure producing the phenomena of recollection. The ambition of accounting for autonoetically flavored experiences by appeal to a dedicated memory system has been given up. It refers rather to states of recollection themselves—Tulving's explananda. With no direct correspondence between systems and phenomena, the nature of episodic memor*ies* can only be explained by a complex, and potentially gerrymandered, assembly of interacting processes.

This development shouldn't catch anyone by surprise. Common sense and phenomenological categories do not form constraints on the domains of explanations of mature sciences (Collins 2007). Hence, we shouldn't expect these domains to map neatly on, or elaborate, commonsense domains:

> Science does not have as its target a complete and coherent description of the world as we find it, the world as delineated by our given categories; instead, its aim is to seek highly abstract 'hidden' laws and mechanisms that unify otherwise heterogeneous phenomena, in light of which our given categories drop out, *at best,* as shallow and partial taxonomic artefacts (*ibid,* p. 632).

This "meta-scientific eliminativism" does not entail that phenomenological or folk notions not aligned with scientific classifications fail to refer; that there really are no memories, imaginings or autonoetic experiences. Nor does it entail that commonsense notions cannot be integrated—if only partially—in the organizational and explanatory schemes of mature sciences. What it does entail is that mature sciences are "not beholden to the categories and modes of explanation...employed in the everyday understanding of the

[42] In reality, the idea that memor*ies* have properties that vary along a number of continuous dimensions has a long history and is, *prima facie* at least, compatible with the existence of functionally distinct underlying memory systems (Tulving 1983, pp. 67-69).

world" (p. 630). For Galileans wary of methodological dualism, this applies as strongly to the sciences of the mind as it does to the rest of the natural sciences.

As memory sciences mature, theorists will face increasingly difficult tasks; empirical *and* conceptual. I have argued that the cataloguing of memory competences is compatible with, and leaves room for, the behavioral and phenomenological variety of memory performance. But this metatheoretical postulate does not issue specific methodological guidelines for theory and concept development. In fact, *no* general doctrine is likely to issue such guidelines (Tulving 1989). Since "there is no logical necessity for a close connection between behaviour and cognition [or] between cognition and conscious awareness", the elucidation of the relations between them is "primarily a matter of empirical study, even if, as is always true in any science, the empirical study must be complemented by rational analysis" (pp. 22-23). The search for basic explanatory principles and crucial data is, at the same time, an exploration of conceptual appropriateness. It involves difficult choices about the roles commonsense and phenomenological categories get to play in sophisticated scientific theories—about accommodation and elimination.

In philosophy, conceptual rigor and empirical adequacy must be accompanied by clarity of methodological commitments. Metaphysicians with a taste for desert landscapes will be content to hitch their wagons to scientific theories, directly identifying the phenomena of recollection with, increasingly abstract, neurocognitive operations. The resultant theories will often be intuitively unsatisfying, eliminating familiar categories and eliding distinctions that we may pre-theoretically care about. Michaelian's (2016) simulation theory, to take the most prominent example, aimed to characterize personal event recollection—a phenomenon purportedly familiar to commonsense theorizing (Ch. 4)—in terms of the operations of an underlying episodic simulation system. Yet, for empirical and conceptual reasons—the former related to the data presented in 4.2., the latter closely linked to McCarroll's (2020) criticism—Michaelian (2022) was led to "radicalize" the theory, concluding that autonoesis is *in*essential to episodic memory. Episodic memory, he tells us in a characteristically eliminativist gesture, "turns out not to be equivalent to what philosophers have sometimes...referred to as "personal memory"" (2022, p. 17). Michaelian is happy to let science dictate his metaphysics. For theorists less happy with this metatheoretical stance, the growing divide between commonsense and scientific notions will present an even bigger challenge. They will not only have to account for the myriad ways in which neurocognitive mechanisms support the familiar activities of remembering and imagination. They will also have to carefully negotiate the conditions under which an appeal to such mechanisms licenses the revision of commonsense notions and categories. Even if phenomena are universally familiar, they may not ultimately be worth saving.

Methodological clarity is also required in philosophical theorizing about autonoesis. Unlike Tulving, philosophers have been primarily concerned with distinguishing not systems but kinds of

conscious memory *states,* with autonoesis frequently characterized as a distinguishing feature of episodic remembering (Dokic 2014; Mahr & Csibra 2018; Perrin et al. 2020). These theories face an uncomfortable dilemma. If they take the phenomenological distinctions to reflect underlying systemic differences, then they have to account not only for the multiple determination of performance but also for the emerging evidence of dissociation between episodic representation and autonoesis. If, on the other hand, they take the phenomenological distinctions to be independent of any systemic differences, then they owe us a detailed, and reasonably plausible, methodological story. What, to put the point simply, justifies talk of personal remembering as a distinct kind*,* given apparent mechanistic heterogeneity? The larger the gap between philosophical and scientific notions, the more urgent the need for methodological clarity and rigor.

5. Conclusion

Multiple systems theories appeal to the activity of complex neurocognitive structures to account for core features of memory and its phenomenology. Memory systems are best understood as idealized competences, abstracted from messy details of interaction for the purpose of gaining explanatory leverage and theoretical insight. Indeed, the positing of episodic and semantic memory has led to an explosion of research, a variety of conceptual and empirical developments that find their bearings by reference to Tulving's project. It has afforded productive investigation of the computational, representational, and neural properties of memory, revealing surprising connections, unifying principles, and mechanisms underlying seemingly heterogenous phenomena. Yet, the idealization is justified only as long as it continues to provide explanatory leverage and insight. With rapid progress, accumulation of problematic and anomalous results, and a renewed vigor in the investigation of neurocognitive and behavioral interactions (Pessoa 2022), we may very well be at the dawn of a new era in memory science. What happens to episodic and semantic memory, as anomalies trigger new idealizations, is anyone's guess.

Bibliography:

Allott, N., Lohndal, T., & Rey, G. (2021). Chomsky's "Galilean" explanatory style. In Allott, N., Lohndal, T., & Rey, G. (Eds.). *A companion to Chomsky*. (pp.517-528). Wiley Blackwell.

Allott, N., & Smith, N. (2021). Chomsky and Fodor on modularity. In Allott, N., Lohndal, T., & Rey, G. (Eds.). *A companion to Chomsky*. (pp.529-543). Wiley Blackwell.

Andelman, F., Hoofien, D., Goldberg, I., Aizenstein, O., and Neufeld, M. Y. (2010). Bilateral hippocampal lesion and a selective impairment of the ability for mental time travel. *Neurocase* 16, 426–435.

Anderson, M. L. (2015). Mining the brain for a new taxonomy of the mind. *Philosophy Compass*, 10(1), 68-77.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior,* 22(3), 261-295.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford: Oxford University Press.

Andonovski, N. (2020). Singularism about episodic memory. *Review of Philosophy and Psychology*, 11(2), 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.

Andonovski, N. (2022). Episodic representation: A mental models account. *Frontiers in Psychology*, *13*.

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., and Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. *Neuron* 65, 550–562.

Anelli, F., Ciaramelli, E., Arzy, S., & Frassinetti, F. (2016). Age-related effects on future mental time travel. *Neural Plasticity* 1867270.

Arzy, S., Molnar-Szakacs, I., & Blanke, O. (2008). Self in time: imagined self-location influences neural activity related to mental time travel. *Journal of Neuroscience,* 28(25), 6502-6507.

Arzy, S., Adi-Japha, E., and Blanke, O. (2009). The mental time line: an analogue of the mental number line in the mapping of life events. *Consciousness & Cognition* 18, 781–785

Barkasi, M., & Rosen, M. G. (2020). Is mental time travel real time travel?. *Philosophy and the Mind Sciences*, 1(1), 1-27.

Benna, M. K., & Fusi, S. (2016). Computational principles of synaptic memory consolidation. *Nature neuroscience*, 19(12), 1697-1706.

Bermúdez, J. L. (2017). Memory and self-consciousness. In K. Michaelian & Bernecker (Eds.), *The*

Routledge handbook of philosophy of memory (pp. 180–191). London: Routledge.

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(4), 657.

Botha, R. P. (1982). On 'the Galilean style' of linguistic inquiry. *Lingua*, 58(1-2), 1-50.

Bower, G. H. (1970). Organizational factors in memory. *Cognitive Psychology*, 1(1), 18-46.

Bower, G. H. (2000). A brief history of memory research. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 3–32). Oxford: Oxford University Press.

Bransford, J. (1979). *Human cognition: Learning, understanding, and remembering.* Thomson Brooks/ Cole.

Broca, P. (1861). Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech). *Bulletin de la Société Anatomique*, 6, 330-57.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, 11(2), 49-57.

Cabeza, R., & Moscovitch, M. (2013). Memory systems, processing modes, and components: Functional neuroimaging evidence. *Perspectives on Psychological Science*, 8(1), 49-55.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford: Oxford University Press.

Cheng, S., & Werning, M. (2016). What is episodic memory if it is a natural kind?. *Synthese*, 193, 1345-1385.

Cheng, S., Werning, M., & Suddendorf, T. (2016). Dissociating memory traces and scenario construction in mental time travel. *Neuroscience & Biobehavioral Reviews*, 60, 82-89.

Chersi, F., & Burgess, N. (2015). The cognitive architecture of spatial navigation: hippocampal and striatal contributions. *Neuron,* 88(1), 64-77.

Chomsky, N. (1978). A theory of core grammar. *Glot* 1(1), 7-26.

Chomsky, N. (1980). *Rules and representations*. Oxford: Blackwell.

Chomsky, N. (2000). *New horizons in the study of language and mind.* Cambridge University Press.

Chomsky, N. (2002). *On nature and language.* Cambridge University Press.

Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science*, 210(4466), 207-210.

Colaço, D. (2022). What counts as a memory? Definitions, hypotheses, and "kinding in progress". *Philosophy of Science*, 89(1), 89-106.

Collins, J. (2007). Meta-scientific eliminativism: A reconsideration of Chomsky's review of Skinner's verbal behavior. *The British Journal for the Philosophy of Science* 58, 625-658.

Coughlin, C., Lyons, K. E., & Ghetti, S. (2014). Remembering the past to envision the future in middle

childhood: Developmental linkages between prospection and episodic memory. *Cognitive Development,* 30, 96-110.

Cowell, R. A., Barense, M. D., & Sadil, P. S. (2019). A roadmap for understanding memory: Decomposing cognitive processes into operations and representations. *Eneuro*, 6(4).

Crowder, R. G. (1986). Remembering experiences and the experience of remembering. *Behavioral and Brain Sciences*, 9, 566-567.

Dafni-Merom, A., and Arzy, S. (2020). The radiation of autonoetic consciousness in cognitive neuroscience: a functional neuroanatomy perspective. *Neuropsychologia* 143:107477.

D'Angelo, M., Frassinetti, F., & Cappelletti, M. (2023). The role of beta oscillations in mental time travel. *Psychological Science*, 09567976221147259.

D'Argembeau, A., & Van der Linden, M. (2004). Phenomenal characteristics associated with projecting oneself back into the past and forward into the future: Influence of valence and temporal distance. *Consciousness and Cognition*, 13(4), 844-858.

D'Argembeau, A., & Van der Linden, M. (2006). Individual differences in the phenomenology of mental time travel: The effect of vivid visual imagery and emotion regulation strategies. *Consciousness and Cognition*, 15(2), 342-350.

De Brigard, F. (2017). Cognitive systems and the changing brain. *Philosophical Explorations,* 20(2), 224-241.

De Brigard, F. & Gessell, B. (2016). Time is not of the essence: Understanding the neural correlates of mental time travel. In Michaelian, K. et al (eds.) *Seeing the future: Theoretical perspectives on future-oriented mental time travel.* (pp.153-179). Oxford: Oxford University Press

De Brigard, F., & Parikh, N. (2019). Episodic counterfactual thinking. *Current Directions in Psychological Science,* 28(1), 59-66.

Delafresnaye, J.F. (1954). *Brain mechanisms and consciousness*. Blackwell

Deleuze, G., & Guattari, F. (1994). *What is philosophy?.* Columbia University Press.

Dokic, J. (2014). Feeling the past: A two-tiered account of episodic memory. *Review of Philosophy and Psychology, 5(3)*, 413-426

Dunn, J. C. (2004). Remember-know: a matter of confidence. *Psychological Review*, 111(2), 524.

Ebbinghaus, H. (1885). *Über das gedächtnis: untersuchungen zur experimentellen psychologie*. Duncker & Humblot.

Eichenbaum, H. (1994). The hippocampal system and declarative memory in humans and animals: Experimental analysis and historical origins. In D. L. Schacter & E. Tulving (Eds.), *Memory systems 1994* (pp. 147–201). The MIT Press

Eichenbaum, H. (2000). A cortical–hippocampal system for declarative memory. *Nature Reviews Neuroscience,* 1(1), 41-50.

Eichenbaum, H., & Cohen, N. J. (2001). *From conditioning to conscious recollection: memory systems of the brain.* New York: Oxford University Press

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, 16(4), 779-788.

Feest, U. (2010). Concepts as tools in the experimental generation of knowledge in cognitive neuropsychology. *Spontaneous Generations: A Journal for the History and Philosophy of Science*, 4(1), 173-190.

Feest, U. (2011). What exactly is stabilized when phenomena are stabilized?. *Synthese,* 182(1), 57-71

Ferbinteanu, J. (2019). Memory systems 2018–Towards a new paradigm. *Neurobiology of Learning and Memory*, 157, 61-78.

Fernández, J. (2019). *Memory: A self-referential account*. Oxford University Press, USA.

Feyerabend, P. K. (1979). *Against method. Outline of an anarchist theory of knowledge.* Verso: London

Francken, J. C., Slors, M., & Craver, C. F. (2022). Cognitive ontology and the search for neural mechanisms: three foundational problems. *Synthese,* 200(5), 378.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, 16, 309-313.

Gardiner, J. M. (2001). Episodic memory and autonoetic consciousness: a first–person approach. *Philosophical Transactions of the Royal Society of London*, *Series B: Biological Sciences*, 356(1413), 1351-1361.

Gauthier, B., & van Wassenhove, V. (2016). Time is not space: core computations and domain-specific networks for mental travels. *Journal of Neuroscience*, 36(47), 11891-11903.

Gazzaniga, M. S. (1991). Interview with Endel Tulving. *Journal of Cognitive Neuroscience,* 3(1), 89-94.

Goodroe, S. C., Starnes, J., & Brown, T. I. (2018). The complex nature of hippocampal-striatal interactions in spatial navigation. *Frontiers in Human Neuroscience*, 12, 250

Hacking, I. (1983). *Representing and intervening. Introductory topics in philosophy of science*. Cambridge: Cambridge University Press

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences,* 104(5), 1726-1731.

Hassabis, D. & Maguire, E.A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences* 11(7): 299-306

Hassabis, D. & Maguire, E.A. (2009). The construction system of the brain. *Philosophical Transactions of the Royal Society B* 364: 1263-1271

Hayman, C. G., Macdonald, C. A., & Tulving, E. (1993). The role of repetition and associative interference in new Semantic learning in amnesia: A case experiment. *Journal of Cognitive Neuroscience,* 5(4), 375-389.

Henke, K. (2010). A model for memory systems based on processing modes rather than consciousness. *Nature Reviews Neuroscience*, 11(7), 523-532.

Hodgetts, C. J., Shine, J. P., Lawrence, A. D., Downing, P. E., & Graham, K. S. (2016). Evidencing a place for the hippocampus within the core scene processing network. *Human Brain Mapping*, 37(11), 3779-3794

Hoerl, C. (2001). The phenomenology of episodic recall. In Hoerl, C., & McCormack, T. (Eds.). *Time and memory: Issues in philosophy and psychology* (No. 1). Oxford University Press

James, W. (1890). *The principles of psychology*, Volume 1. Dover Publications; Reprint edition (1950)

Kapur, S., Craik, F. I., Jones, C., Brown, G. M., Houle, S., & Tulving, E. (1995). Functional role of the prefrontal cortex in retrieval of memories: A PET study. *Neuroreport* 6(14), 1880-1884.

Kim, J. J., & Baxter, M. G. (2001). Multiple brain-memory systems: the whole does not equal the sum of its parts. *Trends in Neurosciences*, 24(6), 324-330.

Klein, S. B. (2013). The complex act of projecting oneself into the future. *Wiley Interdisciplinary Reviews: Cognitive Science,* 4(1), 63-79.

Klein, S. B. (2016). Autonoetic consciousness: Reconsidering the role of episodic memory in future-oriented self-projection. *Quarterly Journal of Experimental Psychology*, 69(2), 381-401.

Klein, S. B., Loftus, J., & Kihlstrom, J. F. (2002). Memory and temporal experience: The effects of episodic memory loss on an amnesic patient's ability to remember the past and imagine the future. *Social Cognition*, 20(5), 353-379.

Klein, S. B., & Nichols, S. (2012). Memory and the sense of personal identity. *Mind,* 121(483), 677-702.

Koyré, A. (1943). Galileo and the scientific revolution of the seventeenth century. *The Philosophical Review,* 52(4), 333-348.

Kuhn, T. S. (1962). Historical structure of scientific discovery: To the historian discovery is seldom a unit event attributable to some particular man, time, and place. *Science*, 136(3518), 760-764.

Kurczek, J., Wechsler, E., Ahuja, S., Jensen, U., Cohen, N. J., Tranel, D., et al. (2015). Differential contributions of hippocampus and medial prefrontal cortex to self-projection and self-referential processing. *Neuropsychologia* 73, 116–126.

Kwan, D., Carson, N., Addis, D. R., & Rosenbaum, R. S. (2010). Deficits in past remembering extend to future imagining in a case of developmental amnesia. *Neuropsychologia*, 48(11), 3179-3186.

Levine, B., Black, S. E., Cabeza, R., Sinden, M., McIntosh, A. R., Toth, J. P., ... & Stuss, D. T. (1998). Episodic memory and the self in a case of isolated retrograde amnesia. *Brain: A journal of Neurology,* 121(10), 1951-1973.

Levine, B., Svoboda, E., Hay, J. F., Winocur, G., & Moscovitch, M. (2002). Aging and autobiographical memory: dissociating episodic from semantic retrieval. *Psychology and Aging,* 17(4), 677.

MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal "time cells" bridge the gap in memory for discontiguous events. *Neuron*, 71(4), 737-749.

Mahr, J. B. (2020). The dimensions of episodic simulation. *Cognition,* 196, 104085.

Mahr, J. B., & Csibra, G. (2018). Why do we remember? The communicative function of episodic memory. *Behavioral and Brain Sciences*, 41

Mahr, J. B., Greene, J. D., & Schacter, D. L. (2021). A long time ago in a galaxy far, far away: How temporal are episodic contents?. *Consciousness and Cognition,* 96, 103224.

Mahr, J. B., & Schacter, D. L. (2022). Mnemicity versus temporality: Distinguishing between components of episodic representations. *Journal of Experimental Psychology: General.*

Martin, M. G. (2001). Out of the past: Episodic recall as retained acquaintance. In Hoerl, C., & McCormack, T. (Eds.). *Time and memory: Issues in philosophy and psychology* (No. 1). Oxford: Oxford University Press

McCarroll, C. J. (2020). Remembering the personal past: Beyond the boundaries of imagination. *Frontiers in Psychology,* 11, 585352. https://doi.org/10.3389/fpsyg.2020.585352

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.

McDonald, R. J., & White, N. M. (1993). A triple dissociation of memory systems: hippocampus, amygdala, and dorsal striatum. *Behavioral Neuroscience*, 107(1), 3.

McDonald, R. J., & Hong, N. S. (2013). How does a specific learning and memory system in the mammalian brain gain control of behavior?. *Hippocampus*, 23(11), 1084-1102.

McMullin, E. (1985). Galilean idealization. *Studies in History and Philosophy of Science Part A*, 16(3), 247-273.

Michaelian, K. (2016). *Mental time travel: Episodic memory and our knowledge of the personal past.* MIT Press.

Michaelian, K. (2022). Radicalizing simulationism: Remembering as imagining the (nonpersonal) past. *Philosophical Psychology*, 1-27.

Millière, R., & Newen, A. (2022). Selfless Memories. *Erkenntnis,* 1-22.

Miloyan, B., & McFarlane, K. A. (2019). The measurement of episodic foresight: A systematic review of assessment instruments. *Cortex*, 117, 351-370.

Miloyan, B., McFarlane, K. A., & Suddendorf, T. (2019). Measuring mental time travel: Is the hippocampus really critical for episodic memory and episodic foresight?. *Cortex*, 117, 371-384.

Moscovitch, M. (1992). Memory and working-with-memory: A component process model based on modules and central systems. *Journal of Cognitive Neuroscience*, 4(3), 257-267.

Moscovitch, M. (2008). The hippocampus as a ‘‘stupid’’ domain-specific module: Implications for theories of recent and remote memory, and of imagination. *Canadian Journal of Experimental Psychology*, 62, 62_79.

Mullally, S. L., & Maguire, E. A. (2014). Memory, imagination, and predicting the future: a common brain mechanism?. *The Neuroscientist*, 20(3), 220-234.

Nyberg, L., Tulving, E., Habib, R., Nilsson, L. G., Kapur, S., Houle, S., ... & McIntosh, A. R. (1995). Functional brain maps of retrieval mode and recovery of episodic information. *Neuroreport*, 7(1), 249-252.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26, 1596-1618.

Pan, S. (2022). What is so special about episodic memory: lessons from the system-experience distinction. *Synthese*, 200(1), 5.

Perrin, D., Michaelian, K., & Sant’Anna, A. (2020). The phenomenology of remembering is an epistemic feeling. *Frontiers in Psychology,* 11, 1531.

Pessoa, L. (2022). *The entangled brain: How perception, cognition, and emotion are woven together.* Cambridge: MIT Press

Pietroski, P., & Rey, G. (1995). When other things aren't equal: Saving ceteris paribus laws from vacuity. *The British Journal for the Philosophy of Science*, 46(1), 81-110.

Poldrack, R. A., & Rodriguez, P. (2004). How do memory systems interact? Evidence from human classification learning. *Neurobiology of Learning and Memory,* 82(3), 324-332.

Popov, V. (2023). If God Handed Us the Ground-Truth Theory of Memory, How Would We Recognize It? *Psyarxiv.*

Potochnik, A. (2017). *Idealization and the aims of science*. University of Chicago Press.

Quon, E., & Atance, C. M. (2010). A comparison of preschoolers' memory, knowledge, and anticipation of events. *Journal of Cognition and Development*, 11(1), 37-60

Ramey, M. M., Henderson, J. M., & Yonelinas, A. P. (2022). Episodic memory processes modulate how schema knowledge is used in spatial memory decisions. *Cognition*, 225, 105111.

Ranganath, C. (2022). Episodic Memory. In Kahana, M. & Wagner, A.D. (Eds.), *Handbook of human memory: Foundations and applications.* Oxford University Press.

Reber, P. J., Knowlton, B. J., & Squire, L. R. (1996). Dissociable properties of memory systems: differences in the flexibility of declarative and nondeclarative knowledge. *Behavioral Neuroscience,* 110(5), 861.

Renoult, L., Irish, M., Moscovitch, M., & Rugg, M. D. (2019). From knowing to remembering: the semantic–episodic distinction. *Trends in Cognitive Sciences*, 23(12), 1041-1057.

Renoult, L., & Rugg, M. D. (2020). An historical perspective on Endel Tulving's episodic-semantic dichotomy. *Neuropsychologia,* 139, 107366.

Rey, G. (2020). *Representation of language: Philosophical issues in a Chomskyan linguistics*. Oxford University Press.

Robin, J., Rivest, J., Rosenbaum, R. S., & Moscovitch, M. (2019). Remote spatial and autobiographical memory in cases of episodic amnesia and topographical disorientation. *Cortex*, 119, 237-257.

Roediger, H.L., Rajaram, S., & Srinivas, K. (1990). Specifying criteria for postulating memory systems. *Annals of the New York Academy of Sciences*, 608(1), 572-595.

Roediger H.L. & McDermott K.B. (1993) Implicit memory in normal human subjects. In: Boller, F., Grafman, J. (Eds.) *Handbook of neuropsychology.* Vol. 8. Amsterdam, Netherlands: Elsevier

Roediger, H.L., Buckner, R.L. & McDermott, K.B. (1999). Components of processing. In: Foster, J.K. & Jelicic, M., (Editors.) *Memory: Systems, process or function?*. Oxford, UK: Oxford University Press

Rosenbaum, R. S., Köhler, S., Schacter, D. L., Moscovitch, M., Westmacott, R., Black, S. E., ... & Tulving, E. (2005). The case of KC: contributions of a memory-impaired person to memory theory. *Neuropsychologia*, 43(7), 989-1021.

Rosenbaum, R. S., Gilboa, A., Levine, B., Winocur, G., & Moscovitch, M. (2009). Amnesia as an impairment of detail generation and binding: evidence from personal, fictional, and semantic narratives in KC. *Neuropsychologia*, 47(11), 2181-2187

Rubin, D. C. (2022). A conceptual space for episodic and semantic memory. *Memory & Cognition*, 50(3), 464-477.

Rubin, D.C. & Umanath, S. (2015). Event memory: A theory of memory for laboratory, autobiographical and fictional events. *Psychological Review 122(1): 1-23*

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

Salmon, W. C. (1984). Scientific explanation: Three basic conceptions. In *PSA: Proceedings of the*

biennial meeting of the philosophy of science association (Vol. 1984, No. 2, pp. 293-305). Cambridge University Press.

Schacter, D. L. (2022). On the evolution of a functional approach to memory. *Learning & Behavior*, 1-9.

Schacter, D. L. & Addis, D.R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transaction of the Royal Society B,* 362: 773-786

Schacter, D. L., & Addis, D. R. (2009). On the nature of medial temporal lobe contributions to the constructive simulation of future events. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 364(1521), 1245-1253.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: Remembering, imagining, and the brain. *Neuron*, 76(4), 677–694.

Schacter, D. L., & Tulving, E. (1994). What are the memory systems of 1994? In D. L. Schacter & E. Tulving (Eds.). *Memory systems 1994.* The MIT Press

Shadish, W., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin

Sheldon, S. A., & Moscovitch, M. (2010). Recollective performance advantages for implicit memory tasks. *Memory,* 18(7), 681-697.

Sherry, D. F., & Schacter, D. L. (1987). The evolution of multiple memory systems. *Psychological Review*, 94(4), 439.

Spreng, R. N., & Mar, R. A. (2012). I remember you: a role for memory in social cognition and the functional neuroanatomy of their interaction. *Brain Research,* 1428, 43-50.

Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of Learning and Memory,* 82(3), 171-177

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science*, 253(5026), 1380-1386.

Suddendorf, T. & Corballis, M. (1997). Mental time travel and the evolution of the human mind. *Genetic Social and Genetic Psychology Monographs* 123, 133–167.

Suddendorf, T. & Corballis, M. (2007). The evolution of foresight: what is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences* 30: 299–313.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, 2(4), i.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson. (Eds.) *Organization of memory*. Oxford, England: Academic Press

Tulving, E. (1983). *Elements of episodic Memory*. Oxford: Oxford University Press.

Tulving, E. (1985a). How many memory systems are there?. *American Psychologist*, 40(4), 385.

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

Tulving, E. (1987). Multiple memory systems and consciousness. *Human Neurobiology,* 6(2), 67-80.

Tulving, E. (1989). Memory: Performance, knowledge, and experience. *European Journal of Cognitive Psychology,* 1(1), 3-26.

Tulving, E. (1991). Concepts of human memory. In *S*quire, L. R., Weinberger, N. M., Lynch, G., & McGaugh, J. L. (Eds.) *Memory: Organization and locus of change*. Oxford University Press

Tulving, E. (1993). What Is Episodic Memory? *Current Directions in Psychological Science,* 2(3), 67-70.

Tulving, E. (1995). Organization of memory: Quo vadis? In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 839–853). The MIT Press.

Tulving, E. (1999). Study of memory: Processes and systems. In Foster, J. K., & Jelicic, M. E. (Eds). *Memory: Systems, process, or function?*. Oxford University Press.

Tulving, E. (2001a). Episodic memory and common sense: how far apart?. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences,* 356(1413), 1505-1515.

Tulving, E. (2001b). Origin of autonoesis in episodic memory. In H. L. Roediger III, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 17–34). American Psychological Association.

Tulving, E. (2002a). Episodic Memory: From Mind to Brain. *Annual Review of Psychology*, 53(1), 1–25.

Tulving, E. (2002b). Chronesthesia: conscious awareness of subjective time. In Stuss, D.T., Knight, R.C. (Eds.) *Principles of frontal lobe function.* New York: Oxford University Press

Tulving, E. (2005). Episodic memory and autonoesis: Uniquely human? In Terrace, H. S., & Metcalfe, J. (Eds.). (2005). *The missing link in cognition: Origins of self-reflective consciousness*. Oxford University Press.

Tulving, E., & Madigan, S. A. (1970). Memory and verbal learning. *Annual Review of Psychology,* 21(1), 437-484.

Tulving, E., & Lepage, M. (2000). Where in the brain is the awareness of one's past?. In Schacter, D. L., & Scarry, E. (Eds.). *Memory, brain, and belief* (Vol. 2). Harvard University Press.

Weinberg, S. (1976). The forces of nature. *Bulletin of the American Academy of Arts and Sciences* 13-29.

Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy*, 104(12), 639-659

Wheeler, M. A., Stuss, D. T., & Tulving, E. (1997). Toward a theory of episodic memory: the frontal lobes and autonoetic consciousness. *Psychological Bulletin*, 121(3), 331.

White, N. M., & McDonald, R. J. (2002). Multiple parallel memory systems in the brain of the rat.

Neurobiology of Learning and Memory, 77(2), 125-184.

White, N. M., Packard, M. G., & McDonald, R. J. (2013). Dissociation of memory systems: The story unfolds. *Behavioral Neuroscience*, 127(6), 813.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2020). The Tolman-Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5), 1249-1263.

Willingham, D. B., & Goedert, K. (2001). The role of taxonomies in the study of human memory. *Cognitive, Affective, & Behavioral Neuroscience,* 1(3), 250-265.

Wimmer, G. E., & Büchel, C. (2016). Reactivation of reward-related patterns from single past episodes supports memory-based decision making. *Journal of Neuroscience*, 36(10), 2868-2880.

Wimmer, G. E., & Büchel, C. (2021). Reactivation of single-episode pain patterns in the hippocampus and decision making. *Journal of Neuroscience*, 41(37), 7894-7908.

Wimmer, G. E., & Shohamy, D. (2012). Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science,* 338(6104), 270-273.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality.* Harvard University Press.

Wisan, W. L. (1978). Galileo's scientific method: A reexamination. In Butts, R. & Pitt, J. (eds.). *New perspectives on Galileo* (pp. 1-57). Springer Netherlands.