Are Neurocognitive Representations 'Small Cakes'?

Olivia Guest[1,2] and Andrea E. Martin[1,3]

[1]Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands
[2]Department of Cognitive Science and Artificial Intelligence, Radboud University, The Netherlands
[3]Language and Computation in Neural Systems Group, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

In order to understand cognition, we often recruit analogies as building blocks of theories to aid us in this quest. One such attempt, originating in folklore and alchemy, is the homunculus: a miniature human who resides in the skull and performs cognition. Perhaps surprisingly, this appears indistinguishable from the implicit proposal of many neurocognitive theories, including that of the 'cognitive map,' which proposes a representational substrate for episodic memories and navigational capacities. In such 'small cakes' cases, neurocognitive representations are assumed to be meaningful and about the world, though it is wholly unclear who is reading them, how they are interpreted, and how they come to mean what they do. We analyze the 'small cakes' problem in neurocognitive theories (including, but not limited to, the cognitive map) and find that such an approach *a*) causes infinite regress in the explanatory chain, requiring a human-in-the-loop to resolve, and *b*) results in a computationally inert account of representation, providing neither a function nor a mechanism. We caution against a 'small cakes' theoretical practice across computational cognitive modelling, neuroscience, and artificial intelligence, wherein the scientist inserts their (or other humans') cognition into models because otherwise the models neither perform as advertised, nor mean what they are purported to, without said 'cake insertion.' We argue that the solution is to tease apart explanandum and explanans for a given scientific investigation, with an eye towards avoiding van Rooij's (formal) or Ryle's (informal) infinite regresses.

Keywords: representation, computational neuroscience, cognitive map, concept neuron, place cell

> One cannot have a recipe for cake that lists a cake, not even a small cake, as an ingredient.
>
> Fred Dretske (1994, p. 469)

Cognitive and computational neurosciences aim to explain how the human organism creates and uses internal representations of external posited structures (see Box 1; Egan, 2020). Representational content is construed as patterns over neuroimaging read-outs (see Figure 1). And this is taken as a stand-in for what is used by the neurocognitive system to carry out the experimental task (cf. Ritchie et al., 2019; Vigotsky et al., 2024). External posited structures, on the other hand, are experimentally manipulated variables such as the organism's location or orientation in physical space. These structures can also be conceptualised as aspects of stimuli, e.g. properties such as colour or tone, that are used by scientists to infer representational content at play for a cognitive capacity if the task involves, e.g. categorisation or language (Egan, 2020; Schellenberg, 2018). In addition, cognitive and neuroscientific conceptualisations serve as scaffolds for artificial intelligence and machine learning research (e.g. Banino et al., 2018; Kitchin, 1994).

On the one hand, practitioners sometimes subscribe to the relatively innocuous claim that "a representation is a state that carries [empirically evaluable] information" (Ritchie et al., 2019, p. 591). On the other, there is also the much stronger claim that decoding models (statistical models that take neuroimaging data as input and output the experimental variables) are "mechanistically interesting" (Vigotsky et al., 2024; also see: Carlson et al., 2018; Chirimuuta, 2013; Chirimuuta, 2024; Popov et al., 2018; Ross and Bassett, 2024; Zednik, 2014). In other words, practitioners believe that statistical analyses provide a mechanistic account of representation and do not constitute epiphenomenal byproducts that "may not be causally related to perceptual experience." (Cohen et al., 2019, p. 11) Where do such commitments leave neurocognitive research on representations? "How do the posited internal representations get their meanings? [And w]hat is it for an internal state or structure to function as a representation, in particular, to serve as a representational vehicle?" (Egan, 2020, pp. 26–27; Camp, 2007; Hurley, 1998b; Millikan, 1991) Neurocognitive theories of representations promise us answers to these questions, but do they deliver?

In this paper, we explore aspects of neurocognitive accounts of representation that have either remained implicit or underexplored (cf. Ritchie et al., 2019; Vigotsky et al., 2024). We explain how certain metatheoretical properties of neurocognitive conceptions of representation can have a detrimental effect on our scientific thinking (Guest, 2024; also Bennett, 1996; Pylyshyn, 1973, 2002, 2003). To do this, we take the case of the *cognitive map*, a type of posited representation, and explain how if used carelessly commits us to the following frustrating but avoidable properties:

1. an infinite regress in the explanatory chain, which rears its head when we try to explain away cognition using the homunculus — the canonical 'smaller cake' per Dretske (1994) — which we explain in **All ghost, no machine**; and

2. explanatory and computational inertness of representation; where representations like cognitive maps provide no mechanistic nor functional analysis, which we explain in **All form, no function**.

When the nominal computationalist theorises like we describe

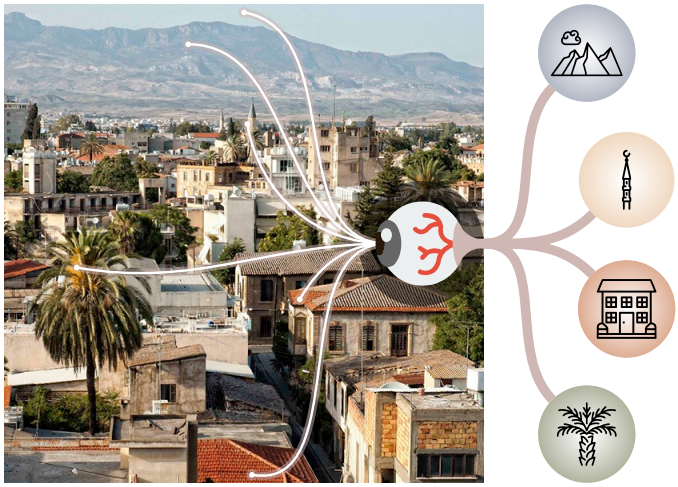


Figure 1: Cartoon depiction of the methodology scientists deploy when they investigate neurocognitive phenomena or capacities. On the left is a scene a participant (represented by an eyeball) might be asked to navigate through. On the right is a series of external posited structures, so-called landmarks in the case of the cognitive map (viz. Maguire et al., 1999): a mountain range, a minaret, a house, a palm tree; which the participant is assumed to be using to navigate. To substantiate the claim that indeed a subset of these landmarks are mentally represented — and therefore, incorporated into a cognitive map — scientists lean on statistical analyses of read-outs from the neurocognitive system, e.g. neuroimaging data. In other words, correlations between the externally posited structures (e.g. landmarks, features of the task or stimuli) and the experimental data (e.g. voxels, single-cell recordings).

above (also see Figure 1), they introduce serious problems in the neurocognitive examination of the hippocampus (viz. Hannula & Duff, 2017) and other brain areas, in the discussion of neurally-based accounts of representation generally, and in the parallelling of connectionist models, i.e. artificial neural networks, with brain, cognition, and behaviour (Guest & Martin, 2023, 2024; Guest et al., 2020). And so when scientists aim to explain, understand, and investigate the human organism's capacity to represent the world, we need to take explicit action and care in order to avoid these traps of our own making.

2 What is a cognitive map?

> The cognitive map has been called an "a priori assumption" that "should be abandoned" (Benhamou, 1996, p. 211), "an unwarranted exercise of anthropomorphism" (Shettleworth, 2010, p. 310), "one of the holy grails of cognition" (Breed, 2017, p. 57), and "one of the most important neuroscientific results in recent decades" (Shea, 2018, p. 113).
>
> Kelle Dhein (2023, p. 62)

In modern mainstream use, *cognitive map* is a phrase that is recruited to cover a broad range of neurocognitive capacities, most often tied to the hippocampus and subserving the capacity of navigation and of episodic memory (see Figure 1; Behrens et al., 2018; Epstein et al., 2017; Jensen, 2006; Maguire et al., 1999; Marozzi and Jeffery, 2012; McNaughton et al., 2006; O'Keefe and Dostrovsky, 1971; O'Keefe and Nadel, 1978; Shettleworth, 2010; Tolman, 1948). Historically, this breadth of reference for the phrase was the case too with Edward Chace Tolman (1948) extolling its virtues to solve inter alia racism, sexism, and bringing about world peace through its use as part of a pedagogical method.[1] Tolman "concluded that the animals must have access to spatial knowledge about the environment, akin to the spatial knowledge obtainable from a map, that could be used to guide behavior in a flexible manner." (Epstein et al., 2017, p. 1504; cf. Simon, 2022)

These cognitive maps are held to be able to represent, or to be representations of, both physical space and non-physical so-called semantic or categorical space (Lisman et al., 2017; O'Keefe & Nadel, 1978). Through its origin with Tolman, the cognitive map as a scientific object has strong links both to his earlier purposive behaviourism (a type of behaviourism that does not rule out cognitive or mental states; Good & Still, 1986; Innis, 1999; Tolman, 1932) and his later cognitive realist views, i.e. "he accepted the notion that a concept such as the cognitive map was a real, albeit unobserved, entity." (Carroll, 2017, p. 181).[2] Perhaps unsurprisingly, the original presentation of the cognitive map came under attack from both operationalists and mainstream behaviourists (viz. Good & Still, 1986; Guthrie, 1935; MacCorquodale & Meehl, 1954). In present cognitive and computational neuroscientific use, however, this type of theory about mental representation passes with little critique (most critique is from previous decades, if not century, and often from outside the computationalist framework: Benhamou, 1996; Bennett, 1996; Jensen, 2006; Mackintosh, 2002; Pylyshyn, 1973; Shettleworth, 2010; Skinner et al., 2003). Indeed, by the 70s, statements such as this pass uncontroversially: "the representational implication of Tolman's 'cognitive maps' is implicit in one form or another in virtually all current cognitive theorizing" (Hammond, 1976, p. 248).

Another important aspect of the contemporary use of cognitive map is its relationship with the hippocampus, spearheaded by John O'Keefe and Lynn Nadel (1978). The cognitive map is called upon to give a more cognitive scientific and less neuroanatomical, or otherwise more theoretical, label to both the representational or functional role of the hippocampus (cf. Konishi, 1986; Mackintosh, 2002) and to support modelling its neuronal mechanisms (viz. Darden, 2006; cf. Chirimuuta, 2018a). And vice versa, the findings of certain cell types in the hippocampal formation (Hafting et al., 2005; M.-B. Moser & Moser, 1998; O'Keefe & Dostrovsky, 1971) enabled the cross-fertilisation and even merger of cognitivism and behaviourism (viz. Thinus-Blanc, 1987, 1996).

So in many ways, and according to multiple experts (viz. Kitchin, 1994; Lisman et al., 2017), the cognitive map appears to furnish the neurocognitive practitioner with all the virtues they ask for (Guest, 2024). Not only is it recruited to explain the contents of the hippocampus in both rodents and humans through analogising them with a map or map-like structure (e.g. GPS system; Craig & McBain, 2015; M.-B. Moser & Moser, 2016), the cognitive map also embodies a proposed mechanistic and functional analysis of both the capacity for episodic mem-

[1] So much for *modern* technoscientific hype needing to be reigned in! Notwithstanding, Tolman was incredibly principled and even accepted being fired for his beliefs on academic freedom (Carroll, 2017).

[2] It is unsurprising if these views appear to be cognitivist from the get-go, since scientific constructs like cognitive maps were developed around the cognitive revolution by essentially proponents of such ideas (Carroll, 2017).

Box 1: A description of the core ideas and terminology to unpick mental representations under computationalism.

Mental representation under computationalism

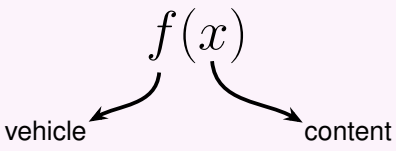
Base computationalism, or just **computationalism**, is the idea that cognition is, in part or whole, explainable through computation. It is an umbrella term for all types of computationalism, and therefore computationalists of all stripes must subscribe to this notion definitionally (Hardcastle, 1995; van Rooij et al., 2024).

Multiple realisability is the idea, and **multiple realisation** is the fact, that the same function can be instantiated by different substrates (Chirimuuta, 2018b; Egan, 2017; Figdor, 2010; Hardcastle, 1995, 1996; Litch, 1997; Polger & Shapiro, 2016; Ross, 2020). For example, a calculator and a person appear to both perform addition, digital and clockwork timepieces both measure time, and a train and a horse both can carry us from A to B. Multiple realisation is core to computationalism.

Naive computationalism is the lack of attention to, ignorance of, or the explicit rejection of the formal repercussions of computationalism, typically manifesting as misrepresenting formal findings, e.g. misapplying the universal approximation theorem (Guest & Martin, 2024), ignoring tractability constraints (van Rooij, 2003, 2008; van Rooij et al., 2019), or neglecting multiple realisability (Guest & Martin, 2023).

Non-naive computationalism bites the bullet and accepts that theoretical computer science must have something to say about cognition under the computationalist paradigm, and that computationalism is a series of beliefs, and not (just) the methodological usage of computers to do our science, e.g. to run models.

In cognitive and computational neuroscience, computationalism is largely, minimally tacitly, accepted. This means that neurons and their assemblies are described as computing, artificial neural networks are widespread as models of neurocognitive capacities, phenomena, and experimental tasks, and generally computationalist vocabulary and methods are often deployed (Guest & Martin, 2021, 2023, 2024). Computationalism binds practitioners to constrained ways of modelling and of metatheoretically adjudicating over models' apparent successes (Guest, 2024).



Under computationalism, capacities can be described in terms of mathematical **functions** (van Rooij, 2008). When it comes to mentally representing, the neurocognitive system can be analysed into representational **vehicle** and the representational **content** (see schematic above; Egan, 2018; Hurley, 1998b; Millikan, 1991). Ideally, emphasis for formal theorising is not on the computationally inert content, which has a more pedagogical or illustrative role as a *gloss*, but on the computationally active vehicle (viz. Egan, 2020, pp. 26–27). For example, in our attempts to formally specify a function (i.e. a cognitive capacity), focussing on subsets of the range (e.g. stimuli) and domain (e.g. actions) while neglecting to specify the operations it performs does not further development of the formal theory (Guest & Martin, 2021).

With respect to **mechanism**, in neuroscience it is often synonymous to substrate (e.g. Darden, 2006; Lisman et al., 2017), but under computationalism a substrate will not, cannot, cut it as a mechanistic analysis. Non-naive computationalists in fact do the opposite, they — through multiple realisability and other principles — completely rule out using substrate and mechanism interchangeably. Any computationalist within cognitive neuroscience will need to thread this needle.

ory and for spacial navigation, e.g. by "providing a spatial and temporal framework" and "neural network mechanisms", and through "computing distances and angles" (Lisman et al., 2017, p. Reid and Staddon, 1997). A small leap from the above gives us all-encompassing versions of the cognitive map, i.e. as a synonym for *all mental representation* — both deflating and inflating the account in various ways (viz. Aly & Turk-Browne, 2017; Bennett, 1996; Downs & Stea, 1973; Gallistel, 1989; Jacobs, 2003; O'Connor, 2019; Schlichting & Preston, 2017; Shettleworth, 2010; Thinus-Blanc, 1987, 1996). This betrays the important fact that "[d]espite using the notion of cognitive maps in a loose and sometimes divergent way, [scholars] agree on the need to identify different forms of cognition, their nature, and how they work and interact with our language-like systems." (Aguilera, 2016, p. 350)

Based on all this, it appears as if explanandum and explanans are interwoven: does the hippocampus supply the mechanisms, or is it what needs to explained? Can mechanisms be supplied by neuroanatomy — is this possible in principle that a scientific explanation can be the thing itself? Does the cognitive map as a theory deliver? How does it offer computational explanation? The next few sections analyse potential answers to these questions, demonstrating how cognitive map as a case study serves us well to understand slippages between the theory sec and the pre-theoretic object of theory, i.e. the label for a cognitive capacity.

3 All ghost, no machine

> The charts do not control the movements of the ship; hence this particular scientific model is seriously lacking in predicting the rat's movements.
>
> Edwin Ray Guthrie (1935, p. 199)

Table 1: A selection of related neurocognitive scientific entities about representation in the brain and their quasi-theoretical properties to help us understand the case study in question: the cognitive map. Concept cells (first row) are held to exist in inter alia the medial temporal lobe and the hippocampus (Calvo Tapia et al., 2020; Quiroga, 2012, cf. Bowers, 2009). While spatial cells (second row), e.g. place and grid cells, are present in a brain area known as the hippocampus and its related structures, and have to do with the cognitive capacities: of navigation; of representing spatial aspects of the organism in its environment; and of episodic memory, of autobiographical sensations, feelings, and events (Marozzi & Jeffery, 2012). In contrast to these examples of more straightforward phenomena or explananda: concept and spatial cells (first two rows) is the cognitive map (last two rows), which some consider a theory or explanans (third row). Each of these entities is analysed into its status as explanandum or explanans (second column) and its mechanistic aspects and functional role (third and fourth columns; recall Box 1).

Scientific entity	Status	Mechanism	Function
CONCEPT CELL as-is	EXPLANANDUM: something we seek to explain, a phenomenon, an observation, a pattern in data.	The "predictable characteristics of an abstract, memory-based representation" (Connor, 2005).	To represent concepts, i.e. invariant properties of stimuli; "building blocks for declarative memory functions." (Quiroga, 2012)
SPATIAL CELLS as-is	EXPLANANDUM: something we seek to explain, a phenomenon, an observation, a pattern in data. There "are the gallimaufry of spatial cell types" (Marozzi & Jeffery, 2012, p. 942), defined based on involvement, firing patterns, in navigation- or spatial-related behaviours.	"[T]heir spatial firing structure reflects computations internally in the system" (E. I. Moser et al., 2008, p. 69) , e.g. "grid-like processing" of navigation in the entorhinal cortex (Horner et al., 2016); a spandrel (viz. Gould & Lewontin, 1979) of a clustering-like process (Mok & Love, 2019).	To represent spatial properties of the agent, of the experiment, or of the environment generally; "these cell types interact to form a cognitive map, and [...] may cooperate in service of both spatial and episodic memory." (Marozzi & Jeffery, 2012, p. 939).
COGNITIVE MAP as-is	EXPLANANS: something that does (or aids in) the explaining, a "metaphor" (Buzsáki in Lisman et al., 2017; Stachenfeld et al., 2017), "hypothesis" (Dhein, 2023; Thinus-Blanc, 1996), "theory" (Bouchekioua et al., 2021; Fenton, 2024; Maguire et al., 1999; O'Keefe, 1994; Tolman, 1959; Wikenheiser & Redish, 2015)	Concept and spatial cells are the cognitive map's "neurobiological substrate" (Farzanfar et al., 2023) or "neurophysiological basis" (Bouchekioua et al., 2021); "the cognitive map is supported by the locale system, a cognitive module located in the hippocampus of vertebrates." (Shettleworth, 2010, p. 297)	The cognitive map is "a 2- or 3-dimensional vector space, on which navigation-relevant vector functions are defined." (Langille and Gallistel, 2020; also Peer et al., 2021); "a model detailing the structure of a decision-making environment [in order] to predict the impact of action choice on potential future rewards." (Moran et al., 2021)
COGNITIVE MAP as should be discussed	EXPLANANDUM: something we seek to explain, a cognitive capacity (Newcombe & Liben, 1982; Thinus-Blanc, 1987; Weisberg & Newcombe, 2018); "[the] ability to return to a goal by processing the location-based (site-dependent) information provided by the current apparent configuration of landmarks" (Benhamou, 1996, p. 201).	WORST CASE: no mechanism can be proposed because other explananda (first and second rows, here), as well as phenomena and observations, cannot play this role, and models of those phenomena are not models of the cognitive map per se (recall Box 1). BEST CASE: see Table 2.	WORST CASE: no functional role can be proposed because a cognitive capacity exists in and of itself (recall Box 1). e.g. the function of the capacity of vision is to perform vision. Vision, i.e. seeing, is to vision, i.e. the capacity, what pump blood is to heart. BEST CASE: see Table 2.

Cognitive maps appear to be a specific case of an accepted but often unacknowledged norm in computational (neuro)cognitive modelling and artificial intelligence, in which the scientist inserts themselves or other humans' cognition into their own models because otherwise models do not perform as advertised e.g. if a computational model is intractable (van Rooij, 2003, 2008; also see Schaeffer et al., 2022). Computational (neuro)cognitive modelling, especially in connectionist tendencies (Guest & Martin, 2023, 2024), so-called neuroAI (Zador et al., 2023), and artificial intelligence broadly, make use of human-labelled data, reinforcement learning from human feedback, "explicitly encoding *human priors* into the training process" (Ilyas et al., 2019,

p. 9), and other human-in-the-loop techniques. This is direct human labour that is obfuscated and repackaged as an automated machine process: human cognition masquerading as a formal model. This human-in-the-loop, "ghost in the machine" (Ryle, 1949), technique is a long-standing thread within the automation of labour and the scientific modelling of cognitive capacities that has its roots in at least the pre-modern period and the industrial revolution (Bainbridge, 1983; Erscoi et al., 2023; Jones, 2021; Pfaffenberger, 1988; Taylor, 2018; van der Gun & Guest, 2023).

Importantly, if too much human-in-the-loop occurs a theory or model is likely less useful than might seem, or even vicious.[3] This is because a core requirement for such scientific objects is to function the same regardless of who controls their running (Guest, 2024). If a non-interested party versus a proponent being in charge of the model changes its ability to perform, alarm bells should ring. That is to say, human-in-the-loop in the general case is inevitable in science. Science is not only a uniquely human activity, but also only people can carry out scientific theorising, i.e. it requires a human-in-the-loop even when parts may seem to be automated (P. Rich et al., 2021; van Rooij et al., 2024). But this requirement does not license portraying a mechanistic or otherwise account as a theory that stands without human guidance for its critical explanatory parts.

This leads us to the case of the *homunculus*: a miniscule human-in-the-loop, originating from folklore and alchemy (Frietsch, 2021; Murase, 2020). In the 17th century, some believed that "sperm cells contained perfectly formed homunculi that were developed into life-sized infants in the matrix of the womb." (Campbell, 2010, p. 5) In its most literal form, in our context, a homunculus is a miniature human who resides in the skull and performs cognition; or performs parts of cognition currently escaping formalisation or otherwise scientific description or understanding. Uncontroversially, such a claim is problematic, as it leads to an infinite regress, also known as Ryle's regress (Bäckström & Gustafsson, 2017; Hornsby, 2011; Nizami, 2018; Ryle, 1949; Tanney, 2011). Appealing to the homunculus does not serve us well scientifically if we are to explain what drives cognition, what cognition is, and how cognition manifests because the homunculus appears to be merely a 'smaller cake' (recall quote from Dretske, 1994; also see: Dawson, 2013; Figdor, 2018a, 2018b; Fodor, 1968).

The criticism of infinite regress for homunculus-like theorising holds unless perhaps if grounded directly, i.e. if framed as an observation or a phenomenon. Nonetheless, even such cases require scientific theorising to be understood and explained , i.e. the scientific entity is an explanandum, something we seek to explain. Such a case appears to be that of *cortical homunculi*, e.g. somatosensory or motor homunculi, which are brain regions that have areas dedicated to sensory input from the skin or internal organs or movement (Dall'Orso et al., 2018; Schmahmann, 2019; Wright and Foerder, 2021; although also controversial Catani, 2017). These brain areas have certain homunculus-like properties, i.e. are described as representing or mapping parts of, e.g. the body in the case of the somatosensory homunculus (Penfield & Boldrey, 1937). Although some aspects remain controversial, such as the gendered aspect, which was only addressed recently (Wright & Foerder, 2021), cortical homunculi remain "largely valid to this day." (Catani, 2017, p. 358; cf. Gordon et al., 2023) In addition, many other such 'topographic maps' appear to emerge in the brain, such as retinotopic and tonotopic maps (Kaas, 1997; cf. Catani, 2017). "The functional role of these maps is difficult to establish, because the coding of spatial information may not be the factor determining their topographic organisation." (Konishi, 1986, p. 163) Such statements betray the scientists' desire to understand which functional roles may be filled by these maps and which proposed mechanisms may be at play that give rise to such topographic maps, that give rise to observations that indicate that, as Susan B Udin and James W Fawcett (1988) describe it, "sensory inputs to the central nervous system (CNS) are topographically arranged." (p. 289) In other words, these maps are not taken to be explanatory, but something to-be-explained, an explanandum.

Another such case of a proposed map-like representation is that of *concept cells* (see first row, Table 1), which "are highly selective neurons [in e.g. the hippocampus and the medial temporal lobe,] that seem to represent the meaning of a given stimulus in a manner that is invariant to different representations of that stimulus." (Reddy and Thorpe, 2014, p. 249; also see: Bausch et al., 2021; Calvo Tapia et al., 2020; Quiroga et al., 2005; Quiroga, 2012; cf. Bowers, 2009) Such cells have also been called 'grandmother' (originally as a joke) or 'gnostic' cells (Gross, 2002). These neurons are also described as performing a mapping, creating a representation, of the task in the abstract that the participant is being asked to perform, i.e. they appear to selectively respond to the same concept or category. With respect to accommodating such findings under a connectionist framework, the logistics of assigning single neurons to each concept, known as localist representations, seems impractical and violates certain tenants of connectionism, especially the parallel distributed processing (PDP) variety (Guest & Martin, 2024). Notwithstanding such localist (e.g. one-hot encoding; Harris and Harris, 2015) as opposed to distributed (Eckhardt, 2004), representations are commonly used by connectionists in their artificial neural network models of capacities (e.g. of vision Krizhevsky et al., 2012; cf. Lycan, 1991).

Additionally, taking the map-likeness even further, "are the gallimaufry of spatial cell types" (Marozzi and Jeffery, 2012, p. 942; see second row, Table 1): so-called place (Dostrovsky & O'Keefe, 1971), head-direction (Ranck, 1985; Taube et al., 1990a, 1990b), border (Barry et al., 2006; Lever et al., 2009; Solstad et al., 2008), and grid (Fyhn et al., 2004; Hafting et al., 2005; Rowland et al., 2016) cells, and more (Grieves & Jeffery, 2017), collectively *spatial cells* (viz. Bush et al., 2014; Grieves & Jeffery, 2017; Jeffery et al., 2018; Marozzi & Jeffery, 2012). This is in contrast to the topographic maps mentioned previously, which are forwarded as representations, maps, or tracings of what the sensory surfaces are experiencing (viz. Camp, 2007; Hurley, 1998a; Schellenberg, 2018) — place cells are not topographically arranged, "adjacent cells do not necessarily have adjacent place fields." (Marozzi & Jeffery, 2012, p. 939).

> [Instead] a place cell fires whenever an animal ventures into, say, a particular region of an enclosure, no matter which direction it approaches from, whether the lights are on or off, etc., and so the signal is remarkably stable no matter what the animal is doing or perceiving.
>
> Kate J Jeffery et al. (2018, p. 96)

As in the case of the concept cells, these spatial cells are described as a representation of the experiment, but in this case not

[3] A useful example of such a vicious theory is physiognomy, which is considered pseudoscientific but nonetheless has been making a comeback in machine learning systems that perform facial recognition (Andrews et al., 2024; Guest, 2024; Spanton & Guest, 2022).

that of the experimental task's abstracted concepts or categories, but that of the experimental environment's physical space or the rodent or human participant's relationship to it, e.g. some mapping of the maze in which we place a rat (Egan, 1999; Epstein et al., 2017). Furthermore, they are held to "provide the building blocks of a cognitive map, including direction, distance, and boundaries" (Marozzi & Jeffery, 2012, p. 942), framing these cell types as a substrate or a mechanism (recall 1) for the cognitive map. Thus, these two proposals for neural representation of space and concept are proposed to be two sides of the same coin and grouped together under the heading 'cognitive map'. In other words, "[t]he spatial positioning system supported by these cells is often taken to be a model system for understanding how the brain processes high-level cognitive information." (Epstein et al., 2017)

Unlike the theoretical content of topographic cortical maps, which are phenomena to be explained, the cognitive map appears to be a very different kind of proposal for the neurocognitive system. In other words, so-called 'cognitive maps', such as spatial and semantic neurons or brain areas are proposed as an explanatory theory for rats', human's, and other species', cognitive capacities. This leaves open the problem of who, or what is reading these maps — and who or what is making these maps? Recall cognitive maps are "akin to the spatial knowledge obtainable from a map" (Epstein et al., 2017, p. 1504). The knowledge obtained from an actual geographic map is not found in nature, directly falling onto our sensory surfaces; if it was, then who needs maps? It is found as the output of cognition, presented on paper or a screen, after another human (or we ourselves) make the map, and as a function of us reading the map. And making maps involves many computationally hard problems, not solvable by just observing nature (viz. E. Rich et al., 2008).

To recapitulate, the cognitive map as a theory does not furnish us with an explanatory account of representation nor indeed of behaviour (viz. Guthrie, 1935; cf. Behrens et al., 2018). In fact quite the opposite, it creates — like the original homunculus as an explanation for sexual reproduction — an infinite regress in the explanatory chain. Amusingly, not only downwards, homunculi-inside-homunculi, but upwards: with so-called meta-learning and "meta-maps" (Ambrogioni & Ólafsdóttir, 2023). These theoretical contortions cause minimally the scientific problem of not explaining while seeming to explain, violating an important desideratum under computationalism dubbed "explaining without assuming" by Marieke Woensdregt et al. (2024), falling under the theoretical vice of underexplaining (viz. Egan, 2020). And thus obfuscating that: we do not have an account for representations in this context; we have inserted a ghost into a machine (i.e. model, theory, or account) in which we promised there was no ghost; and ultimately, we looped in a human, ourselves, to do the dirty work. This is what the "viciously circular" (Fodor, 1968; also see Dawson, 2013) homunculus does to our theorising, i.e. forces us into a type of scientific dishonesty, wittingly or not.

Furthermore, the cognitive map underexplains representational capacities because it presupposes somebody is reading the map. In other words, what we present as formal theory, or otherwise complete, is a description of data that requires much more specification work to function as an explanation at all. And this is the case even if we have a candidate substrate, i.e. the brain as a whole, or various postulated kinds of neurons (recall 1). Notably, this holds for all neurocognitive theories: the substrate is largely known, uncontested. Others have noticed this too, e.g.

> neurons that respond to [environmental] events cannot intrinsically represent anything because they cannot relate or compare their firing to something else. In contrast to the observing human [e.g. the scientist], neurons in, for example, sensory areas driven by environmental signals alone cannot ground their activity to anything meaningful. Grounding refers to the ability of the brain's circuits to assign meaning to changes in neuronal firing patterns that result from sensory inputs.
>
> György Buzsáki and David Tingley (2023, p. 193)

Additionally, of note is that it appears some avoidance of defining these term is present in papers, e.g. while the abstract claims "exploration gives rise to a cognitive map" (p. 191) nowhere else in this paper is the phrase 'cognitive map' used again (Buzsáki & Tingley, 2023). This parallels that "GPS" is not mentioned again outside the title *Navigating the circuitry of the brain's GPS system* by Craig and McBain (2015).

Maximally, this form of obfuscated human-in-the-loop that we have outlined herein also causes problems with respect to scientific pedagogy and practice, making neurocognitive practise poorer if taken as a standard for a typical theoretical account of representation (viz. Guest, 2024; Jensen, 2006). That is to say,

> a psychological theory [that] purports to explain a behavior by postulating an unexplained intelligent process, [...] begins an infinite regress of homunculi within homunculi [and] should be rejected on the grounds that it begs the question
>
> Gualtiero Piccinini (2007, pp. 112–113)

This is in fact the central warning of Iris van Rooij's *The Tractable Cognition Thesis* (2003, 2008) because it remains the case such a problem can occur even in fully formalised theoretical objects such as computational models. When "cognitive tasks that are performed effortlessly by humans are presently being modelled by computationally intractable functions" (van Rooij, 2008, p. 965, also: van Rooij, 2003), the models are falling prey to a formal subtype of Ryle's regress that we dub van Rooij's regress (see Box 2). And so, under computationalism, the practitioner must "concede that [...] the capacities are incorrectly modelled (i.e., [the presence of the regress or the ghost-in-the-machine requires a theoretical revision])" (van Rooij, 2008, p. 965). Or alternatively, the practitioner may conclude that "the hypothesized capacity [e.g. the cognitive map] does not exist at all" (van Rooij, 2008, p. 952). These explanatory regresses are held to be both avoidable and extremely important to avoid under computationalist and cognitivist common sense (Tanney, 2013).

4 All form, no function

> The term "cognitive" refers to an activity, a dynamic process. The word "map" is essentially static; it suggests a static image of the real world. Tolman has said too much and not enough. Too much for his hypothesis to be forgotten. And not enough for his subtleties to outweigh his overstatements.
>
> Catherine Thinus-Blanc (1987, p. 4)

Box 2: The core problems of theorising under computationalism (Box 1), or indeed any formal framework, is that regresses such as those described below inevitably crop up. These cases require acknowledgment and demarcation, and furthermore deep discussion to avoid problems further down the line.

Two types of explanatory regress

Ryle's regress states "if, for any operation to be intelligently executed, a prior theoretical operation had first to be performed and performed intelligently, it would be a logical impossibility for anyone ever to break into the circle." (Ryle, 1949, p. 31; Tanney, 2011, 2013) Theories can result in an infinite regress if they violate the desideratum of "explaining without assuming" (Woensdregt et al., 2024).

Van Rooij's regress is a formal subtype of Ryle's regress. It occurs when "cognitive tasks that are performed effortlessly by humans are presently being modelled by computationally intractable functions" (van Rooij, 2008, p. 965; also: van Rooij, 2003).

Both explanatory regresses can be — and in fact often are — imperceptible to the scientific practitioners involved in the use and deployment of computational accounts and models. In the formal case, especially because of the heightened epistemic status of computational models that appear to produce desirable outcomes, criticisms tend not to stick (Guest, 2024). And so in the case of intractability, or other forms of naive computationalism (see Box 1), infinite regress is hidden in plain sight because code compiles, computational models run, the output correlates with some desirable metric (see Figure 1), the human-in-the-loop does the work: the ghost not only runs the machine, but constitutes it as such (also see the frame problem, Shanahan, 2016). In this context, the success-to-truth inference is fallacious; claims such as "if the program works then we can be certain that all homunculi have been discharged from the theory" (Dennett, 2017, p. 81) are wrong prima facie (see Nizami, 2018, specifically on Dennett, 2017; and generally: Chirimuuta, 2021; Guest and Martin, 2023, 2024; Morris, 1991; van Rooij et al., 2024). The trap of the success-to-truth inference ensnares many (from Attneave, 1961 to Dennett, 2017) because a recipe that lists a cake as an ingredient still results in a cake (recall opening quote; Dretske, 1994). If one needs a cake, and takes the first and only step of purchasing one from a confectioner, then one will certainly have the desired output. Such an account provides no theory for how the capacity of cake making is realised; it merely has superficially appropriate output. Thus providing neither mechanistic nor functional understanding for the capacity at hand, be it cake-making or cognition broadly.

Resolution may be out of reach (P. Rich et al., 2021; van Rooij et al., 2024) using modern methodologies and frameworks, but that is not licence to halt all cognitive scientific theorising nor to assume formalism is useless nor that anything goes and that success-to-truth is the only way forward. Quite the opposite: if our formal tools and proposed theories give way to van Rooijan and Rylean regresses understanding more about why we as a field fall victim to such sleights of hand is protective and productive. Our metatheoretical calculus, our adjudication over our theories, cannot contain success-to-truth under computationalism without risking the whole enterprise (Guest, 2024; Guest & Martin, 2023). On the contrary, it should be explicitly embracing non-naive computationalism (recall Box 1). Catherine Thinus-Blanc (1987) underlines: "These issues will not be solved in a day. This does not mean however that we should be unaware of them or be crippled by the difficulties involved instead of facing the promising challenge of inter-disciplinary exchanges." (p. 15)

That is to say, we are endangering our science both in the general case that Gilbert Ryle warns against (Bäckström & Gustafsson, 2017; Hornsby, 2011; Ryle, 1949; Tanney, 2011, 2013, 2022) and the specific formal case that Iris van Rooij (2003, 2008) proves is at play under computationalism (recall Box 2). In this section, we tackle questions such as: What do proponents advertise cognitive map as offering? What do practitioners get when they deploy cognitive maps? What neurocognitive or computationalist mechanisms or functions does it furnish us with?

Cognitive and computational neuroscience deals in two broad types of function (recall Box 1). First, it attempts to describe or delimit the *functional role* an entity plays, which is the purpose of an entity within the system, e.g. the heart pumps blood (also known as natural function; Dretske, 1994; Guest & Martin, 2023; Kristan & Katz, 2006; Millikan, 2021; Schellenberg, 2018). Second, it tries to formally capture the *mathematical function* — what input-output mappings hold for a system that performs computations — i.e. "the interpretation of a computational system should connect the formal apparatus of the theory with its pre-theoretic explananda." (Egan, 1999, p. 183; also see Blokpoel, 2018; Guest and Martin, 2021; Hardcastle, 1996; van Rooij and Baggio, 2021)

Proponents of the cognitive map, as described, promise a functional explanation (recall section 2, **What is a cognitive map?**), i.e.

> Tolman conceived of cognitive maps as extending generally to mapping life's experiences in any behaviorally relevant domain. He conceived the function of these maps as organizing specific events in systematic fashion appropriate to the dimensions of the relevant context, and he argued strongly that the function of cognitive maps is to support expectancies and planning of behavior to obtain sought goals.
>
> Daniela Schiller et al. (2015, p. 13909)

It appears as if the only thing provided by including cognitive map here is: First, the idea of a map-like representation that is

held to exist because behaviour appears to be like that of somebody reading a map just like how it appears to be the case that a small human exists that could be controlling our cognition. Second, a synonym for the contents of the hippocampus, the entorhinal cortex, and other related structures. For example, "hippocampal neural representations can be thought of as cognitive maps" (Ambrogioni & Ólafsdóttir, 2023, p. 702). However, such a label is not necessary for the mechanistic nor functional understanding of a brain area. And so, just like how the contents of random access memory, or any specific subset of the input-output pairs of a function, are not what is relevant to describing, specifying, and ultimately understanding what these systems do, the same goes for labels or what are called contents, the cognitive map (recall Box 1). For example, notice how when describing the goal of such research, as "unravelling the circuitry of the hippocampal formation navigation system" (Craig & McBain, 2015, p. 737), it is apparent that phrases like "navigation system" perform only the service of labelling the capacity and synonymising it with the hippocampus and related brain structures. The same usage pattern can be seen here: "Spatial navigation is thought to be guided by the internal representation of spatial relations in a specific environment, referred to as a cognitive map" (Farzanfar et al., 2023, p. 64). Finally, here is another example, wherein the cognitive map is deployed in framing empirical findings:

> if both spatial and temporal inferences are driven from the same cognitive map, the distortion should similarly affect sketch-maps and travel-time estimations. But if the temporal and spatial aspects of cognitive map are represented or processed separately, distortions on temporal and spatial expressions may dissociate. [...] In conclusion, we found dissociation between effects of familiarity on the spatial and temporal estimations of an environment, which we suggest may relate to differences in temporal and spatial tuning of cognitive maps or the speed of accessing source memories.
>
> Anna Jafarpour and Hugo Spiers (2017, p. 12 & 16)

Cognitive map is then either a label or synonym with the capacities under study, navigation, planning, and so on, or a label or synonym for the hippocampus or any collection of relevant neural substrates, hence mentioning "tuning" in the extract above. This synonimising role was also noted in MacCorquodale and Meehl (1954, p. 30): "Tolman has shown a tendency to restate the system by revising its vocabulary (i.e., 'sign-gestalt-expectations' become 'cognitive maps')". But a synonym is neither necessary nor sufficient for building a theoretical account.

Experimentalists wish to use cognitive map to enrich our understanding through framing their results — a sensible request to lay at the feet of a purported theoretical construct. But this means, by definition, we are falling short of what is required of us. In other words:

> Cognitive scientists attempt to model such capacities by constructing precise characterizations of the hypothesized inputs and outputs of cognitive capacities as well as the functional mappings between them. This is what David Marr (1982) called the computational-level theory of a cognitive process.
>
> Iris van Rooij (2008, p. 939)

But if cognitive map is the name of a proposed capacity, then what is it adding in of itself to functional or mechanistic descriptions and understandings of the cognitive system? Is it delivering explanatory power? If cognitive map is a synonym of hippocampus, what is it adding? It is to-be-explained, and neither a sufficient nor necessary part of an explanation. Quite the opposite. Others have noticed this too, e.g.

> For a cognitive map to be useful, the organism must have a mechanism for connecting map coordinates to fixed aspects of the environment that can be identified by perceptual systems. [...] A second requirement for a cognitive map to be useful is that it must include a mechanism for planning a route to one's destination.
>
> Russell A Epstein et al. (2017, pps. 1506–1508)

These are yet to be found. And claims such as those that some make along the lines of place cells provide "neural network mechanisms" (p. 1437) or a "neuronal embodiment" (Lisman et al., 2017, p. 1444) for the cognitive map do not hold water because these are phenomena. The thing to-be-explained cannot be explained by another thing to-be-explained. So when we see in literature claims such as, Alexandra O. Constantinescu et al. (2016, p. 1465) claiming that "[t]he ability to interact with knowledge in this flexible and generalizable fashion is the central advantage of maintaining an explicit cognitive map", we must ask what do these framings of experimental results offer us? For a cognitive map to be a useful theory, we must know where such map-like structures come from and what computations they perform?

So even if mounting correlational evidence supports the idea of map-like constructs being readable off the neurocognitive system, this does not mean it is in and of itself evidence for a cognitive map (for analyses of this argument, and why it is problematic, see: Carlson et al., 2018; Chirimuuta, 2013; Cohen et al., 2019; Guest & Martin, 2023; Popov et al., 2018; Ritchie et al., 2019; Ross & Bassett, 2024; Vigotsky et al., 2024; Zednik, 2014). "The semantic interpretation of these states in the envisioned models would play a purely heuristic role, allowing us to keep track of what the network is doing" (Egan, 1995, p. 183). This is "because [these so-called representations] play no characterizable causal roles in connectionist models" (Egan, 1995, p. 185).

This framing of the existence of correlations — between what falls on the sensory surfaces or between neural or neuroimaging data and the structure of the experimental task participants are carrying out — as somehow constituting a theory for their existence, that a cognitive map is explanatory, is problematic. The map is being created and read by the *human-in-the-loop*, the scientist. In and of itself, the map explains nothing, it is something the scientist should seek to explain. A copy of the environment or even a richer redescription of it, uncovered by correlation, is not an explanatory account of how "brain systems and computations support concept learning, memory, and spatial navigation." (Mok & Love, 2019, p. 2) So to recruit such a metaphor — of the map — is to ignore the thing to be explained, shifting to the infinite regress wherein the explanation is relegated outside the system under study (e.g. spatial navigation) even though it is the capacity we aim to explain and understand.

Cognitive maps, like other neurocognitive representations, are often evaluated using (2nd order) isomorphisms (correlations over correlations). This is problematic as such isomorphisms do not evaluate everything or indeed much of anything,

because they confuse the map for the territory and invite fallacies of causation, or the attribution of causation based on correlation. They can only tell us that the current task demands are isomorphic to an ideal that is hand-selected or is at least similarly requested from another source of data (recall Figure 1). If we think of cognition or the brain as something close enough to a universal Turing Machine, i.e. if we are computationalist, a match says only that our postulated map is possible, but not how it is used or arises. Something which we could a priori know: the task contains the however derived correlation or tracing, the external posited structures, by definition, that is how we found that map to begin with. Finding an isomorphism with the organisation of our home and the so-called organisation of our brain, e.g. knives and forks are stored closer to each other than they are to bedsheets, means very little. Meaningful findings involving isomorphisms exist, indubitably — but they are not coherent theories, almost entirely useless on their own (recall, e.g. the somatosensory homunculus). The cognitive map does not even constrain the space of possible maps, importantly, it merely states maps are somehow computed/produced by the neurocognitive system. In contrast, basic computationalist assumptions do offer useful constraints, but the map adds none to them to aid us in theorising.

Furthermore, for a simple case of understanding why reading something off an entity is not relevant to the entity's representational powers, consider the following case: A brick in a wall exposed to the elements can be subject to measurement of its temperature and moisture, and these measurements will correlate with the weather. Does this correlation allow us to conclude that the brick represents the weather? If yes, then it is possible one is either a panrepresentationalist, i.e. everything represents. And under the computationalist flavour one could thus ascribe to pancomputationalism, i.e. everything computes something more than the identity function. Alternatively, one does not believe in representations at all, and therefore one allows this absurd conclusion as a reductio argument.

> The [person] on the street acknowledges that minds are rather mysterious, but [they are] definitely sure that a mind is something that you either have or you haven't. Bricks haven't.
>
> Edwin Ray Guthrie (1935, p. 1)

Trying to neither fully deflate representation nor assign everything representational abilities, neurocognitive scientists likely want to be very careful when they conclude from brain read-outs that representation is indeed taking place. In this context, there is nothing that licenses correlational read-outs to be indicative of the nature of, structure of, or even existence of, representations, map-like or otherwise.

We are forced to conclude that the cognitive map, in its current instantiation, is frustratingly computationally inert (Rescorla, 2009; cf. Aguilera, 2016, 2018; Camp, 2007; Rescorla, 2020). It is either a label for a capacity as we see in developmental investigations of children's abilities, or a synonym for the representations (however deficient) themselves that are used by the broader capacities of e.g. navigation, route planning, and so on. And this is why it is not a proper theory, and certainly not a computational one, and never can be without drastic additions (recall Box 1 and Table 1). "The theory proper comprises a specification of the function (in the mathematical sense) computed by the mechanism[. Cognitive content] is ascribed to facilitate the explanation of the relevant cognitive capacity." (Egan, 2020, p. 33) She goes on to elaborate:

> [T]he computational theory proper can fully explain the interaction between organism and environment, and hence the organism's success, without adverting to cognitive content. The [cognitive content] characterizes the interaction between the organism and its environment that enables the cognitive capacity in terms of the former representing elements of the latter; the theory does not.
>
> Frances Egan (2020, p. 34)

The computational theory is tasked with explaining the cognitive capacity, i.e. how does the cognitive system achieve performance in, e.g. navigation. The cognitive content, such as the cognitive map, does not and cannot do this if it is a representation of the environment, be it the abstract properties: of the experimental task, e.g. concept cells (recall Table 1), or of (some posited structure of) what falls on the sensory surfaces, e.g. light intensity (recall Figure 1).

Forwarding the cognitive map as-is as a theory of *how* the brain represents — i.e. over and above evidence *that* it represents — offers little above a meagre piece of evidence that the neurocognitive system might have some tracking of the experimental task, because even this is not certain as such readouts do not guarantee such representations are formed or used (also see Mackintosh, 2002). As part of a theory the cognitive map is minimally unfinished, just the gloss per Frances Egan (2020) and not needed for a computational theory, and maximally confusing, since it offers nothing theoretically. Computationalists already commit theoretically to the brain having representational content, which is why we are on the hunt for a theory on how this is done (viz. Hardcastle, 1996).

5 Here be dragons

> The human brain is, by all accounts, the most complex and wonderful object in the world. But so far we are left with no reason to suppose that an answer to the question about why our beliefs line up with our actions is one that can be given by looking at second-order properties that supervene on matter that is to be found inside the agent's skull.
>
> Julia Tanney (2011, p. 7)

Frances Egan (2018) opens with: "Much of computational cognitive science construes human cognitive capacities as representational capacities, or as involving representation in some way." (p. 247) Indeed, many computationalist connectionist theories and models — from classical models of the hippocampus (e.g. Zipser, 1985) to models of cognitive capacities such as categorisation, vision, language (see examples in Guest & Martin, 2024) — all make use of representation (Egan, 2010; cf. Lycan, 1991). Since 'representation' as a cognitive scientific term undergirds so much of our science, we should strive to untangle it from harmful conceptualisations. Alarm bells should ring in cases where connectionist accounts capture "critical behaviours[, but these] are driven by statistical regularities in the model's input" (Guest et al., 2020, p. 293); the content-vehicle distinction is being violated (recall Box 1). Herein, the cognitive map has

Table 2: Potential (re)conceptualisions for 'cognitive map' to avoid the traps. In the column 'Taken to be' are the types of potential non-mutually exclusive uses of cognitive map. The darker the box, the more work the contents imply for the theoretician with respect to both formalisation and verbal theorising. The 2nd and 3rd columns give examples and constraints imposed as a function of the scientific status, e.g. 'cognitive map taken as theory' binds us to constructing a theory proper, a representational vehicle (recall Box 1).

Taken to be	Examples	Constraints
PHENOMENON	To the experimenter it looks like an organism is reading a map to navigate.	None because within reason, anything can look like anything. It definitely *looks like* the Sun goes round the Earth. See relevant analyses by Dhein (2023) for insects' cognitive maps.
CAPACITY	A synonym for the cognitive capacity of mental representation, memory generally, or specifically of navigation and spacial cognition.	Computationalist methods and analyses can be employed (viz. Blokpoel, 2018; Guest & Martin, 2021; van Rooij et al., 2019). Primal is the need to explain how this is different from extant named capacities, e.g. navigation, spatial learning, episodic memory.
FUNCTIONAL ROLE	The hippocampus, or another candidate brain region, computes cognitive maps as part of its raison d'être. And so 'cognitive map' is to hippocampal formation what 'pumps blood' is to heart. This can be teleological (e.g. Millikan, 2021).	The broader theory must explain if this is a unique functional role and specify if functional roles are definitive for a given brain region (see Mackintosh, 2002). Does computing cognitive maps define if a brain region is part of, e.g. the hippocampal formation in the same way that an organ (even if artificial) which pumps blood is (therefore) a heart? Functional role may therefore be unintuitive or less useful for the cognitive and computational neuroscientist.
SUB-THEORY	A synonym for (a subtype of) 'mental representation' with a commitment to map-like content. A static representation of, e.g. relationships between landmarks.	The broader theory must: explain why map-like and not, e.g. linguistic representations (see Aguilera, 2016, 2018), are at play; avoid descent into such gloss (content details) remaining a stand-in for formal theory because content is not vehicle (recall Box 1; Egan, 2020). Coexistence with non-map-like representations is possible (Aguilera, 2021; Camp, 2007).
THEORY	A label for a future theory, or a synonym for an existing theory.	The theory must focus on the vehicle (recall Box 1; e.g. Aguilera, 2016; Camp, 2007; Egan, 2020) and avoid Ryle's and van Rooij's regresses (recall Box 2).

provided us with a case study with conclusions and potential lessons learned that apply to any proposed theory of mental representations or of cognition in general. First, let us look at the lessons specifically for the cognitive map, before taking stock of what we could learn for theorising and metatheorising about mental representations in general.

Potential repairs, or indeed complete reconceptualisation and re-formalisation, of the cognitive map, such as the work by Mariela Aguilera (2021) and Elisabeth Camp (2007), are likely to be the only way forward, one of many steps on a long journey for those who wish to retain cognitive maps. Other attempts to theorise using the cognitive map, as described above, regardless of the use of implementation-level (viz. Guest & Martin, 2021) — both senses: appealing to substrates in the brain, and appealing to computational mechanism (recall Box 1) — computational models, do not provide a relevant or solid basis for continued theorising.

To wit, cognitive map as-is is likely a rhetorical spandrel (Gould & Lewontin, 1979), i.e. a remnant of problematic otherwise abandoned theorising, and therefore has a deleterious effect on our scientific reasoning. It is usefully framed as a historical remnant from a time when behaviourism was staunchly anti-cognitivist (recall **What is a cognitive map?**). Importantly, this role, which the cognitive map played arguably well in Tolman's time, has ended; the field moved to a more compatibilist stance, thus mainstreaming the use of mental representation (e.g. Favela and Machery, 2023; cf. Amundson, 1983). Notwithstanding, and recalling the careful threading of the needle described in Box 1:

> It is a paradox that the "Tolmaniacs" [Tolman's students] from Berkeley who tend to speak of rats as "little furry people" are much more likely to search for central neural mechanisms than S-R[i.e. stimulus-response] theorists who speak of rats as "little furry machines."
>
> John Garcia (1976, p. 81)

To recapitulate our main point, herein we have built the case for cognitive map needing careful use, mindful deployment when verbally and formally theorising. And so in the case of neurocognitive mental representations cognitive maps can be "systematically misleading" per Ryle, i.e.

> the sense in which such quasi-ontological statements are misleading is not that they are false and not even that any word in them is equivocal or vague, but only that they are formally improper to the facts of the logical form which they are employed to record and proper to facts of quite another logical form.
>
> Gilbert Ryle (1931, p. 150)

Solutions encompass more than throwing away the presently improper cognitive map completely (viz. Bennett, 1996). Such drastic action is not required, especially if a scientist is committed to map-like representations or similar (viz. Aguilera, 2016, 2018; Camp, 2007; Casati & Varzi, 1999; Schellenberg, 2018). To put it simply, the practitioner who wishes to be a non-naive computationalist and to keep such phraseology in their modelling and theorising needs to shift focus to formally specifying the neurocognitive computational mechanisms and functions, the vehicle itself (recall Box 1; cf. Egan, 2017; Smortchkova et al., 2020). A zooming out from the status quo described in Table 1 is required to then zoom back in with fresh eyes. In the most methodologically and terminologically committed case, the scientist can think deeply about what it is that 'cognitive map' instantiates, if not (just) a theory (see Table 2).

To provide some rectification to cognitive map, we must answer: is cognitive map just a synonym for mental representation, under a scheme of map-like or spatially-based mental representations? Then the practitioner should take heed of the constraints for such a sub-theory case (see relevant row of Table 2). Such mutually inclusive (re)conceptualisations of cognitive map as phenomenon (first row, Table 2) all the way to full-blown theory (last row, Table 2) require increasingly more theoretical and formal work as we move down the table. Notwithstanding, choosing at least one row of Table 2 is required for intra- and interdisciplinary communication, theorising, and modelling endeavours, if we wish to retain the cognitive map. Because when, for example,"we know little about how the acquisition of cognitive maps is shaped by different features of exploratory behavior" (Brunec et al., 2023, p. 1), what is most important is how to house them in a theory than continue collecting data (viz. Buzsáki & Tingley, 2023; Guest & Martin, 2021; Hardcastle, 1996; Hardcastle & Stewart, 2002; van Rooij & Baggio, 2021).

Our analyses have been in service of understanding how scientists within our fields explain mental representations of externally posited structures. Given that we have used the scientific entity of the cognitive map as a case study for theorising with respect to mental represention, what general lessons can be drawn out?

First, we warn that externally posited structures, such as landmarks (recall Figure 1) or other features of experiments, such as of tasks and of stimuli are representational content. Content is a pedagogical or illustrative gloss (recall Box 1; Egan, 2020) that cannot constitute the theory. Thus confusing representational content for representational vehicles not only causes homonculus-like infinite regress in the explanatory chain (recall Box 2), it also causes the related problem for evaluating formal theory through obfuscating the human in the loop who does the work under a computational veneer. In other words, when van Rooijan or Rylean regresses take place, solving them through human-in-the-loop techniques, as discussed in **All ghost, no machine** is computational modelling only in name. In practice, such formalisms and computational accounts are no different to the titular aphorism and opening quote, wherein a recipe for a cake lists 'buy a cake' as the first and only step (Dretske, 1994; recall Box 2). Indubitably, such a method works in terms of output, but offers no insight into cake making — in this context the success-to-truth inference is a fallacious one (recall Box 2). And this holds no matter how much we might study the process of travelling to the confectioner's, picking out a cake, and buying it. Cognitive map as-is is a non- or even anti-mechanistic account portrayed as mechanistic.

Second, we warn that any scientific entity, not just the cognitive map, is misleading if use by the broader field is not openly discussed — especially to focus on what it is the entity is capturing (recall Table 1). Furthermore, in the case of neurocognitive theorising, we are on the hunt for an explanation as to how the representations are created and used. These problems likely exist in, and the critiques we make can apply to, many cases, e.g. from mirror neurons (neurons are held to fire, represent, when the organism both performs and observes another perform an action; Heyes, 2010; Rizzolatti & Sinigaglia, 2008) to predictive coding (brain areas are held to represent predictions of other brain ares or of aspects of the environment; Grush, 2004; Rao & Ballard, 1999) — when we implicate neurocognitive accounts of mental representation, we must consider what such accounts claim to offer versus their status as-is. In our case study, cognitive map could be vehiculized (Aguilera, 2021; Camp, 2007), but this is not something reflected in the extant research within cognitive and computational neuroscience, as we have shown in **All form, no function**. In other words, if the cognitive map becomes commonly accepted and formally ensconced as a vehicle, we will be on a much firmer footing for reasoning about such mental representations. But such a vehiculization will still be agnostic as to how such presentations emerge, which appears a mechanistic desideratum broadly. Nonetheless, retaining the cognitive map but reconceptualising it is possible (recall Table 2), e.g. preserving a commitment to map-like representations (as content), or towards reconciling such representations (as both content and vehicle) with other types.

Third, in the case of reading mental representation off the brain or other similar framings of data, we caution against reasoning that takes as the antecedent correlational matching between external posited structures, e.g. features of task or stimuli, and the data we collect. This form of metatheoretical reasoning harms our thinking as theory is underdetermined by data, multiple realisability interferes with correlational matches, and computationalism in general should never descend into the naive variety (recall Box 1). Metatheoretical calculi, ways we adjudicate over theories, that permit such reasoning are minimally errant and maximally damaging (recall Box 2; Guest, 2024; Guest and Martin, 2023, 2024). No amount of correlations will ever license the conclusion that the neurocognitive system represents in a certain way — correlation does not imply cognition (viz. Guest & Martin, 2023). Cognitive map as-is is a non- or even anti-theoretical account portrayed as theoretical.

Thus, it is both the case that:

> Picture theorists often explicitly deny the claim that there are literally pictures in the brain. Yet appealing to a pictorial format to explain experimental phenomena invariably requires such a literal picture. [...] What has sometimes been called a 'functional space' (such as a matrix data structure) will not do because such a space, being a fiction, can have any properties we like.
>
> Zenon W. Pylyshyn (2003, p. 114)

And that:

> Setting the problem in functional terms does not mean that maps, spatial representations, images of "the world in the head" do not exist. We will never be able to see what they are, but we already have some notion of how they work. Furthermore, according to the dynamic nature of the cognitive mapping system, maps

> should be subjected to continuous changes and hence difficult for investigators to grasp.
>
> Catherine Thinus-Blanc (1987, p. 14)

In other words, from the correlations nothing follows. Our formal and verbal theories are the antecedents (Guest & Martin, 2023, 2024). And it is those theories that we should strive to formalise, improve, and curate. What remains true from the time of Tolman to now is that if "the only sure criterion is to have fun", (Tolman, 1959, p. 152) then not all our ideas will be formally defensible within the hard bounds of non-naive computationalism.

References

Aguilera, M. (2016). Cartographic systems and non-linguistic inference. *Philosophical Psychology*, *29*(3), 349–364 (cit. on pp. 3, 9–11).

Aguilera, M. (2018). Why the content of animal thought cannot be propositional. *Análisis Filosófico*, *38*(2), 183–207 (cit. on pp. 9–11).

Aguilera, M. (2021). Heterogeneous inferences with maps. *Synthese*, *199*(1), 3805–3824 (cit. on pp. 10, 11).

Aly, M., & Turk-Browne, N. B. (2017). How hippocampal memory shapes, and is shaped by, attention. In D. E. Hannula & M. C. Duff (Eds.), *The hippocampus from cells to systems: Structure, connectivity, and functional contributions to memory and flexible cognition* (pp. 369–403). Springer International Publishing. (Cit. on p. 3).

Ambrogioni, L., & Ólafsdóttir, H. F. (2023). Rethinking the hippocampal cognitive map as a meta-learning computational module. *Trends in Cognitive Sciences*, *27*(8), 702–712 (cit. on pp. 6, 8).

Amundson, R. (1983). Ec tolman and the intervening variable: A study in the epistemological history of psychology. *Philosophy of Science*, *50*(2), 268–282 (cit. on p. 10).

Andrews, M., Smart, A., & Birhane, A. (2024). The reanimation of pseudoscience in machine learning and its ethical repercussions. *Patterns*, *5*(9) (cit. on p. 5).

Attneave, F. (1961). In defense of homunculi. *Sensory communication*, 777–782 (cit. on p. 7).

Bäckström, S., & Gustafsson, M. (2017). Skill, drill, and intelligent performance: Ryle and intellectualism. *Journal for the History of Analytical Philosophy* (cit. on pp. 5, 7).

Bainbridge, L. (1983). Ironies of automation. In *Analysis, design and evaluation of man–machine systems* (pp. 129–135). Elsevier. (Cit. on p. 5).

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M. J., Degris, T., Modayil, J., et al. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429–433 (cit. on p. 1).

Barry, C., Lever, C., Hayman, R., Hartley, T., Burton, S., O'Keefe, J., Jeffery, K. J., & Burgess, N. (2006). The boundary vector cell model of place cell firing and spatial memory. *Reviews in the Neurosciences*, *17*(1-2), 71–98 (cit. on p. 5).

Bausch, M., Niediek, J., Reber, T. P., Mackay, S., Boström, J., Elger, C. E., & Mormann, F. (2021). Concept neurons in the human medial temporal lobe flexibly represent abstract relations between concepts. *Nature communications*, *12*(1), 6164 (cit. on p. 5).

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509 (cit. on pp. 2, 6).

Benhamou, S. (1996). No evidence for cognitive mapping in rats. *Animal Behaviour*, *52*(1), 201–212 (cit. on pp. 2, 4).

Bennett, A. T. (1996). Do animals have cognitive maps? *Journal of Experimental Biology*, *199*(1), 219–224 (cit. on pp. 1–3, 11).

Blokpoel, M. (2018). Sculpting computational-level models. *Topics in cognitive science*, *10*(3), 641–648 (cit. on pp. 7, 10).

Bouchekioua, Y., Blaisdell, A. P., Kosaki, Y., Tsutsui-Kimura, I., Craddock, P., Mimura, M., & Watanabe, S. (2021). Spatial inference without a cognitive map: The role of higher-order path integration. *Biological Reviews*, *96*(1), 52–65 (cit. on p. 4).

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological review*, *116*(1), 220 (cit. on pp. 4, 5).

Breed, M. D. (2017). *Conceptual breakthroughs in ethology and animal behavior*. Academic Press. (Cit. on p. 2).

Brunec, I. K., Nantais, M. M., Sutton, J. E., Epstein, R. A., & Newcombe, N. S. (2023). Exploration patterns shape cognitive map learning. *Cognition*, *233*, 105360 (cit. on p. 11).

Bush, D., Barry, C., & Burgess, N. (2014). What do grid cells contribute to place cell firing? *Trends in neurosciences*, *37*(3), 136–145 (cit. on p. 5).

Buzsáki, G., & Tingley, D. (2023). Cognition from the body-brain partnership: Exaptation of memory. *Annual review of neuroscience*, *46*(1), 191–210 (cit. on pp. 6, 11).

Calvo Tapia, C., Tyukin, I., & Makarov, V. A. (2020). Universal principles justify the existence of concept cells. *Scientific reports*, *10*(1), 7889 (cit. on pp. 4, 5).

Camp, E. (2007). Thinking with maps. *Philosophical perspectives*, *21*, 145–182 (cit. on pp. 1, 5, 9–11).

Campbell, M. B. (2010). Artificial men: Alchemy, transubstantiation, and the homunculus. *Republics of Letters: A Journal for the Study of Knowledge, Politics, and the Arts*, *1*(2), 4–15 (cit. on p. 5).

Carlson, T., Goddard, E., Kaplan, D. M., Klein, C., & Ritchie, J. B. (2018). Ghosts in machine learning for cognitive neuroscience: Moving from data to theory. *NeuroImage*, *180*, 88–100 (cit. on pp. 1, 8).

Carroll, D. W. (2017). *Purpose and cognition: Edward tolman and the transformation of american psychology*. Cambridge University Press. (Cit. on p. 2).

Casati, R., & Varzi, A. C. (1999). *Parts and places: The structures of spatial representation*. MIT press. (Cit. on p. 11).

Catani, M. (2017). A little man of some importance. *Brain*, *140*(11), 3055–3061 (cit. on p. 5).

Chirimuuta, M. (2013). Minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. *Synthese*, *191*(2), 127–153 (cit. on pp. 1, 8).

Chirimuuta, M. (2018a). Explanation in computational neuroscience: Causal and non-causal. *The British Journal for the Philosophy of Science* (cit. on p. 2).

Chirimuuta, M. (2018b). Marr, Mayr, and MR: What functionalism should now be about. *Philosophical Psychology*, *31*(3), 403–418 (cit. on p. 3).

Chirimuuta, M. (2021). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*, *199*(1), 767–790 (cit. on p. 7).

Chirimuuta, M. (2024). *The brain abstracted: Simplification in the history and philosophy of neuroscience*. MIT Press. (Cit. on p. 1).

Cohen, M. A., Dilks, D. D., Koldewyn, K., Weigelt, S., Feather, J., Kell, A. J., Keil, B., Fischl, B., Zöllei, L., Wald, L., et al. (2019). Representational similarity precedes category selectivity in the developing ventral visual pathway. *NeuroImage*, *197*, 565–574 (cit. on pp. 1, 8).

Connor, C. E. (2005). Friends and grandmothers. *Nature*, *435*(7045), 1036–1037 (cit. on p. 4).

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*(6292), 1464–1468 (cit. on p. 8).

Craig, M. T., & McBain, C. J. (2015). Navigating the circuitry of the brain's gps system: Future challenges for neurophysiologists. *Hippocampus*, *25*(6), 736–743 (cit. on pp. 2, 6, 8).

Dall'Orso, S., Steinweg, J., Allievi, A., Edwards, A., Burdet, E., & Arichi, T. (2018). Somatotopic mapping of the developing sensorimotor cortex in the preterm human brain. *Cerebral cortex*, *28*(7), 2507–2515 (cit. on p. 5).

Darden, L. (2006). Discovering mechanisms in neurobiology: The case of spatial memory with carl f. craver. In *Reasoning in biological discoveries: Essays on mechanisms, interfield relations, and anomaly resolution* (pp. 40–64). Cambridge University Press. (Cit. on pp. 2, 3).

Dawson, M. R. (2013). *Mind, body, world: Foundations of cognitive science*. Athabasca University Press. (Cit. on pp. 5, 6).

Dennett, D. C. (2017). *Brainstorms: Philosophical essays on mind and psychology*. MIT press. (Cit. on p. 7).

Dhein, K. (2023). The cognitive map debate in insects: A historical perspective on what is at stake. *Studies in History and Philosophy of Science*, *98*, 62–79 (cit. on pp. 2, 4, 10).

Dostrovsky, J., & O'Keefe, J. (1971). The hippocampus as a spatial map. preliminary evidence from unit activity in the freely moving rat. *Brain research*, *34*(1), 171–175 (cit. on p. 5).

Downs, R. M., & Stea, D. (Eds.). (1973). *Image and environment: Cognitive mapping and spatial behavior*. Routledge. (Cit. on p. 3).

Dretske, F. (1994). If you can't make one, you don't know how it works. *Midwest Studies In Philosophy*, *19*(1), 468–482 (cit. on pp. 1, 5, 7, 11).

Eckhardt, B. V. (2004). Connectionism and the propositional attitudes. In C. E. Erneling (Ed.), *The mind as a scientific object: Between brain and culture*. Oxford University Press. (Cit. on p. 5).

Egan, F. (1995). Folk psychology and cognitive architecture. *Philosophy of Science*, *62*(2), 179–196 (cit. on p. 8).

Egan, F. (1999). In defence of narrow mindedness. *Mind & Language*, *14*(2), 177–194 (cit. on pp. 6, 7).

Egan, F. (2010). Computational models: A modest role for content. *Studies in History and Philosophy of Science Part A*, *41*(3), 253–259 (cit. on p. 9).

Egan, F. (2017). Function-theoretic explanation. *Explanation and integration in mind and brain science*, 145–163 (cit. on pp. 3, 11).

Egan, F. (2018). The nature and function of content in computational models. In *The routledge handbook of the computational mind* (pp. 247–258). Routledge. (Cit. on pp. 3, 9).

Egan, F. (2020). A deflationary account of mental representation. *What are mental representations*, 26–53 (cit. on pp. 1, 3, 6, 9–11).

Epstein, R. A., Patai, E. Z., Julian, J. B., & Spiers, H. J. (2017). The cognitive map in humans: Spatial navigation and beyond. *Nature neuroscience*, *20*(11), 1504–1513 (cit. on pp. 2, 6, 8).

Erscoi, L., Kleinherenbrink, A. V., & Guest, O. (2023). Pygmalion displacement: When humanising ai dehumanises women. *SocArXiv. February*, *11* (cit. on p. 5).

Farzanfar, D., Spiers, H. J., Moscovitch, M., & Rosenbaum, R. S. (2023). From cognitive maps to spatial schemas. *Nature Reviews Neuroscience*, *24*(2), 63–79 (cit. on pp. 4, 8).

Favela, L. H., & Machery, E. (2023). Investigating the concept of representation in the neural and psychological sciences. *Frontiers in Psychology*, *14*, 1165622 (cit. on p. 10).

Fenton, A. A. (2024). Remapping revisited: How the hippocampus represents different spaces. *Nature Reviews Neuroscience*, 1–21 (cit. on p. 4).

Figdor, C. (2010). Neuroscience and the multiple realization of cognitive functions. *Philosophy of Science*, *77*(3), 419–456 (cit. on p. 3).

Figdor, C. (2018a). The fallacy of the homuncular fallacy. *Belgrade Philosophical Annual*, *31*(31), 41–56 (cit. on p. 5).

Figdor, C. (2018b). *Pieces of mind: The proper domain of psychological predicates*. Oxford University Press. (Cit. on p. 5).

Fodor, J. A. (1968). The appeal to tacit knowledge in psychological explanation. *The Journal of Philosophy*, *65*(20), 627–640 (cit. on pp. 5, 6).

Frietsch, U. (2021). Alchemy and the early modern university: An introduction. (Cit. on p. 5).

Fyhn, M., Molden, S., Witter, M. P., Moser, E. I., & Moser, M.-B. (2004). Spatial representation in the entorhinal cortex. *Science*, *305*(5688), 1258–1264 (cit. on p. 5).

Gallistel, C. R. (1989). Animal cognition: The representation of space, time and number. *Annual review of psychology* (cit. on p. 3).

Garcia, J. (1976). I. krechevsky and i. In L. Petrinovich & J. L. McGaugh (Eds.), *Knowing, thinking, and believing: Festschrift for professor david krech* (pp. 71–84). Springer US. (Cit. on p. 10).

Good, J., & Still, A. (1986). Tolman and the tradition of direct perception. *British Journal of Psychology*, *77*(4), 533–539 (cit. on p. 2).

Gordon, E. M., Chauvin, R. J., Van, A. N., Rajesh, A., Nielsen, A., Newbold, D. J., Lynch, C. J., Seider, N. A., Krimmel, S. R., Scheidter, K. M., et al. (2023). A somato-cognitive action network alternates with effector regions in motor cortex. *Nature*, *617*(7960), 351–359 (cit. on p. 5).

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of san marco and the panglossian paradigm: A critique of the adaptationist programme. *Proc. R. Soc. Lond. B*, *205*, 581–598 (cit. on pp. 4, 10).

Grieves, R. M., & Jeffery, K. J. (2017). The representation of space in the brain. *Behavioural processes*, *135*, 113–131 (cit. on p. 5).

Gross, C. G. (2002). Genealogy of the "grandmother cell". *The Neuroscientist*, *8*(5), 512–518 (cit. on p. 5).

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and brain sciences*, *27*(3), 377–396 (cit. on p. 11).

Guest, O. (2024). What makes a good theory, and how do we make a theory good? *Computational Brain & Behavior*, 1–15 (cit. on pp. 1–3, 5–7, 11).

Guest, O., Caso, A., & Cooper, R. P. (2020). On simulating neural damage in connectionist networks. *Computational brain & behavior*, *3*, 289–321 (cit. on pp. 2, 9).

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802 (cit. on pp. 3, 7, 10, 11).

Guest, O., & Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior* (cit. on pp. 2–4, 7, 8, 11, 12).

Guest, O., & Martin, A. E. (2024). "a metatheory of classical and modern connectionism" (cit. on pp. 2–5, 7, 9, 11, 12).

Guthrie, E. R. (1935). *Psychology of learning.* Harper. (Cit. on pp. 2, 3, 6, 9).

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801–806 (cit. on pp. 2, 5).

Hammond, K. R. (1976). The social implementation of cognitive theory. *Knowing, thinking, and believing: Festschrift for Professor David Krech*, 245–260 (cit. on p. 2).

Hannula, D. E., & Duff, M. C. (2017). *The hippocampus from cells to systems: Structure, connectivity, and functional contributions to memory and flexible cognition*. Springer. (Cit. on p. 2).

Hardcastle, V. G. (1995). Computationalism. *Synthese*, *105*, 303–317 (cit. on p. 3).

Hardcastle, V. G. (1996). *How to build a theory in cognitive science*. State University of New York Press. (Cit. on pp. 3, 7, 9, 11).

Hardcastle, V. G., & Stewart, C. M. (2002). What do brain data really show? *Philosophy of Science*, *69*(S3), S72–S82 (cit. on p. 11).

Harris, S., & Harris, D. (2015). *Digital design and computer architecture*. Morgan Kaufmann. (Cit. on p. 5).

Heyes, C. (2010). Where do mirror neurons come from? *Neuroscience & Biobehavioral Reviews*, *34*(4), 575–583 (cit. on p. 11).

Horner, A. J., Bisby, J. A., Zotow, E., Bush, D., & Burgess, N. (2016). Grid-like processing of imagined navigation. *Current Biology*, *26*(6), 842–847 (cit. on p. 4).

Hornsby, J. (2011). Ryle's knowing how and knowing how to act. In J. Bengson & M. A. Moffett (Eds.), *Knowing how: Essays on knowledge, mind, and action* (p. 80). Oxford University Press USA. (Cit. on pp. 5, 7).

Hurley, S. L. (1998a). *Consciousness in action*. Harvard University Press. (Cit. on p. 5).

Hurley, S. L. (1998b). Vehicles, contents, conceptual structure, and externalism. *Analysis*, *58*(1), 1–6 (cit. on pp. 1, 3).

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, *32* (cit. on p. 4).

Innis, N. K. (1999). Edward c. tolman's purposive behaviorism. (cit. on p. 2).

Jacobs, L. F. (2003). The evolution of the cognitive map. *Brain, behavior and evolution*, *62*(2), 128–139 (cit. on p. 3).

Jafarpour, A., & Spiers, H. (2017). Familiarity expands space and contracts time. *Hippocampus*, *27*(1), 12–16 (cit. on p. 8).

Jeffery, K. J., Grieves, R., & Donnett, J. (2018). Recording the spatial mapping cells: Place, head direction, and grid cells. In *Handbook of behavioral neuroscience* (pp. 95–121, Vol. 28). Elsevier. (Cit. on p. 5).

Jensen, R. (2006). Behaviorism, latent learning, and cognitive maps: Needed revisions in introductory psychology textbooks. *The Behavior Analyst*, *29*, 187–209 (cit. on pp. 2, 6).

Jones, P. (2021). *Work without the worker: Labour in the age of platform capitalism*. Verso Books. (Cit. on p. 5).

Kaas, J. H. (1997). Topographic maps are fundamental to sensory processing. *Brain research bulletin*, *44*(2), 107–112 (cit. on p. 5).

Kitchin, R. M. (1994). Cognitive maps: What are they and why study them? *Journal of environmental psychology*, *14*(1), 1–19 (cit. on pp. 1, 2).

Konishi, M. (1986). Centrally synthesized maps of sensory space. *Trends in Neurosciences*, *9*, 163–168 (cit. on pp. 2, 5).

Kristan, W. B., & Katz, P. (2006). Form and function in systems neuroscience. *Current biology*, *16*(19), R828–R831 (cit. on p. 7).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25* (cit. on p. 5).

Langille, J. J., & Gallistel, C. R. (2020). Locating the engram: Should we look for plastic synapses or information-storing molecules? *Neurobiology of Learning and Memory*, *169*, 107164 (cit. on p. 4).

Lever, C., Burton, S., Jeewajee, A., O'Keefe, J., & Burgess, N. (2009). Boundary vector cells in the subiculum of the hippocampal formation. *Journal of Neuroscience*, *29*(31), 9771–9777 (cit. on p. 5).

Lisman, J., Buzsáki, G., Eichenbaum, H., Nadel, L., Ranganath, C., & Redish, A. D. (2017). Viewpoints: How the hippocampus contributes to memory, navigation and cognition. *Nature neuroscience*, *20*(11), 1434–1447 (cit. on pp. 2–4, 8).

Litch, M. (1997). Computation, connectionism and modelling the mind. *Philosophical Psychology*, *10*(3), 357–364 (cit. on p. 3).

Lycan, W. G. (1991). Homuncular Functionalism Meets PDP. In W. Ramsey, S. P. Stich, & D. M. Rumelhart (Eds.), *Philosophy and connectionist theory*. Lawrence Erlbaum. (Cit. on pp. 5, 9).

MacCorquodale, K., & Meehl, P. E. (1954). Edward c. tolman. *Modern learning theory*, 177–266 (cit. on pp. 2, 8).

Mackintosh, N. J. (2002). Do not ask whether they have a cognitive map, but how they find their way about. *Psicológica*, *23*(1) (cit. on pp. 2, 9, 10).

Maguire, E. A., Burgess, N., & O'Keefe, J. (1999). Human spatial navigation: Cognitive maps, sexual dimorphism, and neural substrates. *Current opinion in neurobiology*, *9*(2), 171–177 (cit. on pp. 2, 4).

Marozzi, E., & Jeffery, K. J. (2012). Place, space and memory cells. *Current Biology*, *22*(22), R939–R942 (cit. on pp. 2, 4–6).

Marr, D. (1982). *Vision*. W. H. Freeman. (Cit. on p. 8).

McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I., & Moser, M.-B. (2006). Path integration and the neural basis of the'cognitive map'. *Nature Reviews Neuroscience*, *7*(8), 663–678 (cit. on p. 2).

Millikan, R. G. (1991). Perceptual content and fregean myth. *Mind*, *100*(4), 439–459 (cit. on pp. 1, 3).

Millikan, R. G. (2021). Neuroscience and teleosemantics. *Synthese*, *199*(1), 2457–2465 (cit. on pp. 7, 10).

Mok, R. M., & Love, B. C. (2019). A non-spatial account of place and grid cells based on clustering models of concept

learning. *Nature communications*, *10*(1), 5685 (cit. on pp. 4, 8).
Moran, R., Dayan, P., & Dolan, R. J. (2021). Human subjects exploit a cognitive map for credit assignment. *Proceedings of the National Academy of Sciences*, *118*(4), e2016884118 (cit. on p. 4).
Morris, M. (1991). Why there are no mental representations. *Minds and Machines*, *1*, 1–30 (cit. on p. 7).
Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, *31*(1), 69–89 (cit. on p. 4).
Moser, M.-B., & Moser, E. I. (1998). Distributed encoding and retrieval of spatial memory in the hippocampus. *Journal of Neuroscience*, *18*(18), 7535–7542 (cit. on p. 2).
Moser, M.-B., & Moser, E. I. (2016). Where am i? where am i going? *Scientific American* (cit. on p. 2).
Murase, A. (2020). The homunculus and the paracelsian liber de imaginibus. *ambix*, *67*(1), 47–61 (cit. on p. 5).
Newcombe, N., & Liben, L. S. (1982). Barrier effects in the cognitive maps of children and adults. *Journal of Experimental Child Psychology*, *34*(1), 46–58 (cit. on p. 4).
Nizami, L. (2018). Reductionism ad absurdum: Attneave and dennett cannot reduce homunculus (and hence the mind). *Kybernetes*, *47*(1), 163–185 (cit. on pp. 5, 7).
O'Connor, M. (2019). *Wayfinding: The science and mystery of how humans navigate the world*. St. Martin's Press. (Cit. on p. 3).
O'Keefe, J. (1994). Cognitive maps, time and causality (cit. on p. 4).
O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain research* (cit. on p. 2).
O'Keefe, J., & Nadel, L. (1978). The hippocampus as a cognitive map. (Cit. on p. 2).
Peer, M., Brunec, I. K., Newcombe, N. S., & Epstein, R. A. (2021). Structuring knowledge with cognitive maps and cognitive graphs. *Trends in cognitive sciences*, *25*(1), 37–54 (cit. on p. 4).
Penfield, W., & Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, *60*(4), 389–443 (cit. on p. 5).
Pfaffenberger, B. (1988). Fetishised objects and humanised nature: Towards an anthropology of technology. *Man*, 236–252 (cit. on p. 5).
Piccinini, G. (2007). Computationalism, the church–turing thesis, and the church–turing fallacy. *Synthese*, *154*, 97–120 (cit. on p. 6).
Polger, T. W., & Shapiro, L. A. (2016). *The multiple realization book*. Oxford University Press. (Cit. on p. 3).
Popov, V., Ostarek, M., & Tenison, C. (2018). Practices and pitfalls in inferring neural representations. *NeuroImage*, *174*, 340–351 (cit. on pp. 1, 8).
Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological bulletin*, *80*(1), 1 (cit. on pp. 1, 2).
Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioral and brain sciences*, *25*(2), 157–182 (cit. on p. 1).
Pylyshyn, Z. W. (2003). Return of the mental image: Are there really pictures in the brain? *Trends in cognitive sciences*, *7*(3), 113–118 (cit. on pp. 1, 11).
Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102–1107 (cit. on p. 5).
Quiroga, R. Q. (2012). Concept cells: The building blocks of declarative memory functions. *Nature Reviews Neuroscience*, *13*(8), 587–597 (cit. on pp. 4, 5).
Ranck, J. B. (1985). Head direction cells in the deep cell layer of dorsolateral pre-subiculum in freely moving rats. *Electrical activity of the archicortex* (cit. on p. 5).
Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79–87 (cit. on p. 11).
Reddy, L., & Thorpe, S. J. (2014). Concept cells through associative learning of high-level representations. *Neuron*, *84*(2), 248–251 (cit. on p. 5).
Reid, A. K., & Staddon, J. E. (1997). A reader for the cognitive map. *Information sciences*, *100*(1-4), 217–228 (cit. on p. 3).
Rescorla, M. (2009). Predication and cartographic representation. *Synthese*, *169*, 175–200 (cit. on p. 9).
Rescorla, M. (2020, December). 135Reifying Representations [_eprint: https://academic.oup.com/book/0/chapter/338900569/chapter-pdf/57910625/oso-9780190686673-chapter-6.pdf]. In *What are Mental Representations?* Oxford University Press. (Cit. on p. 9).
Rich, E., et al. (2008). *Automata, computability and complexity: Theory and applications*. Pearson Prentice Hall Upper Saddle River. (Cit. on p. 6).
Rich, P., de Haan, R., Wareham, T., & van Rooij, I. (2021). How hard is cognitive science? *Proceedings of the annual meeting of the cognitive science society*, *43*(43) (cit. on pp. 5, 7).
Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British journal for the philosophy of science* (cit. on pp. 1, 8).
Rizzolatti, G., & Sinigaglia, C. (2008). *Mirrors in the brain: How our minds share actions and emotions*. Oxford University Press, USA. (Cit. on p. 11).
Ross, L. N. (2020). Multiple realizability from a causal perspective. *Philosophy of Science*, *87*(4), 640–662 (cit. on p. 3).
Ross, L. N., & Bassett, D. S. (2024). Causation in neuroscience: Keeping mechanism meaningful. *Nature Reviews Neuroscience*, *25*(2), 81–90 (cit. on pp. 1, 8).
Rowland, D. C., Roudi, Y., Moser, M.-B., & Moser, E. I. (2016). Ten years of grid cells. *Annual review of neuroscience*, *39*(1), 19–40 (cit. on p. 5).
Ryle, G. (1931). Systematically misleading expressions. *Proceedings of the Aristotelian society*, *32*, 139–170 (cit. on p. 10).
Ryle, G. (1949). *The concept of mind*. Barnes & Noble. (Cit. on pp. 5, 7).
Schaeffer, R., Khona, M., & Fiete, I. (2022). No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. *Advances in neural information processing systems*, *35*, 16052–16067 (cit. on p. 4).
Schellenberg, S. (2018). *The unity of perception: Content, consciousness, evidence*. Oxford University Press. (Cit. on pp. 1, 5, 7, 11).
Schiller, D., Eichenbaum, H., Buffalo, E. A., Davachi, L., Foster, D. J., Leutgeb, S., & Ranganath, C. (2015). Memory and space: Towards an understanding of the cognitive map. *The Journal of Neuroscience*, *35*(41), 13904–13911 (cit. on p. 7).

Schlichting, M. L., & Preston, A. R. (2017). The hippocampus and memory integration: Building knowledge to navigate future decisions. In D. E. Hannula & M. C. Duff (Eds.), *The hippocampus from cells to systems: Structure, connectivity, and functional contributions to memory and flexible cognition* (pp. 405–437). Springer International Publishing. (Cit. on p. 3).

Schmahmann, J. D. (2019). The cerebellum and cognition. *Neuroscience letters*, *688*, 62–75 (cit. on p. 5).

Shanahan, M. (2016). The Frame Problem. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2016). Metaphysics Research Lab, Stanford University. (Cit. on p. 7).

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press. (Cit. on p. 2).

Shettleworth, S. J. (2010). *Cognition, evolution, and behavior, second edition* (2nd ed.). (Cit. on pp. 2–4).

Simon, C. (2022). "you can be a behaviorist and still talk about the mind–as long as you don't put it into a person's head": An interview with howard rachlin. *Journal of the Experimental Analysis of Behavior*, *119*(1), 9 (cit. on p. 2).

Skinner, D. M., Etchegary, C. M., Ekert-Maret, E. C., Baker, C. J., Harley, C. W., Evans, J. H., & Martin, G. M. (2003). An analysis of response, direction and place learning in an open field and t maze. *Journal of Experimental Psychology: Animal Behavior Processes*, *29*(1), 3 (cit. on p. 2).

Smortchkova, J., Murez, M., Dolega, K., & Schlicht, T. (2020). Representational kinds. *What are Mental Representations*, 213–241 (cit. on p. 11).

Solstad, T., Boccara, C. N., Kropff, E., Moser, M.-B., & Moser, E. I. (2008). Representation of geometric borders in the entorhinal cortex. *Science*, *322*(5909), 1865–1868 (cit. on p. 5).

Spanton, R. W., & Guest, O. (2022). Measuring trustworthiness or automating physiognomy? A comment on Safra, Chevallier, Grèzes, and Baumard (2020). *arXiv preprint arXiv:2202.08674* (cit. on p. 5).

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature neuroscience*, *20*(11), 1643–1653 (cit. on p. 4).

Tanney, J. (2011). Ryle's regress and the philosophy of cognitive science. *JL Austin et La Philosophie du Langage Ordinaire*, 447–67 (cit. on pp. 5, 7, 9).

Tanney, J. (2013). Ryle's conceptual cartography. In *The historical turn in analytic philosophy* (pp. 94–110). Springer. (Cit. on pp. 6, 7).

Tanney, J. (2022). Gilbert Ryle. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2022). Metaphysics Research Lab, Stanford University. (Cit. on p. 7).

Taube, J. S., Muller, R. U., & Ranck, J. B. (1990a). Head-direction cells recorded from the postsubiculum in freely moving rats. i. description and quantitative analysis. *Journal of Neuroscience*, *10*(2), 420–435 (cit. on p. 5).

Taube, J. S., Muller, R. U., & Ranck, J. B. (1990b). Head-direction cells recorded from the postsubiculum in freely moving rats. ii. effects of environmental manipulations. *Journal of Neuroscience*, *10*(2), 436–447 (cit. on p. 5).

Taylor, A. (2018). The automation charade. *Logic Magazine*, *5*(1) (cit. on p. 5).

Thinus-Blanc, C. (1987). The cognitive map concept and its consequences. In *Cognitive processes and spatial orientation in animal and man: Volume i experimental animal psychology and ethology* (pp. 1–19). Springer. (Cit. on pp. 2–4, 6, 7, 12).

Thinus-Blanc, C. (1996). *Animal spatial cognition: Behavioural and brain approach*. World Scientific Publishing Company. (Cit. on pp. 2–4).

Tolman, E. C. (1932). *Purposive behavior in animals and men*. Appleton-Century. (Cit. on p. 2).

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, *55*(4), 189 (cit. on p. 2).

Tolman, E. C. (1959). Psychology: A study of a science. (Cit. on pp. 4, 12).

Udin, S. B., & Fawcett, J. W. (1988). Formation of topographic maps. *Annual review of neuroscience*, *11*(1), 289–327 (cit. on p. 5).

van Rooij, I. (2003). *Tractable cognition: Complexity theory in cognitive psychology* [Doctoral dissertation]. (Cit. on pp. 3, 4, 6, 7).

van Rooij, I. (2008). The tractable cognition thesis. *Cognitive science*, *32*(6), 939–984 (cit. on pp. 3, 4, 6–8).

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science [PMID: 33404356]. *Perspectives on Psychological Science*, *16*(4), 682–697 (cit. on pp. 7, 11).

van der Gun, L., & Guest, O. (2023, July). Artificial intelligence: Panacea or non-intentional dehumanisation? (Cit. on p. 5).

van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (2019). *Cognition and Intractability: A Guide to Classical and Parameterized Complexity Analysis*. Cambridge University Press. (Cit. on pp. 3, 10).

van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., & Rich, P. (2024). Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior*, 1–21 (cit. on pp. 3, 5, 7).

Vigotsky, A. D., Iannetti, G. D., & Apkarian, A. V. (2024). Mental state decoders: Game-changers or wishful thinking? *Trends in Cognitive Sciences* (cit. on pp. 1, 8).

Weisberg, S. M., & Newcombe, N. S. (2018). Cognitive maps: Some people make them, some people struggle. *Current directions in psychological science*, *27*(4), 220–226 (cit. on p. 4).

Wikenheiser, A. M., & Redish, A. D. (2015). Decoding the cognitive map: Ensemble hippocampal sequences and decision making. *Current opinion in neurobiology*, *32*, 8–15 (cit. on p. 4).

Woensdregt, M., Blokpoel, M., van Rooij, I., & Martin, A. E. (2024). Challenges for a computational explanation of flexible linguistic inference. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46* (cit. on pp. 6, 7).

Wright, H., & Foerder, P. (2021). The missing female homunculus. *Leonardo*, *54*(6), 653–656 (cit. on p. 5).

Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A., Clopath, C., et al. (2023). Catalyzing next-generation artificial intelligence through NeuroAI. *Nature communications*, *14*(1), 1597 (cit. on p. 4).

Zednik, C. (2014). Are systems neuroscience explanations mechanistic? https://philsci-archive.pitt.edu/10859/ (cit. on pp. 1, 8).

Zipser, D. (1985). A computational model of hippocampal place fields. *Behavioral neuroscience*, *99*(5), 1006 (cit. on p. 9).